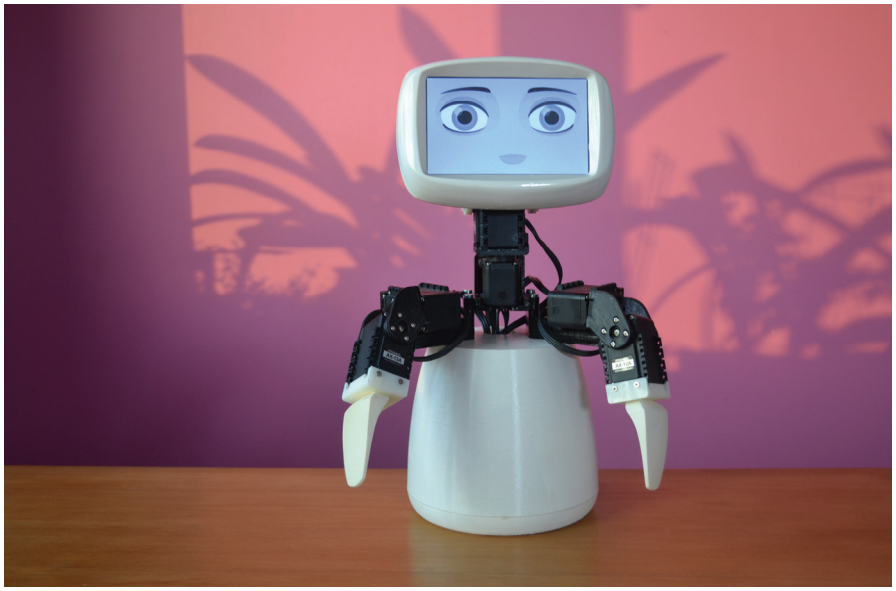
International conference

LINGUISTIC FORUM 2020:
LANGUAGE AND ARTIFICIAL INTELLIGENCE

12–14 November 2020

Institute of Linguistics RAS, Moscow, Russia



Международная конференция

ЛИНГВИСТИЧЕСКИЙ ФОРУМ 2020:
ЯЗЫК И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

12–14 ноября 2020 г.

Москва, Институт языкознания РАН

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ ЯЗЫКОЗНАНИЯ РАН

International conference “Linguistic Forum 2020: Language and Artificial Intelligence”. 12–14 November 2020. Institute of Linguistics RAS, Moscow, Russia: Abstracts / Ed. by Andrej A. Kibrik, Valentin Gusev, Dmitry Zalmanov. — 164 p. — Moscow: Institute of linguistics RAS, 2020.

Международная конференция «Лингвистический форум 2020: Язык и искусственный интеллект». 12–14 ноября 2020 г. Институт языкознания РАН, Москва: Тезисы докладов / Под ред. А. А. Кибрика, В. Ю. Гусева, Д. А. Залманова. — 164 с. — М.: Институт языкознания РАН, 2020. — ISBN 978-5-6045633-3-5

Программный комитет:
А. С. Бердичевский
П. И. Браславский
А. А. Кибрик (председатель)
Э. С. Клышинский
А. А. Котов
С. О. Кузнецов
А. Б. Кутузов
В. К. Логачева
Н. В. Лукашевич
О. Н. Ляшевская
Д. И. Муромцев
Б. В. Орехов
† Г. С. Осипов
Н. К. Рябцева
В. А. Салимовский
В. П. Селегей
И. В. Смирнов
В. Д. Соловьёв
Д. Талбот
Е. А. Филатова
С. А. Шаров

Program committee:
Aleksandrs Berdicevskis
Pavel Braslavski
Elena Filatova
Andrej A. Kibrik (chair)
Eduard Klyshinskiy
Artemy Kotov
Andrey Kutuzov
Sergei Kuznetsov
Varvara Logacheva
Natalia Loukachevitch
Olga Lyashevskaya
Dmitry Muromtsev
Boris Orekhov
† Gennady Osipov
Nadezhda Riabtseva
Vladimir Salimovsky
Vladimir Selegey
Sergei Sharov
Ivan Smirnov
Valery Solovyev
David Talbot

ISBN 978-5-6045633-3-5
9 785604 563335

© Authors
© Institute of linguistics RAS, 2020

© Авторы
© Институт языкознания РАН, 2020

CONTENTS СОДЕРЖАНИЕ

С. О. Кузнецов. Искусственный интеллект. Основные направления и вызовы 6
Aleksandrs Berdicevskis. Corpora, vectors, models: computational approaches to linguistic typology 7
Н. В. Лукашевич. Автоматические методы построения таксономий по текстам 8
David Talbot. Learning about language from data 9
Ilia Afanasev. Corpus-dictionary system: introducing a concept 10
Angelina Bolshina, Natalia Loukachevitch. Monosemous relatives approach to automatic data labelling for word sense disambiguation in Russian 12
Mariya Khudyakova, Natalya Gronskaya, Natalia Antonova, Maria Nelyubina, Anastasia Surova, Maria Shibanova, Anna Vorobyova, Olga Dragoy. Characteristics of spoken discourse across different elicitation tasks 14
Mariya Khudyakova, Galina Ryazanskaya, Tatyana Shishkovskaya. Verbal fluency task in schizophrenia and schizoaffective disorder: automated analysis 17
Artemiy Kotov. Consciousness in computer architectures: The theory of expelled scripts 19
Irina Ovchinnikova. Impact of the human factor on errors in computer-assisted translation 21
Natalia Petrova. Originality of speech perspectivization in cinematic arts 23
Nadezhda Riabtseva. Artificial vs. human intelligence in digital vs. human translation 25
Timur Sadykov, Timur Zhukov. Machine learning approach towards automatic text synthesis in Russian 28
Leonardo Sanna. Evaluating word embeddings for computational pragmatics 30
Vera Zabotkina, Dmitri Orlov, Elena Pozdnyakova. Topic modeling and expert assessment as tools for concepts identification in the THREATCorpus 32
Liudmila Zaidelman, Artemiy Kotov, Zakhar Nosovets, Vadim Ushakov, Vera Zabotkina, Boris Velichkovsky. Semantic spaces in dictionary and in the brain 34
Н. Ш. Александрова. Понимание речи человеком: двойственность организации 37
З. Н. Афинская. Умный (*smart*) в сфере искусственного интеллекта 39
Е. А. Бауэр. Некоторые аспекты невербальной коммуникации в психотерапии 40
В. П. Белянин. К построению субъективных тезаурусов 42
Л. Б. Бойко. Восприятие названия поликодового текста в аспекте перевода 44
Е. Г. Борисова. Модель для обработки недискретных данных в лингвистическом описании 46
В. В. Бочкарев, А. В. Шевлякова. Метод корпусного исследования дистрибуции и семантики неологизмов 48
В. В. Бочкарев, А. В. Шевлякова, Г. В. Парамей, Е. В. Рахилина. Микродиахроническое исследование терминов *бурый* и *коричневый* на материале корпуса Google Books Ngram 50
Е. Л. Боярская, Е. В. Шевченко, И. В. Томашевская. Динамика концептуальной структуры аксиологической категории 52
Е. В. Буденная, Ю. В. Николаева, А. А. Евдокимова. Эвиденциальные маркеры в кинетических паттернах говорящего 54
В. Л. Введенский, К. Г. Гуртовой. Структура тезауруса русских прилагательных, определенная по измерениям времени распознавания слов на слух 56
А. В. Вдовиченко. Семиотическое (воз)действие и возможности искусственного интеллекта: условный «тест Тьюринга» как тестирование испытателя 58
Б. М. Величковский, В. И. Заботкина, Л. Я. Зайдельман, А. А. Котов, З. А. Носовец, В. Л. Ушаков. Нейросемантика: латерализация репрезентаций менталистских понятий 62

Л. Л. Волкова, Д. Ю. Мальцева. О расширении запросов для поиска фактов на естественном языке 64
Л. Л. Волкова, А. А. Котов. Эмоциональные реакции робота-компаньона на фразы на русском языке 66
О. А. Волошина. Приемы мнемотехники в Ригведе (о способах заучивания, хранения в памяти и функционировании древнеиндийских гимнов) 68
В. В. Глебкин. Ономасиологический подход к проблеме лексических и семантических изменений 70
Ю. А. Грибер, Г. В. Парамей. Возрастные различия репертуара цветонаименований у современных носителей русского языка 72
Н. Э. Гронская, О. В. Драгой, Н. Ю. Антонова, М. В. Шибанова, М. С. Нелюбина, К. С. Яшин, И. А. Медяник. Дефицит связной речи у пациентов с опухолями головного мозга в раннем послеоперационном периоде 74
А. А. Гурин. Анализ эмоциональной тональности текстов на русском языке 76
М. В. Дайтер, Е. И. Риехакайнен. Использование методов машинного обучения для изучения фонетической редукции слов в русской устной речи 79
Д. А. Девяткин, Л. А. Каджая, В. А. Мишланов, В. А. Салимовский. Автоматическая идентификация ментальных действий, конституирующих речевые жанры научного теоретического текста 81
О. В. Донина, Д. С. Дмитриев. Возможность использования методов машинного обучения для автоматического выявления стертых метафор 83
К. А. Елистратова. О некоторых компонентах «картины мира» современного подростка ... 85
Д. А. Залманов, А. А. Кибрик. Уровень активации vs. специальные ограничения при математическом моделировании референциального выбора 88
Н. Л. Зелянская, К. И. Белоусов, Р. В. Чумаков. Цифровое измерение «наивной» географии..... 90
В. В. Иванов, М. И. Солнышкина. Метод оценки сложности текста, основанный на онтологиях 92
О. Ю. Инькова. Несвязность текста, или О некоторых подводных камнях на пути автоматической обработки текста 94
В. В. Казаковская, Н. К. Онипенко. Вводно-модальные слова как недискретное явление в языке 96
А. А. Кибрик, Г. Б. Добров, Н. А. Коротаев. Мультиканальный ресурс и моделирование естественной коммуникации 98
П. О. Кисель. Раздельно-слитные биномы *лихэцы* как проявление недискретности в современном китайском языке 100
А. В. Колмогорова, А. А. Калинин, А. В. Маликова. Специфика объективации категории наблюдателя в нейтральных и эмоциональных текстах через призму сентимент-анализа 102
Е. Н. Колпачкова, Е. И. Мовшович. Средства обеспечения связности в современном китайском языке 104
А. Д. Кошелев. Иерархия человеческих понятий, заданная значениями предметных существительных 106
С. А. Крылов, Е. Н. Никитина, Н. К. Онипенко, М. А. Станкевич. Типы грамматической информации в «Семантико-грамматическом словаре русских глаголов» и их применение в компьютерном анализе текстов (на материале глаголов с каузативно-эмотивной семантикой) 109
Ю. М. Кузнецова. Особенности жанра популярных медицинских рекомендаций в контексте задач когнитивного ассистирования 111
Н. В. Лавров. Контекстное поле диалога и платформа эмоциональных переживаний в аспекте алгоритмов коммуникации 113

Ю. Е. Лещенко, Т. И. Доценко, Т. С. Остапенко. Комбинированный трилингвизм как источник моделирования искусственной нейронной сети 115
О. В. Ломтатидзе, А. С. Алексеева, Е. С. Исакова. Использование показателей движения глаз при изучении восприятия текстов на русском и английском языках 117
А. А. Лопухина, С. М. Гольдина, А. К. Лауринавичюте, О. В. Драгой. Специфика движений глаз при чтении у детей с дислексией 119
Н. Е. Лысенко. Способность давать показания у лиц с нарушениями слуха и речи 121
С. А. Малютина, Е. Д. Лапин, М. Д. Терехина. Лексическое разнообразие у молодых и пожилых людей: эксперимент на заканчивание предложений 123
В. А. Мишланов, Л. А. Каджая, Е. Н. Махмутова, И. В. Смирнов. Об автоматической идентификации текстов с заданной модальностью (на материале текстов сетевой коммуникации с тематикой профессионального кризиса) 125
Е. А. Огородникова, О. В. Лабутина, И. Г. Андреева, А. П. Гвоздева, Ю. А. Баулин. Фактор просодики в восприятии коммуникативной сцены с пространственным разделением источников речи и речеподобной помехи 127
П. Б. Паршин. «Язык и искусственный интеллект»: неформатные соображения о формулировке темы форума 129
М. А. Пильгун. Речевое поведение акторов социальных медиа в городских конфликтах: нейросетевой подход 131
И. Г. Подгорбунская. К вопросу об измерении количества и ценности информации в языковой единице вторичной номинации (на примере идиом английского языка) 132
К. К. Поливанов, О. А. Казакевич, Н. В. Сердобольская, З. М. Халилова, Е. М. Будянская, А. П. Евстигнеева, К. О. Мищенкова, Д. Д. Мордашова, С. В. Покровская, Е. А. Ренковская. Интернет-ресурс «Малые языки России»: способы визуализации языковых данных 134
В. С. Сазонова, Е. И. Гавришина. Инструменты для создания корпуса литературных текстов на малоресурсных языках 136
М. В. Самулеева , А. А. Смирнова, З. А. Зорина. Формирование эквивалентных отношений между знаком и референтом зависит от типа последнего 138
У. В. Смирнова. Транслингвальные особенности знаков (на примере рекламных вывесок центральных улиц города Москвы) 140
В. Д. Соловьев, Н. В. Лукашевич. Распределение квазисинонимов в тезаурусе для автоматической обработки текстов 142
А. И. Тимофеев. О способе исчисления параметров в образном представлении 145
С. Ю. Толдова, Т. В. Давыдова, М. В. Кобозева, Д. Б. Писаревская. Неканонические элементарные дискурсивные единицы (ЭДЕ) в текстах блогов (на материале разметки Ru-RSTreebank) 147
М. Н. Усачёва. Жанровые и модусные особенности элицитированных дискурсов 150
О. Л. Филипеня, В. В. Ткаченко. Лингвокогнитивный метод алгоритмизации процессов логического мышления 152
А. Б. Хомяков. О видах смыслов и применении их в искусственном интеллекте 154
Е. В. Чистова. Менеджмент мультимодальной коммуникации синхронных переводчиков 156
С. А. Шаповал. Качественный анализ ошибок в текстах нейросети Порфирьевич 158
С. Х. Шихалиева. Корпус диалектов и изучение вариативности говоров в табасаранском онлайн-словаре 160
И. А. Щирова, А. Г. Гурочкина. Проблема категорий текста: сложность, динамика и некоторые перспективы решения 162

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ. ОСНОВНЫЕ НАПРАВЛЕНИЯ И ВЫЗОВЫ

С. О. Кузнецов

skuznetsov@hse.ru

НИУ Высшая школа экономики, Российская ассоциация искусственного интеллекта

В докладе, посвященном памяти двух первых президентов Российской ассоциации искусственного интеллекта (РАИИ) Д. А. Поспелова и Г. С. Осипова, будут обсуждаться вопросы, связанные с содержанием понятия «Искусственный интеллект» (ИИ), историей развития ИИ, основными направлениями и последними достижениями в области ИИ, а также вызовами разного рода.

CORPORA, VECTORS, MODELS: COMPUTATIONAL APPROACHES TO LINGUISTIC TYPOLOGY

Aleksandrs Berdicevskis
aleksandrs.berdicevskis@gu.se
University of Gothenburg

In this paper, I will make several strong claims about language change and typology. The claims should be perceived not as reliably established truths, but rather as my views of how we can productively approach certain research questions. The claims are as follows:

1. When studying language change, we should try to explain it, but be aware that it may not always be possible (see Claims 2 and 3). Explanations can involve both language-internal and language-external (such as social, cultural, cognitive or even geographical) factors.

2. Sometimes the change cannot be explained since there is no «explanation» as such, just a complex interaction of random processes. Even in these cases, however, it can sometimes be established whether certain kinds of change are more likely than others.

3. Sometimes the change cannot explained simply because there is not enough data. Explanation typically involves causal claims, and establishing causal links is difficult. Even if change A in language X occurs after event B, this fact *per se* is usually not enough to claim that B causes A. This fact, however, can be enough to hypothesize about a causal link, and the hypothesis should considered true if:

a) it is supported by nomothetic (large-scale) quantitative cross-linguistic studies;

b) the causal mechanism (how B can cause A) is convincingly described, and it has been shown that this mechanism really works (for instance, via psycholinguistic experiments);

c) there is converging evidence from various other sources (corpus studies; diachronic studies; computational modelling etc).

4. The distribution of properties across the world's languages is partly shaped by weak universal pressures on language change generated by extra-linguistic factors (cf. Claims 1 and 2). Thus, understanding language change is the key to answering many questions traditionally viewed as belonging to language typology (and vice versa). To an extent, sociolinguistics, diachronic linguistics and typology should be viewed as facets of the same research direction, sometimes labeled *language evolution*.

5. One productive way of approaching the evolutionary questions is using the opportunities of modern natural language processing: corpus linguistics and beyond.

I will illustrate how this research agenda can be pursued using three examples from my own work.

АВТОМАТИЧЕСКИЕ МЕТОДЫ ПОСТРОЕНИЯ ТАКСОНОМИЙ ПО ТЕКСТАМ

Н. В. Лукашевич

louk_nat@mail.ru

МГУ имени М. В. Ломоносова

Для повышения качества автоматической обработки текстов часто нужно использовать знания: лексические (синонимы, антонимы и др.), знания о мире или о конкретной предметной области. Для представления таких знаний вручную создаются специализированные компьютерные ресурсы: таксономии, онтологии, тезаурусы. При этом часто возникает вопрос, насколько можно автоматизировать извлечение такого рода знаний из текстовых коллекций. В докладе будут описаны методы автоматического извлечения родо-видовых (таксономических) отношений из текстов, представлены подходы к их тестированию и достигаемые результаты.

LEARNING ABOUT LANGUAGE FROM DATA

David Talbot
talbot@yandex-team.ru
Yandex Translate

Modern neural network-based methods, known as “deep learning”, have transformed natural language processing (NLP) over the past 10 years with unprecedented progress on tasks such machine translation, question answering, dialogue systems and text generation. This isn’t the first time that statistical learning has taken the field of NLP hostage, leaving apparently little room for linguistics, but somehow this time it feels different.

In this talk, I will summarize how the field of NLP has changed over the past 10 years under the influence of deep learning, how this is similar to previous waves of empiricism and how it differs, which problems have been solved, which remain illusive and why deep learning, while ostensibly pushing linguistics out of picture, may in fact be opening up new research directions for linguists.

CORPUS-DICTIONARY SYSTEM: INTRODUCING A CONCEPT

Ilia Afanasev
szrnamerg@gmail.com
St. Petersburg State University

Dictionaries and corpora are, and have been since the emergence of the latter, probably the main source of understanding the language structure, and language models verification. Dictionaries tend to give more information about what a particular language unit is, while these are corpora that help to contextualize it, understand its place in the system as a whole. Thus, we are dealing with paradigmatic and syntagmatic type of interaction between a researcher and a language. But while using only one of them we are not able to acquire the full-fledged description, and this leads us to the idea of merging two ways. Basically, what if we create a parallel corpus-and-dictionary system?

Creating a dictionary from a corpus of representative size is a hard task, which is hardly possible to be done manually by a single researcher, a small group, or even a big group within the adequate timeframe, not being prolonged for dozens of years. Humans need help for such task, and this help might be provided by a suitable kind of software. However, what research might require this technology?

It is the quite wide in recent years discussion on relationship and hybridization between the Old Croatian and the Croatian Church Slavonic languages on the territory of modern Croatia during Middle Ages (Kozak et al 2019).

The question of the research being conducted is the following: how does one distinguish an Old Croatian text from a Croatian Church Slavonic (or is it even possible)? To do that, one should do both dictionary-based and corpus-based research. And the best way to do that would be to create a system that prepares both dictionary and corpus, and dealing with results of its job as a linguist afterwards.

The lack of software that can automatically build such a system, in case a researcher wants to do it for any language or even languages, became quite obvious, as well as the way to create it.

Firstly, the universal system (a corpus-dictionary creating software blueprint) is brought into existence. Secondly, it is adapted for languages required. Thirdly, the corpus-based and dictionary-based linguistic research on the material collected is held. In the conclusion, there is a full (or as close to full as possible) description of a text, on basis of which one can draw conclusions on text nature, and continue moving to understanding of relationship between Old Croatian, Croatian Church Slavonic and their hybrids. In fact, from this corpus-dictionary system the description of languages is born.

Thus, the main requirement for the system is that it should be able to create both corpus (with at least tokenization function) and dictionary as separate units.

To create corpus, the system should acquire a text, provided by user, conduct preprocessing (for instance, in Urdu it leads to resolving “context sensitivity, where the letters change their shape with respect to their letters” (Daud et al 2017: 283)), sentence (optionally) and word tokenization (which in some languages is quite a hard process to perform (Attia 2007: 65)); after this phase is finished, everything should be put in a specially organized database, either web or desktop, which can be accessed by user via suitable graphical user interface (GUI).

To create dictionary, the system should acquire a list of words, normalize and lemmatize it, then place in order, designed by user (default option would be alphabetical), enhancing with some sort of additional information, which is also provided to it. Access to the information via GUI is desirable as well.

The following step is an interface for interaction between these two units. It can be achieved by performing two separate tasks. Firstly, the program gets a raw text, preprocesses and tokenizes it, producing a text entity, a list of sentence (or clause, or paragraph, or sequence) entities, and a list of lists of token entities, organized by sentence entities. Secondly, it merges all token entities into a single list, normalizes, lemmatizes, and sorts it, thus creating a dictionary for a single text, with each dictionary item possessing data on the tokens to which it is related. Information on normalization and lemmatization of each token entity is returned back to the corpus. Optionally, a user might add information either to the text in the corpus, or to the items in the dictionary, and this information should be redirected to the other side of the system. Finally, there should be a dictionary for the whole corpus, which is going to be merged and recompiled each time a change in a collection happens. However, the immediateness of these changes is yet to be considered, due to the potential size of the created corpora, and potential time for its recompilation.

The method of data storage data is under consideration. Generally, language corpora use XML markup, as it is for a big amount of other data in systems all over the world. However, nowadays there is a growing tendency of using JSON for the same purposes, which is applied in our solution.

The solution is to be represented via web or desktop application, integrated to the whole system.

The workflow was tested on the material of a fragment of a Glagolitic missal from the XIV c. (The National Library of Russia, Berčić II 36).

As a conclusion, it is safe to say that by now the results of the experiment on a particular dataset are satisfactory and are to be presented. A trial version of a corpus-dictionary system passed preliminary tests. However, it still needs to get more flexible and robust, which is the next step of the development. Also, to achieve the overall aim of the research, being disambiguation of Old Croatian, Croatian Church Slavonic and their hybrids, more texts should be analyzed and compared.

References

Attia M. A. 2007. Arabic Tokenization System // Proceedings of the 5th Workshop on Important Universal Matters. Madison: Omnipress, pp. 65—72.

Daud A., Khan W., Che D. 2017. Urdu language processing: a survey // Artificial Intelligence Review, vol. XLVII, pp. 279–311.

Kozak V. V., Makarova A. L., Sobolev A. N. 2019. O jezicima hrvatskoglagoljske liturgije u svjetlu djela dr. Luke Jelića», Luka Jelić (1864.–1922.) — istraživač i znanstvenik: valorizacija znanstvenog opusa. Zadar.

OpenCorpora. Доступ: http://opencorpora.org/ (22.04.2020)

MONOSEMOUS RELATIVES APPROACH TO AUTOMATIC DATA LABELLING FOR WORD SENSE DISAMBIGUATION IN RUSSIAN[1]

Angelina Bolshina
angelina_ku@mail.ru
Lomonosov Moscow State University

Natalia Loukachevitch
louk_nat@mail.ru
Lomonosov Moscow State University
Kazan Federal University

Word Sense Disambiguation (WSD) is a key task in computational semantics and aims to identify the correct sense of polysemous words. Various supervised methods including kNN, Naive Bayes, SVM, neural networks are applied to this task (Navigli 2009), but to achieve high performance, supervised WSD algorithms require large sense-annotated datasets. The annotation of such corpora demands considerable time and human resources, that is why supervised machine learning approaches suffer from a knowledge acquisition bottleneck. Not for all languages hand-crafted sense-annotated corpora are available, and this also holds true for the Russian language. Automatic acquisition of training samples can help to tackle this problem. Our research is focused on the monosemous relatives approach, which exploits a set of unambiguous words (or phrases) related to particular senses of a polysemous word.

Usually WordNet (Miller 1995) is used as a source for such relatives. WordNet is a lexical-semantic resource for the English represented in form of semantic graph. All words in this network are grouped into sets of synonyms that are called synsets. Monosemous relatives are the words or phrases that have only one sense (i.e. belong only to one synset) and are related to the target ambiguous word through some connection in WordNet. In our research we exploit Russian wordnet RuWordNet (Loukachevitch et al. 2016).

A fragment of monosemous relatives list for the word *такса* ‘dachshund’:

(1) «*охотничий пёс, охотничья собака, пёсик, четвероногий друг, псина,* …» / ‘hunting dog, hunting dog, doggie, four-legged friend, dog …’

In this work we utilized word2vec embedding models to extract the most similar words to each monosemous relative according to word contexts. These similar words represent a distributional set of close words with the respective similarity measures.

The central idea of our candidate selection and ranking algorithm is based on the assumption that a training collection can be built not only with the direct relations like synonymy, hypernymy and hyponymy but also with far more distant words, such as cohyponyms. The principal features of our approach are as follows:

1. We take into consideration not only the closest relatives to a target word sense, as it was done in previous works, but also more distant relatives located within 4 steps from a target polysemous word sense. A target word sense is a sense of a polysemous word that we want to disambiguate.
2. We utilize similarity scores between a candidate monosemous relative and synsets close to a sense of a target polysemous word in order to evaluate how well this candidate can represent the sense of an ambiguous word. Candidate monosemous relatives are unambiguous words and phrases, that can be located in up to four-step relation paths to a polysemous word.
3. To measure the relevance and suitability of a monosemous candidate, we exploit a thesaurus set of words similar to a target sense which we call ***a synset nest***. The group of syno-

[1] The work of Loukachevitch N. in the current study concerns formulation of the disambiguation approach for RuWordNet data, calculation of paths between synsets, criteria for selecting contexts; this work is supported by the Russian Science Foundation grant no. 19-71-10056 financed through Kazan Federal University.

nyms to a target sense and all the words from directly related synsets within 2 steps from a target word comprise ***the synset nest*** for a target sense.

4. We check similarity scores to the synset nest for both closest and further located monosemous relatives because a word described as monosemous in the thesaurus can actually have polysemous usage in a corpus. These similarity scores are also exploited to calculate the resulting weight of a monosemous relative.

The candidate ranking algorithm identifies which monosemous relatives are most similar to the target ambiguous word's sense. As a result of this procedure, all monosemous relatives are sorted by the weight they obtained. Then we extract from the corpus the contexts in which these relatives occur, substitute them with the target ambiguous word in these texts and add the samples to a training collection.

The following fragment of list of monosemous relatives with similarity scores (given in brackets) was obtained for the noun *гвоздика* 'clove':

(2) *чёрный перец* 'black pepper' (7.5), *кардамон* 'cardamom' (6.8)*, корица* 'cinnamon' (6.5)*, имбирь* 'ginger' (6.4)*, мускатный орех* 'nutmeg' (6) ….

We utilized two corpora in the research. A news corpus consists of news articles harvested from various news sources. Another corpus is Proza.ru, a segment of Taiga corpus (Shavrina and Shapovalova 2017), which is compiled of works of prose fiction. We exploit these two corpora to train the respective word2vec models and to compare the performance of the WSD models trained on the collections obtained with these resources. For evaluation of our algorithm of training data generation, we used three distinct RUSSE'18 datasets for Russian (Panchenko et al. 2018). From these datasets we have taken only those words and senses which have one-to-one correspondence with senses in RuWordNet.

We conducted several experiments to determine whether our text collections can be used as training datasets for a WSD model. We used nearest neighbor classifier (kNN) based on the contextualized word embeddings ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019). The model based on the ELMo pre-trained embeddings by RusVectōrēs outperformed all other models achieving 0.857 F1 score.

References

Devlin J., Chang M.-W., Lee K., Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.

Loukachevitch N. V., Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V. 2016. Creating Russian WordNet by Conversion // Proceedings of Conference on Computational Linguistics and Intellectual Technologies Dialog-2016, pp. 405–415.

Miller G. 1995. WordNet: A Lexical Database for English // Communications of the ACM, vol. 38(11), pp. 39–41.

Navigli R. 2009. Word sense disambiguation: A survey. ACM computing surveys (CSUR), vol. 41(2), 10.

Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Arefyev N., Leontyev A., Loukachevitch N. 2018. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", pp. 547–564.

Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. 2018. Deep contextualized word representations // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2227–2237.

Shavrina T., Shapovalova O. 2017. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser // Proceedings of "CORPORA2017", international conference, Saint-Petersburg.

CHARACTERISTICS OF SPOKEN DISCOURSE ACROSS DIFFERENT ELICITATION TASKS[1]

Mariya Khudyakova
mariya.kh@gmail.com

Natalya Gronskaya
ngronskaya@hse.ru

Natalia Antonova
natalie.eskadron@gmail.com

Maria Nelyubina
marnelyubina@gmail.com

Anastasia Surova
asurova909@gmail.com

Maria Shibanova
mshibanova4@gmail.com

Anna Vorobyova
vorobyovaaa2015@gmail.com

Olga Dragoy
olgadragoy@gmail.com

National Research University Higher School of Economics

Introduction. Analysis of spontaneous speech is a crucial part of language assessment in clinical populations since it allows to evaluate different aspects of language production on micro- and macrolinguistic levels (see cf. Bryant et al. 2017; Linnik et al. 2015; Prins, Bastiaanse 2004). The most common tasks for speech elicitation include conversations (e.g., Armstrong et al. 2011; Goodglass et al. 2001), personal stories (Armstrong et al. 2011; Behrns et al. 2009; Olness, Ulatowska 2011) or picture-elicited discourse (e.g Howard et al. 2010; Williams et al. 2010).

However, there is evidence that the quantitative characteristics of connected speech are influenced by the genre of discourse (Fergadiotis et al. 2011; Fergadiotis, Wright 2011; Olness 2006). Given that, we argue that for comprehensive assessment of discourse production normative samples of a number of discourse genres should be collected and analyzed. Moreover, for proper assessment of language abilities at different time points (e.g. before and after treatment) several equivalent elicitation tasks for each discourse genre should be developed.

We present design for the analysis of the normative sample of discourse across three elicitation tasks in a group of non-brain damaged individuals.

Method and Material. Discourse samples were collected online on the Finding Five platform (FindingFive 2019) from 27 native Russian speakers with no neurological or psychiatric disorders from three age groups: young adults (8 females, 2 males; mean age — 21.20, SD = 3.01; mean N of year of education — 14,20, SD = 2.30), middle-aged adults (5 females, 3 males; mean age — 43.25, SD = 5.34; mean N of year of education — 18.13, SD = 1.89), and older adults (5 females, 4 males; mean age — 56.33, SD = 3.71; mean N of year of education — 15.89, SD = 1.69).

For elicitation of different discourse genres, we prepared three types of tasks: picture-elicited narratives, personal stories and procedural discourse. For each task type three different equivalent tasks were selected. As stimuli for picture-elicited narratives three comics by Herluf Bidstrup with a single main character were selected ("Superman", "Discovery of the World", "Wonderful Day"). For personal story production three questions about a significant personal event were selected (*What was the best or most memorable gift you have received? What was the best or most memorable trip you have taken? What was the best or most memorable party you have had?*). Procedural discourse production is based on a sequence of pictures adapted from IKEA furniture assembly instructions without any accompanying text (a chair, a table and a bench).

Each participant was presented with one task of each type in a fixed order. To account for priming effects the combinations of tasks were balanced across experimental lists.

Annotation. The discourse will be annotated ELAN (Wittenburg et al. 2006). The annotation scheme includes orthographic transcription and segmentation into elementary discourse units (EDUs) and utterances based on syntactic characteristics (similar to Bergelson, Khudyakova 2020).

[1] The study was funded by the Russian Science Foundation, project No. 20-18-00399

Pauses, false-starts, perseverations, automatic expressions and errors are annotated on several tiers, as well as lemmas and part-of-speech information. Each EDU is rated based on its grammatical completeness on a three-point scale, and the grammatical complexity of each utterance is rated on a three-point scale. Additionally, each EDU is annotated as pertaining to one of the key macrocomponents of spoken discourse: mainline, background, comment, meta-comment and discourse marker (cf. Bergelson, Khudyakova 2020; Cahana-Amitay, Jenkins 2018; Labov 2008; Olness, Ulatowska 2011).

Data analysis. For the analysis the following measures will be used: speech rate measured as words per minute, mean length of EDU, percentage of pauses, rate of disfluencies (false starts and perseverations), lexical diversity, average grammatical completeness of EDUs, average grammatical complexity of utterances, the rate of different macrocomponents.

Analysis will be conducted with linear mixed-effect models with task type and task variant as fixed effects.

References

Armstrong E., Ciccone N., Godecke E., & Kok B. 2011. Monologues and dialogues in aphasia : Some initial comparisons // Aphasiology, 25(11), pp. 1347–1371.

Behrns I., Wengelin Å., Broberg M., & Hartelius L. 2009. A comparison between written and spoken narratives in aphasia // Clinical Linguistics & Phonetics, 23(7), pp. 507–528.

Bergelson M. B., & Khudyakova M. V. 2020. Narrative discourse segmentation in clinical linguistics // In Search of Basic Units of Spoken Language. A Corpus-Driven Approach, pp. 258–284.

Bryant L., Spencer E., & Ferguson A. 2017. Clinical use of linguistic discourse analysis for the assessment of language in aphasia // Aphasiology, 31(10), pp. 1105–1126.

Cahana-Amitay D., & Jenkins T. 2018. Working memory and discourse production in people with aphasia // Journal of Neurolinguistics, 48, pp. 90–103.

Fergadiotis G., & Wright H. H. 2011. Lexical diversity for adults with and without aphasia across discourse elicitation tasks // Aphasiology, 25(11), pp. 1414–1430.

Fergadiotis G., Wright H. H., & Capilouto G. J. 2011. Productive vocabulary across discourse types // Aphasiology, 25(10), pp. 1261–1278.

FindingFive. 2019. FindingFive: A web platform for creating, running, and managing your studies in one place. FindingFive Corporation (nonprofit). findingfive.com

Goodglass H., Kaplan E., & Barresi B. 2001. Boston Diagnostic Aphasia Examination. Record booklet.

Howard D., Swinburn K., & Porter G. 2010. Putting the CAT out: What the Comprehensive Aphasia Test has to offer // Aphasiology, 24(1), pp. 56–74.

Labov W. 2008. Oral narratives of personal experience // Cambridge Encyclopedia of the Language Sciences, pp. 546–548. https://doi.org/10.1037/a0022354

Linnik A., Bastiaanse R., & Höhle B. 2015. Discourse production in aphasia: a current review of theoretical and methodological challenges. 7038 (December).

Olness G. S. 2006. Genre, verb, and coherence in picture-elicited discourse of adults with aphasia // Aphasiology, 20(2/3/4), pp. 175–187.

Olness G. S., Ulatowska H. K. 2011. Personal narratives in aphasia: Coherence in the context of use // Aphasiology, 25(11), pp. 1393–1413.

Prins R., Bastiaanse R. 2004. Analyzing the spontaneous speech of aphasic speakers // Aphasiology, 18(12), pp. 1075–1091.

Williams C., Thwaites A., Buttery P., Geertzen J., Randall B., Shafto M., Devereux B., Tyler L. K. 2010. The Cambridge Cookie-Theft Corpus: A corpus of directed and spontaneous speech of brain-damaged

patients and healthy individuals // LREC 2010. Proceedings of the 7th International Conference on Language Resources and Evaluation, November 2015, pp. 2824–2830.

Wittenburg P., Brugman H., Russel A., Klassmann A., Sloetjes H. 2006. ELAN: a Professional Framework for Multimodality Research // Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

VERBAL FLUENCY TASK IN SCHIZOPHRENIA AND SCHIZOAFFECTIVE DISORDER: AUTOMATED ANALYSIS

Mariya Khudyakova
mariya.kh@gmail.com

Galina Ryazanskaya
galka1999@gmail.com

National Research University Higher School of Economics

Tatyana Shishkovskaya
tszyszkowska@gmail.com

Mental Health Research Center

Thought disorder (TD) is a set of specific disturbances in thought, speech, and communication that is typical for psychosis in schizophrenia and schizoaffective disorder. It includes positive symptoms such as the flight of ideas, neologisms, loose associations and tangentiality (Hart & Lewine 2017). The loose associations symptom, being one of the most common signs of TD, is rather problematic to measure in a fast and reliable manner. One of the possible solutions is to automatically measure the typicality of the patient's associations in a verbal fluency (VF) task.

VF task is a standardized screening tool used in psychiatry, neurology and clinical linguistics. It consists of naming as many words from one category (semantic or phonetic) as possible under a time limit (usually one minute). The task is simple and sensitive to language disorders and thus it is included in many cognitive batteries (e.g. Keefe et al. 2004). People with schizophrenia are known to produce fewer words on VF task than controls (Bokat et al. 2003; Juhasz et al. 2012), and with patterns of semantic relations between words different from healthy speakers (e.g. Robert et al. 1998). One of the methods to measure the loosening of associations is the analysis of clusters in the VF task output. Automated methods of such analysis include averaging over cosine similarities of adjacent words (Elvevåg et al. 2007) and computing a cluster fraction (Rosenstein et al. 2015).

In this study we present automated analysis of VF task data by people with schizophrenia and schizoaffective disorders, and healthy speakers. We compared automated and manual clustering of the VF output and analyzed the correlation of cluster size and number with scores on a psychiatric assessment scale.

Method and Material. Participants with schizophrenia and schizoaffective disorder (10 females, 5 males, mean age — 28.46 years, SD = 7.41) were recruited at the Russian Mental Health Research Center. Healthy participants (18 females, 3 males, mean age — 24.10 years, SD = 4.37) passed the threshold on the scores on the general mental health questionnaire, SCL–90-R — Symptom Checklist-90-Revised (Derogatis & Savitz 2000). All participants were assessed by a psychiatrist with the Positive and Negative Syndrome Scale (PANSS; Kay et al. 1987).

The participants were asked to name as many animals as they could in one minute. Subjective manual clustering of the transcribed responses was performed by three raters individually (interrater agreement: Cohen's K = 0.7). The gold standard clustering was defined as the set of all cluster boundaries that at least two annotators agreed on.

For preliminary automated analysis we used word2vec model tayga_none_fasttextcbow_300_10_2019 from RusVectores website (Kutuzov & Kuzmenko 2017) to calculate the average cosine similarity across a response and the cluster fraction (number of clusters in relation to the number of words produced). The number of clusters was calculated using several various threshold values on cosine similarity, as described by Kim and colleagues (2019).

Results. Splitting at the global mean cosine similarity was shown to be the best method for approximating manual (strong correlation with cluster number, rho = 0.75; good model performance for cluster boundary positioning, f_1-measure = 0.73).

Independent-sample t-test has revealed significant differences between groups in unique word count ($p_{adj} < 0.001$), mean cluster length ($p_{adj} < 0.001$), but not in mean cosine similarity ($p_{adj} > 0.05$). Only unique word count was significantly correlated with PANSS scores (rho = –0.68, $p < 0.001$).

Discussion. The automated approximation of clustering proved to be successful, and thus the threshold-based methods, can be gradually introduced as a reliable assessment tool for the VF task.

Confirming previous findings (Bokat & Goldberg 2003; Juhasz et al. 2012), the number of unique words produced on the verbal fluency task was a good predictor for the diagnosis and was strongly negatively correlated with all PANSS subscales. Clustering measures, though not correlated with any PANSS subscale, differed significantly between the groups. This might mean that PANSS misses some relevant TD phenomena, which should be investigated further.

Our results suggest that automated analysis of VF task can be used for assessment of people with schizophrenia and schizoaffective disorder.

References

Bokat C. E., Goldberg T. E. 2003. Letter and category fluency in schizophrenic patients: a meta-analysis. Schizophrenia Research, 64(1), pp. 73–78.

Derogatis L. R., Fitzpatrick M. 2004. The SCL–90-R, the Brief Symptom Inventory (BSI), and the BSI–18 // Handbook of Psychological Assessment in Primary Care Settings. Lawrence Erlbaum Associates Publishers, pp. 297–334.

Elvevåg B., Foltz P. W., Weinberger D. R., Goldberg T. E. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia // Schizophrenia Research, 93(1–3), pp. 304–316.

Hart M., Lewine R. R. 2017. Rethinking thought disorder // Schizophrenia Bulletin, 43(3):514–522.

Juhasz B. J., Chambers D., Shesler L. W., Haber A., Kurtz M. M. 2012. Evaluating lexical characteristics of verbal fluency output in schizophrenia // Psychiatry Research, 200(2–3), pp. 177–183.

Kay S. R., Fiszbein A., Opler L. A. 1987. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia // Schizophrenia Bulletin, 13(2), pp. 261–276.

Keefe R. 2004. The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery // Schizophrenia Research, 68(2–3), pp. 283–297.

Kim N., Kim J.-H., Wolters M. K., MacPherson S. E., Park J. C. 2019. Automatic Scoring of Semantic Fluency // Frontiers in Psychology, 10.

Kutuzov A., Kuzmenko E. 2016. WebVectors: a toolkit for building web interfaces for vector semantic models // International Conference on Analysis of Images, Social Networks and Texts, pp. 155–161.

Robert P. H., Lafont V., Medecin I., Berthet L., Thauby S., Baudu C., Darcourt G. U. Y. 1998. Clustering and switching strategies in verbal fluency tasks: comparison between schizophrenics and healthy adults // Journal of the International Neuropsychological Society, 4(6), pp. 539–546.

Rosenstein M., Foltz P. W., DeLisi L. E., & Elvevåg B. 2015. Language as a biomarker in those at high-risk for psychosis // Schizophrenia Research, 165(2–3), pp. 249–250.

CONSCIOUSNESS IN COMPUTER ARCHITECTURES: THE THEORY OF EXPELLED SCRIPTS[1]

Artemiy Kotov

kotov@harpia.ru

Russian State University for the Humanities, Kurchatov Institute

Consciousness is a phenomenon, experienced by a person subjectively. As a naïve language concept *consciousness* may be characterized by different features, in particular: (a) it provides some space for subjective imaging, including imagination, (b) it distinguishes 'self' from 'non-Self' and simulates *the theory of mind*, (c) it may resolve conflicts and contradictions in representations by considering different options, preferring one and rejecting the others, (d) it grounds voluntary actions as deliberately selected and responsible. Consciousness has a long study history, well summarized by M. Overgaard and J. Mogensen (2017). Following B.J. Baars (1988) consciousness operates as a searchlight, highlighting concepts within the current scope of attention. Another classic view is *the multi-level approach* by D. Rosenthal, where higher-order thoughts operate on the ordinary thoughts, previously unconsciousness. Such level architectures are currently implemented in many theoretical and computer models (Minsky 1968; Sloman, Chrisley 2003). At the same time, the stipulation of several processing levels is not sufficient to capture the nature of consciousness. We start from the assumption, that consciousness is a phenomenon, appearing within some information processing architectures. While the architectures rely on classic causality, they may run not only in the human brain, but also in some artificial processors. Modern approaches to brain modelling assume, that consciousness should appear within the accurate computation model of the entire brain. However, we suggest, that some effects of consciousness may be simulated within a simplified model — *a minimal architecture of consciousness*. We try to simulate these effects withing our computer architecture for companion robots.

The architecture contains a speech processor, central processor and a robot management system, which may output speech and basic nonverbal behavior on a companion robot F-2. The central processor consists of *scripts* — if-then operators or productions. Scripts simulate inferences, speech replies and basic emotions. A script includes two model representations: M^i (initial) and M^f (final). As soon as a semantic representation (frame) of an input text matches the premise M^i, the script invokes and generates the output representation M^f, forcing the robot to perform a speech response or an emotional reaction, attached to M^f. Multiple scripts may invoke at the same time, thus generating a compound behavioral pattern on the robot. The model presently consists of 1600 *r-scripts* for rational processing of input and 79 *d-scripts* for emotional processing. In particular, the utterances *Someone is looking at you* or *a man is always interested in the life of the beloved girl* invoke d-scripts ATTENTION (*Everybody appreciates me!*), PLAN (*They plan something against me!*) and SUBJV (*He always does one and the same, nothing else! He always thinks about one and the same!*). The performance of scripts, tested within processing of 80 mln. wordform text corpus, suggests an important view on the essence of consciousness. We shall concentrate on the ability of scripts to classify an input situation into several classes at the same time.

As shown in Fig. 1a, a simple reactive system compares an input S_1 to the premises (M^i) of all the available scripts. The best script (here — *d-scr1*) is selected for processing, while the alternative scripts (like *d-scr2*) are suppressed. It means, that conceptualization M_1^i, suggested by the best script *d-scr1* will be the only processing route, like, for the example above: *everybody appreciates me!* (ATTENTION). In this case the processing scope of attention ("searchlight") will include, first, the initial model M_1^i of *d-scr1* (see scope A), and then — its final model M_1^f (see scope B).

[1] The present study has been supported by the Russian Science Foundation, project No 19-18-00547.

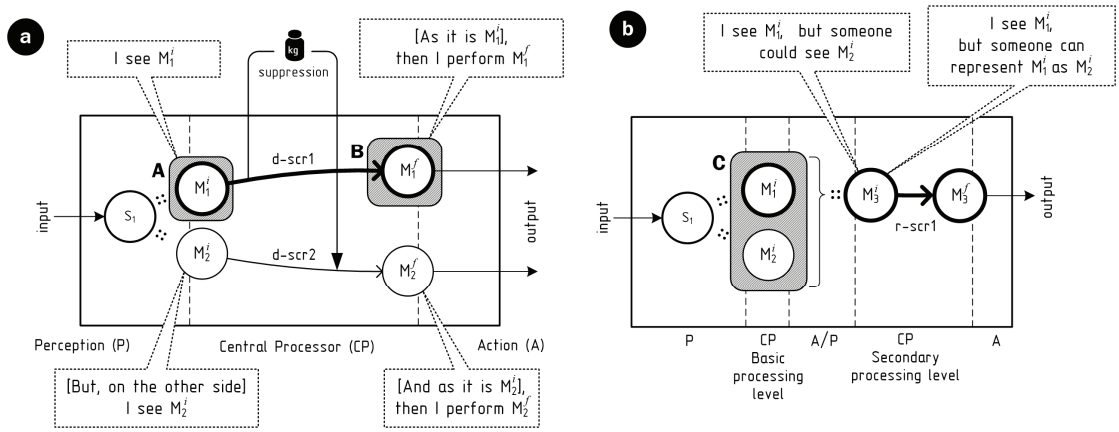


Fig. 1. (a) Simple architecture: the leading script *d-scr1* suppresses other scripts, like *d-scr2*, (b) The minimal architecture of consciousness: script *r-scr1* of the secondary *deliberative* level gains access to the diverse representations M_1^i, M_2^i, constructed by scripts (not shown) on the basic processing level

Within the simple architecture, the alternative script *d-scr2* may not be completely suppressed, but instead delayed in time. The system will process *d-scr1*, represent S_1 as a positive situation (via ATTENTION), then return to *d-scr2*, as to the second activated script, and represent S_1 as a negative situation (via PLAN and SUBJV scripts). In this case the system processes scripts separately from each other and combines their cues within the behavior. Although a system, which selects only one, the most relevant script, could be quite effective, secondary or *expelled* scripts (like *d-scr2*) may be useful to give an alternative view on an incoming situation S_1. Within this architecture we have simulated irony: the robot received an event 'somebody hits you', invoked DANGER script as the most relevant, suppressed its expression and searched for the best positive d-script for an ironic answer. The result could be the ATTENTION with an output *It's good you paid attention at me!* For the simulation of irony, a separate *irony script* has been checking the list of activated scripts (scope C on Fig. 1b), looking for an alternative script in case the relevant script is suppressed. We assume, that within this mechanism a higher-order rational script *r-scr1* (Fig. 1b) should check the most relevant primary script, identify M_1^i as the most relevant ('self') representation of S_1, and use an alternative ('non-Self') representation M_2^i for (a) irony, (b) self/non-Self distinction (M_1^i is 'my view' and M_2^i is 'not my view'), (c) for imagination and the theory of mind ('in another situation I would represent S_1 as $M_2^{i'}$ or 'someone else could represent S_1 as $M_2^{i'}$). Thus, we suggest, that within the model of alternative representations, the expelled scripts may serve as a basis to simulate irony, imagination, theory of mind, conflict handling and other features of consciousness, suggested at the beginning of this note, thus, *representing a minimal architecture of consciousness*.

References

Overgaard M., Mogensen J. 2017. An integrative view on consciousness and introspection // Rev. Philos. Psychol. Vol. 8. No 1, pp. 129–141.

Baars B. J. 1988. A cognitive theory of consciousness.

Rosenthal D. 2005. Consciousness and mind.

Sloman A., Chrisley R. 2003. Virtual Machines and Consciousness // J. Conscious. Stud. Vol. 10. No 4–5, pp. 133–172.

Minsky M. 1968. Matter, Mind and Models // Semantic information processing. Cambridge, Mass.: MIT Press, pp. 425–431.

IMPACT OF THE HUMAN FACTOR ON ERRORS IN COMPUTER-ASSISTED TRANSLATION

Irina Ovchinnikova
ira.ovchi@gmail.com
I. M. Sechenov First Moscow State Medical University

In the last decade, linguistic software has acquired features of the business software products that were worked out for online business communication and joint projects. The linguistic software was developed for translation industry to provide access to translation tools and customer relationship management. Computer-assisted translation (CAT) cloud platforms allow searching a project, performing translation and localization, discussing an output with a customer, and transferring payments. Besides machine translation (MT) systems and shared translation memory (TM), the CAT platforms combine all of the translation tools in the single place. A CAT platform includes dictionaries and thesauruses, concordances and access to an MT algorithm. It is also able to accumulate translation memory based on the translator's individual projects. Since the industry demands fast delivery of accurate translated documents, CAT platforms offer a solution to enhance the translators' productivity. However, the enhancing generates grounds for new errors in translation.

The paper presents the analysis of the errors in CAT that reveals failures in applying the tools and in post-editing the output as well as disadvantages of the platforms design and peculiarities of Computer-mediated Communication (CMC). The errors in translation projects on Smartcat (more than 43 000 word forms of English and Hebrew source documents translated into Russian) were classified according to the harm they brought to target texts. The harm is considered in the aspects of readability, accuracy, semantic coherence, and functionality of the target text. Fluency errors damage the target text readability; accuracy errors misrepresent the content; functional and discourse errors distort the objective of the source text translation and obstruct perception of the target text in the target culture. Translation tools affect the distribution of errors in the final product.

Technological environment forces translators to acquire computer tools and to participate in CMC where English functions as a lingua franca. The English language is involved in the translators' professional communication due to the necessity of the information retrieval while verifying data from a source text since translators are responsible for the accuracy of information in a target text (Byrne 2007). The communication within CAT platforms requires English as well. While working with CAT, translators are forced to code-switching, which involves three languages. Therefore, the translator is exposed to the cross-linguistic influence of two active languages (English and a source language) on a target language. The influence produces the ground for inter-language interference and errors in translation that reflect the influence of English (Zheng et al. 2018). The interference is revealed in the syntactic borrowings from English, errors in capitalization and punctuation that penetrate into the CAT final product. Foremost, the interference evokes fluency errors.

The design of CAT platforms provides fast access to the tools. The working window shows a table with segments of a source text that affects the source text perception by the translator. The segmentation leads the translator to apply the bottom up linear strategy to the source text comprehension that provokes discourse and semantic errors in the target text. The translator erred in expressing semantic coherence among text segments while working on the long text. The errors are recognizable by applying the top bottom and non-linear strategy to the target text editing.

Since the CAT platforms include MT and TM, the translator needs to choose the appropriate strategy: manual translation from the scratch vs selecting suggestions of the algorithm or samples from the shared translation memory. The latter options require revising. Selecting revising of the MT and TM suggestions alongside with the manual translation, the translator has to alternate code-switching and post-editing. Grammar errors in segments with the post-edited TM and MT

output show the attention attrition due to multitasking. Translators make mistakes in long-distance syntactic dependencies, in a chain of clauses in complex and compound sentences, etc. The errors are typical of MT output. While accuracy and fluency errors are equally probable in MT outputs, TM contains less fluency errors as compared to accuracy ones (Barby et al. 1916). The grammar errors on the CAT platform show traces of the insufficient proofreading of the MT output.

Spell-checkers and grammar checkers are able to recognize the errors; automatic evaluation systems can detect errors in syntax and morphology; some of the systems are implemented into CAT platforms. The discourse error of giving different names to the same entity in different segments of the target text (as *ресторан 'Рыба от Иди' / 'Иди фиш' / 'Иди дагим'*) might be obviated by applying the particular tool. Nevertheless, automatic evaluation systems are unable to recognize all of the errors. However, translators neglect the advantages of CAT platforms. Adjusting the design of the platforms to the human perception, developers might improve the human-computer interaction in the industry.

The factors that cause errors in CAT:

(1) Inter-language interference between English and a target language (alongside with the cross-linguistic influence of the source language on the target language)
(2) Source text segmentation in CAT platforms working window
(3) Multitasking on CAT platforms: switching from the text generating to the text evaluation while combining manual translation with post-editing of MT or TM output
(4) Underestimation of the computer tools of the CAT platform by translators.

Thus, the technologies in the translation industry enhance the speed but still unable to improve the quality of translation. The errors on the CAT platform reveal the impact of the human factor.

References

Barbu E., Escartín C. P., Bentivogli L., Negri M., Turchi M., Orasan C. and Federico M. 2016. The First Automatic Translation Memory Cleaning Shared Task // Machine Translation, 30(3–4), pp. 145–166.

Byrne J. 2007. Caveat Translator: Understanding the Legal Consequences of Errors in Professional Translation // Journal of Specialised Translation, 7, pp. 2–24.

Zheng X, Roelofs A., Farquhar J., and Lemhöfer K. 2018. Monitoring of language selection errors in switching: Not all about conflict // PLoS ONE 13(11): e0200397. https://doi.org/10.1371/journal. pone.0200397

ORIGINALITY OF SPEECH PERSPECTIVIZATION IN CINEMATIC ARTS

Natalia Petrova
natalia-yu-petrova@yandex.ru
Moscow State Linguistic University

The main point of this performance is to draw attention to the most recent research in the field of cognitive studies, that is to the linguistic **theory of perspectivization** (Langacker 1999; Talmy 2003) as applied to cinematic arts. This theory sets a new challenge in the "homo sapiens — artificial mind" relationships by means of voice and video interfaces which have already become standard means of communication. This time the focus is not only on machine better understanding of speech patterns but on correct interpretation of inferences.

The process of pespectivization, or **viewpoint,** is often recognized as a necessary prerequisite of human communication and sometimes stands **pars pro toto** for *viewpoint*, *focus of interest*, *foregrounding*, *horizon*, *orientation*, *position*, *relief*, *stance*, *standpoint*, and *style* (Graumann, Kallmeyer 2002; Verhagen 2007). In lexicographic sources perspectivization is regarded as "the cognitive mechanism of construal of an object in discourse from the point of view of the speaker (conceptualiser)" (Petrova, Iriskhanova 2020). Undoubtedly, perspectivization is related to the notion of *perspective* in painting, which in its turn refers to the Latin *perspicere* — «to see through», thus making the term **an interdisciplinary transfer**.

One of the ways to highlight the illocutive force of perspectivization is to follow step-by-step perspective-building in a dialogical unity, which can be interpreted and processed by the machine. Here films or cartoons may serve a productive basis for research as they are largely built on two-party or multi-party speech. The term "cinematic arts" is chosen as it covers not actually filmmaking as **the result** of the process of creating visual art-form, but also **the process** itself. We argue that montage, dubbing, adaptation, etc. are of great value in our understanding of speech perspectivization.

In a film the simplest speech perspective structure, or **monoperspective**, is seemingly a separate utterance of a speaker. Following this logic, the growth of the number of speakers may lead to multiplication of perspectives. But this is not always so. Recent research has shown many-faceted possibilities of **staged intertextuality** (Kotthoff 2002) when a single speaker embraces several voices at a time, some of them being quite contradictory. Such dialogues may multiply the illocutive force of the utterances due to the isomorphic nature of the participants, who are both characters of the film and actors performing their roles. There are many other varieties, e.g. film-in-a-film structure, just like in the opening scene of the comedy "Mrs. Doubtfire" where a talented actor Daniel Hillard, acted by Robin Williams, is involved in the process of voicing an animated cartoon. He synchronizes two voice-over roles presented in opposition — the one of a cat and the one of a parrot:

(1) **Daniel the Cat**: *Salutations, snack.*
Daniel the Parakeet: *Yikes! On second thought... Police! Civic authorities! ASPCA! ASAP! Murder! Betrayal! Kidnapped! No, birdnapped!*
Daniel the Cat: *A cup of garlic, a twist of parakeet. Eat your heart out, Julia Child.*

When the Cat allows the parakeet "to have the last soothing cigarette" before he dies Daniel puts his own words into the bird's mouth to protest.

(2) **Daniel the Cat**: <…> *I'm not wholly without heart. How about a nice soothing cigarette?*
Daniel the Parakeet: *Oh, I will not do this. I cannot! Oh, what a foul way for a bird to die! I don't want to get beak cancer. No! My lungs are blackened!* ("Mrs. Doubtfire", 00:01 — 03:42)

Here Daniel introduces his own opinion (*I* wanted *to* comment *on the situation; Shoving a* cigarette *into Pudgie's mouth is* morally *irresponsible*) and acquires the role of a metacommentator, which doubles his humane perspective and costs him his job. In this case we trace a polyperspective, or cleft perspective. In perspectival terms, the whole scene represents splitting tactics based on the implicit doubling of self by means of staged intertextuality. Even based on contextualization through visual support, it can be hardly detected by artificial intelligence which calculates answers on I-He-She-It deixis.

The skill for doing voices helps Daniel excel at staged intertextuality. During his interview with the court liaison Mrs Sellner he shows 10 different heroes at a time. They are stereotypical characters of a Martian, an immigrant, and a monster. They are also renowned people: actors like Humphrey Bogart or Sean Connery and even the US president Ronald Reagan. Here the effect of polyperspectivity is built on their continuous change in the eye of the viewer.

(3) 1. *We're looking for intelligent life. Oops, we've made a mistake* (as a Martian). 2. *Happy to be in America. Don't ask for a green card* (as an immigrant). 3. *I want you in the worst way* (as a monster). 4. *This is certainly a rough meeting. It's not going very well for me* (as Groucho Mars). 5. *Look at me, Moneypenny. I want to undo that bow and get to know you* (as Sean Connery in the role of James Bond). 6. *I'm crazy to make a deal with you* (as a used-car salesman). 7. *Nancy and I are still looking for the other half of my head* (as Ronald Reagan in "Spitting Image"). 8. *They're doin' it! I'm sittin' on a gold mine* (as Walter Brennan). 9. *Don't make me smack you, sweetheart* (as Humphrey Bogart). 10. *Fiiigarrro!* (as Pudgie the parakeet in "Mrs. Doubtfire") (ibid., 18:35–20:20).

Our short review shows that the subject of speech perspectivization in cinematic arts deserves genuine scientific interest and can be analysed in many spheres of film industry. Artificial intelligence is highly demanded in the matter of split-perspectives and multivocality which are not only problems of interhuman communication but of its social orientation and re-orientation.

Reference

Graumann C. F., Kallmeyer W. 2002. An Introduction // Perspective and Perspectivation in Discourse (Human Cognitive Processing) / Ed. by C. F. Graumann and W. Kallmeyer. Amsterdam; Philadelphia: John Benjamins Pub Co, 2002, pp. 1–11.

Kotthoff H. 2002. Irony, quotation, and other forms of staged intertextuality // Perspective and Perspectivation in Discourse (Human Cognitive Processing) / Ed. by C. F. Graumann W. Kallmeyer. Amsterdam; Philadelphia: John Benjamins Pub Co, pp. 201–229.

Langacker R. W. 1999. Grammar and conceptualization. Berlin; New York: Mouton de Gruyter.

Petrova N. Yu., Iriskhanova O. K. 2020. The Glossary. Viewpoint. scodis.com/?q=ru/perspectivization (16.05.2020).

Talmy L. 2003. Toward a Cognitive Semantics. Vol. 1: Concept Structuring Systems. Cambridge, Mass., London: A Bradford Book, The MIT Press.

Verhagen A. 2007. Construal and perspectivisation // The Oxford Handbook of Cognitive Linguistics (Oxford Handbooks) / Ed. by D. Geeraerts H. Cuyckens. Oxford: Oxford University Press, pp. 48–81.

ARTIFICIAL VS. HUMAN INTELLIGENCE IN DIGITAL VS. HUMAN TRANSLATION

Nadezhda Riabtseva
nadia_riabceva@mail.ru
Institute of Linguistics, Russian Academy of Sciences

Artificial Intelligence (AI) has a great number of various advantages over Human Intelligence (HI), though it is a common presupposition that AI just mimics those corresponding cognitive functions that humans associate with the human mind, such as remembering, learning, problem solving, computing, text searching, "object identification / recognition", etc. Primarily, in early 1950s, the field was founded on the assumption that «human intelligence» can be precisely described so that a computer can be made to simulate it. At present, these two rather complement each other so that AI-based systems, devices and programs, drawing upon computer science, information engineering, mathematics, etc. have become most common help in contemporary life. They are used in information science and practice, industries, transport, robotics, logistics, business, science, art, and many other important activities and, mostly, techniques. In the twenty-first century, AI techniques have experienced a resurgence following concurrent advances in computer power, compiling and using "Big Data" instruments, etc. AI techniques have become an essential part of the information and technology industries as well, assisting in solving many challenging problems in computer science, software engineering, operations research, etc. The most advanced field in this respect is the perspective in applying artificial neural networks to different tasks, particularly those that are connected with formal comparison of data, compilation of big data sets dealing with linguistic activities, etc.

AI is also helpful in a lot of translation tasks: in compiling translation memories, bilingual glossaries of recurrent terms, concordances of specialized collocations, translation technology tools, translation workflow, hybrid translation technologies, terminology identification and rendering procedures, as well as in all kinds of terminology lexicographic work, computer-based linguistic support, and similar computer aids and digital linguistic resources for translators. They all can assist translators in their tasks: oral, written or simultaneous, cf. Computer-Assisted-Interpreting-Training (CAIT) tools, etc. (Goldsmith 2018; Santamaría 2017). And yet, in spite of various, multiple, wide-spread and effective AI applications, there is a number of linguistic activities that are still difficult to computerize. The reason is that AI can simulate human intelligence in some particular linguistic cases, mostly to some extent and mostly rendering help, as AI cannot function like HI and substitute it in understanding natural human language and its usage in the direct sense of the word: because human language uses too much indirect, implicit, analogous, syncretic, supplementary, compressed, etc. linguistic tools to achieve human goals. That is why natural language coherent text processing, including translation, still presents a most difficult task for AI, as here the way to rely on (big data) statistical methods in many cases fails to get adequate results. It concerns translating not only poetry and analogous "esthetic works", but some other creatively charged texts, such as academic papers, which, like many others, demonstrate unsurpassable cross-linguistic asymmetry. They exclude literal translation. It is demonstrated here by the comparison of a digitally translated abstract (by *Google Translator*) to an academic paper (Hanks, Moz'e 2019) from English into Russian, which exposes a lot of failures: terminological, logical, stylistic, etc; their corresponding pairs in both texts (1), (2) are given in *italics*, while preferable (non-literal, adequate, that is, authentic) Russian expressions are given in brackets in ***(bold italics)***:

(1) "Traditionally, dictionaries are *meaning-driven*—that is, they list *different senses* (or *supposed senses*) of *each* word, but do not say much about the *phraseology that distinguishes*

one sense from another. Grammars, on the other hand, are *structure-driven*: they attempt to describe all possible structures of a language, but say little about meaning, phraseology, or collocation. In both disciplines during the 20th century, the practice of inventing evidence rather than discovering it led to *intermittent* and *unpredictable distortions* of fact. *Corpora* now provide empirical evidence on a large *scale* for *lexico-syntactic description*, but *there is still a long way to go*. Many *cherished beliefs* must be abandoned before a synthesis between empirical lexical analysis and grammatical theory can be achieved. In this paper, by empirical analysis of just one word (*the noun 'way'*), we show how *corpus evidence* can be used to *tackle* the *complexities* of lexical and *constructional* meaning, *providing new insights* into the *lexis-grammar interface*."

(2) "Традиционно словари *ориентированы на смысл* (ср. ***описывают значения слова***) — то есть они перечисляют различные *смыслы* (ср. ***значения***) (или *предполагаемые смыслы,* ср. ***возможные значения***) (~~***каждого)***~~ слова, но мало говорят о *фразеологии, которая отличает один смысл от другого* (ср. ***о фразеологических выражениях, которые отличают одно значение слова от другого***). Грамматики, с другой стороны, *управляются структурой* (ср. ***описывают структуру языка***): они пытаются описать все возможные структуры языка, но мало говорят о значении, фразеологии или словосочетании. В обеих дисциплинах в течение 20-го века практика изобретения доказательств вместо их обнаружения приводила к **периодическим* (ср. ***спорадическим***) и **непредсказуемым* (ср. ***неожиданным****) искажениям* фактов. В настоящее время **корпусы предоставляют эмпирические данные в большом *масштабе* (ср. ***корпусные данные*** *предоставляют эмпирический* ***материал*** *в большом* ***объеме***) для *лексико-синтаксического описания* (ср. ***для исследования лексики и грамматики***), *но *впереди еще долгий путь* (ср. ***но в этом отношении еще многое предстоит сделать***). *Многие *заветные убеждения должны быть отброшены* ((ср. ***От многих широко распространенных заблуждений следует избавиться***) прежде, чем будет достигнут синтез между эмпирическим лексическим анализом и грамматической теорией. В этой статье путем эмпирического анализа только одного слова *('*путь' существительного*) (ср. ***существительного 'путь'***) мы показываем, как **корпусные доказательства* (ср. ***корпусные данные***) могут быть использованы для **решения* (ср. ***установления****)* сложных лексических и **конструктивных* (ср. ***конструкционных***) *значений*, обеспечивая **новое понимание* (ср. ***новые принципы понимания***) **лексико-грамматического интерфейса.*" (ср. ***взаимодействия лексики и грамматики***)".

The main problem here is rather not that Russian academic style differs drastically from the English one, but that Russian and English grammars are incompatible (as sometimes terminologies are). That is why digitally translated "academic text" is "quasi-academic", has "quasi-meaning" and causes "pseudo understanding". Though AI is much more operational, fast, mathematically and statistically exact and solid, many-faceted, does not make human mistakes, etc., it works algorithmically, that is why it cannot (still?) generate a creative non-discrete/analogous decision, without which many linguistic tasks cannot be fulfilled. The human mind surpasses AI in quality at least because it can think, recall (what is not fixed anywhere beside his mind), dream, create, make subjective inferences and many other humanly tasks which are not quite algorithmic. Thus, literal or "quasi-literal" translation of academic papers from one natural language to another natural language makes it sound "quasi-natural", distorts its meaning, logics, sense, terminology and objectives, always has a number of incorrect word co-occurrences, and, as a whole, does not sound natural/authentic. It is because natural language has a number of qualities that are difficult to formalize and present in explicit, discrete and definite forms, these qualities are (1) language ability to express meaning implicitly and (2) analogously, (3) indiscreetly, (4) rather indefinitely (in cer-

tain speech positions), (5) (sometimes) creatively, that is “unprecedently”, etc. Generated by human mind, human language is congruent to it and thus is not fully and formally describable and predictable.

References

Goldsmith J. 2018. The Interpreter's Toolkit: Interpreters’ Help — a one-stop shop in the making? // AIIC.net. http://aiic.net/p/8499 [accessed May 2nd 2020].

Hanks P., Može S. 2019. The way to analyse ‘way’ // International Journal of Lexicography, Volume 32, Issue 3, September, pp. 247–269, https://doi.org/10.1093/ijl/ecz005

Santamaría Ciordia L. 2017. Implementing safe and effective collaborative environments in technology enhanced interpreter training // Clina. 3–1, pp. 35–55.

MACHINE LEARNING APPROACH TOWARDS AUTOMATIC TEXT SYNTHESIS IN RUSSIAN

Timur Sadykov
Sadykov.TM@rea.ru

Timur Zhukov
Zhukov.TA@rea.ru

Plekhanov Russian University of Economics

Automatic inflection of words in a natural language is necessary for a variety of theoretical and applied purposes like parsing, topic-to-question generation (Chali, Hasan 2015), speech recognition and synthesis, machine translation, tagset design (Kuzmenko 2017), information retrieval (Iomdin 2003), content analysis (Belonogov 2010), and natural language generation (Cerutti et al. 2017).

Despite substantial recent progress in the field (Korobov 2015), automatic inflection still represents a problem of formidable computational complexity for many natural languages in the world. Most state-of-the-art approaches make use of extensive manually annotated corpora that currently exist for all major languages (Segalovich 2003).

Russian is a highly inflected language whose grammar and rules of inflection are known for their complexity (Sorokin 2016). There exist several other approaches towards automated Russian inflection and synthesis of grammatically correct Russian text, e.g. Korobov 2015. Besides, numerous programs attempt automated inflection of a particular part of speech or synthesis of a document with a rigid predefined structure (Chernikov and Karminsky 2014). Judging by publicly available information, most of such programs make extensive use of manually annotated corpora which might cause failure when the word to be inflected is different enough from the elements in the database.

In the talk to be presented at the conference, we describe a fully algorithmic dictionary-free approach towards automatic inflection and text synthesis in Russian. The developed algorithms have been implemented in $C^{\#}$ programming language. The described functionality is freely available online at www.passare.ru through both manual entry of a word to be inflected and by API access of main functions for dealing with big amounts of data.

The web-service passare.ru offers a variety of functions for the inflection of single Russian words, word matching, and synthesis of grammatically correct Russian text. In particular, the inflection of a Russian noun by number and case, the inflection of a Russian adjective by number, gender, and case, the inflection of a Russian adverb by the degrees of comparison are implemented. Russian verb is the part of speech whose inflection is by far the most complicated in the language. The presented algorithm provides inflection of a Russian verb by tense, person, number, and gender. It also allows one to form the gerunds and the imperative forms of a verb. Besides, functions for forming and inflecting active present and past participles as well as passive past participles are implemented.

The algorithmic coverage of the Russian language provided by the web-service passare.ru aims to balance grammatical accuracy and easiness of use. We use machine learning approach for the verb inflection function as well as for classification of verb endings. In addition, elements of artificial intelligence are used in other functions, such as parts of speech recognition and normalization of the word form.

References

Belonogov G. G., Horoshilov A. A., Horoshilov A. A. 2010. Automation of the English-Russian bilingual phraseological dictionaries based on arrays of bilingual texts (bilingual) // Automatic Documentation and Mathematical Linguistics. 44:3, pp. 103–110.

Cerutti F., Toniolo A., Norman T. J. 2017. On natural language generation of formal argumentation, arXiv: 1706.04033.

Chali Y. and Hasan S. A. 2015. Towards topic-to-question generation // Computational Linguistics, 41:1.

Chernikov B. V., Karminsky A. M. 2014. Specificities of lexicological synthesis of text documents // Procedia Computer Science, 31, pp. 431–439.

Iomdin L. L. 2003. Natural language processing as a source of linguistic knowledge // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, pp. 68–74.

Korobov M. 2015. Morphological analyzer and generator for Russian and Ukrainian languages // Communications in Computer and Information Science 542, pp. 330–342.

Kuzmenko E. A. 2017. Morphological analysis for Russian: Integration and comparison of taggers // Communications in Computer and Information Science 661, pp. 162–171.

Segalovich I. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, pp. 273–280.

Sorokin A. 2016. Using longest common subsequence and character models to predict word forms // Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 54–61.

EVALUATING WORD EMBEDDINGS FOR COMPUTATIONAL PRAGMATICS

Leonardo Sanna
leonardo.sanna@unimore.it
University of Modena and Reggio Emilia

Word embeddings are a popular set of machine learning techniques that are used in natural language processing to create a semantic model of a corpus, representing semantic relationship within a geometric space. In this representation every word is mapped to a vector so that words with similar meanings are represented close to one another while words with different meanings are distant.

In this work we briefly present a novel use of word embedding for computational pragmatics in which we used word2vec Skip-Gram (henceforth SG) (Mikolov et al. 2013) to explore the inferential model generated by algorithmic personalization on Facebook (Hargreaves et al. 2018). This experiment was performed in a recent (yet unpublished[1]) study in which we implemented Eco's model reader (1979) as an inferential model. In a nutshell, we found that the model reader can be simulated with word embedding to create computational inferential models, that are then compared to identify a possible divergence of interpretation paths. Our experiment conducted on Facebook showed that algorithmic personalization has an influence on model readers, differentiating inferences across different timelines.

Though, for pragmatic goals, word embeddings have two major issues that are 1) accuracy 2) variability. The standard way of evaluating accuracy of word embeddings is to use intrinsic evaluation, using predefined analogies and similarities such as WordSim-353 (Finkelstein et al. 2002) and SimLex (Hill et al. 2014). However, this accuracy is quite hard to estimate while inquiring pragmatic aspects as we do not have any predefined pragmatic output.

On the other hand, we have a variability problem, meaning that each training of our SG will generate a slightly different model, as word2vec is trained with a probabilistic neural network. This is of course an issue for reliability and reproducibility of the model. For this problem, one of the most recent solutions was proposed by Hellrich (2019), who experimented a Singular Value Decomposition of a PPMI matrix with weighting-based downsampling to generate reliable word embeddings.

The aim of this work is to review these methods for a computational pragmatic approach, highlighting possible limitations and adaptations. The review will cover three aspects:

1) Reproducibility of computational pragmatics
2) Reliability assessment for computational pragmatics inferences
3) Accuracy test for computational pragmatics goals

References

Eco U. 1979. The Role of the Reader, Bloomington: Indiana UP.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G., & Ruppin E. (2001). Placing search in context: The concept revisited // Proceedings of the 10th international conference on World Wide Web, pp. 406–414.

[1] The study has been prepared for JADT 2020: 15th International Conference on Statistical Analysis of Textual Data. Due to COVID-19 pandemic the conference has been postponed to further date, along with the proceedings. The data used for the experiment are the same data used by Hargreaves et al. (2018).

Hill F., Reichart R., Korhonen A. 2014. SimLex-999: Evaluating semantic models with (Genuine) similarity estimation // Computational Linguistics, 41(4), pp. 665–695.

Hargreaves E., Agosti C., Menasché D., Neglia G., Reiffers-Masson A., Altman E (2018). Biases in the Facebook News Feed: a Case Study on the Italian Elections // International Conference on Advances in Social Networks Analysis and Mining, August 2018, Barcelona.

Hellrich J. 2019. Word embeddings: reliability & semantic change (Vol. 347). IOS Press.

Lai S., Liu K., Xu L., Zhao J. 2016. How to generate a good word embedding // IEEE Intelligent Systems, 31.6, pp. 5–14.

Lenci A. 2008. Distributional semantics in linguistic and cognitive research // Italian journal of linguistics, 20(1), pp. 1–31.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. 2013. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems, December 2013, Lake Tahoe, pp. 3111–3119.

Parisier E. 2011. The Filter Bubble: What the Internet is hiding from you. New York: The Penguin Press.

Rehurek R., Sojka P. 2010. Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Wokshop on New Challenges for NLP Frameworks.

Sahlgren M., Lenci A. 2016. The effects of data size and frequency range on distributional semantic models. arXiv preprint arXiv:1609.08293

Wang B. et al. 2019. Evaluating word embedding models: methods and experimental results // APSIPA Transactions on Signal and Information Processing 8.

TOPIC MODELING AND EXPERT ASSESSMENT AS TOOLS FOR CONCEPTS IDENTIFICATION IN THE THREATCorpus[1]

Vera Zabotkina
zabotkina@rggu.ru

Dmitri Orlov
dm-e-orlov@yandex.ru

Russian State University for the Humanities

Elena Pozdnyakova
helenpozdnyakova@yandex.ru

Moscow State Institute of International Relations (University), Russian State University for the Humanities

The research was carried out using the methods of automated topic modeling based on the THREATCorpus (http://www.threatcorpus.org), which is a domain-oriented corpus of Russian-language texts (10.5 mln words), including three types of discourse (fiction, nonfiction / historiographic discourse, media discourse). Each text entry is hypothesized as a conceptualization of various types of threats in the modern world. Application of topic modeling tools, combined with expert assessment and markup of historiographic and media texts make up the methods of the research. To introduce text parts into the database of the research, elements and relations have been added (Document — Fragment — Thesaurus Concept, see Fig. 1). PLSA, LDA, BigARTM, etc. are used in the algorithmic formation of a thematic profile of document collection. Due to the fact that each document may relate to several topics, topic modeling methods are referred to as “soft clustering” tools.

To conduct the experiment (using the LDA topic modeling algorithm), corpus documents were preliminarily converted from PDF, DOCX, DOC and HTML formats to plain text. Further, the data structures necessary for the operation of the algorithm have been prepared: 1) a dictionary of unique terms reduced to normal form (pymorphy2 library is used for normalization), 2) a term-document matrix.

Topic modeling application results showed a higher occurrence of “hard” threats in news texts, while the “soft” sociocultural threats are more clearly expressed in the texts of historiographic discourse.

Specifically, for historiographic discourse, topics were identified that are characteristic of the terminology: (ANALYSIS, HISTORICAL, MEMORY), (STATE, IDEA, ORDER), (KNOWLEDGE, SOCIAL, HISTORICAL), (HISTORY, EVENT, NORTHERN), (PEOPLE, SOVIET, RUSSIAN). For the news collection (2018–2019), topics such as: (VISA, MODE, RUSSIA), (ROCKET, TEST, C–400), (FIRE, CHILD, CAMP), (PLAGIATE, EUROVISION, CHARGE) are automatically detected; (SEAFRONT, TYUMEN, RADIATION), (TRUMP, VENEZUELA, NATO).

At the same time, thematic modeling must necessarily be accompanied by a meaningful interpretation of the results. Thus, the analysis of news reports related to the topic (TRUMP, VENEZUELA, NATO) showed that this topic is a combination of two independent stories: 1) Donald Trump's participation in the NATO summit (July 2018), 2) US imposition of sanctions on Venezuela (August 2017). This example shows that when conducting thematic modeling, it is necessary to take into account the fundamental possibility of various partitions into topics (clusters).

The second part of the study demonstrates the results of expert markup of text fragments representing sociocultural threats. During the expert assessment, the fragments were tied to the concepts of the “Sociocultural threats” thesaurus (see Fig. 2), developed during the project lifetime. It was shown that historiographic discourse is characterized by a high frequency of the fol-

[1] This study was supported by the Russian Science Foundation, RScF grant 17-78-30029

lowing topics: the confrontational nature of historical politics, the politicization of discussions, wars, the growth of international tension, triggering racial, national and religious hatred, the aggravation of economic contradictions and risks. News discourse is characterized by such topics as: terrorism, growing international tension, the aggravation of economic contradictions and risks, nationalism, cyber aggression. The analysis of the similarities and differences between automated and expert identification of thematic structures and related concepts demonstrates the multiplicity of discursive representations of sociocultural threats.

Further research, based on topic modeling and expert data analysis will be aimed at systemic description of conceptual scope and language representations of sociocultural threats, based on Russian-language discursive practices. Also, it is planned to further adjust the parameters of topic modeling to increase the interpretability of the results, which will allow for automated analysis of the trends in the discourse of sociocultural threats.

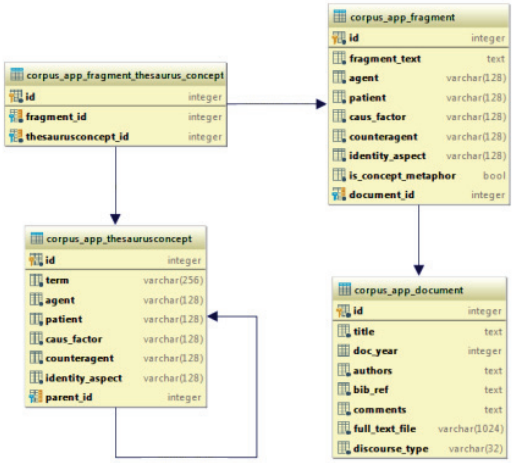


Fig. 1. Database layout for storing expert markup data (fragment)

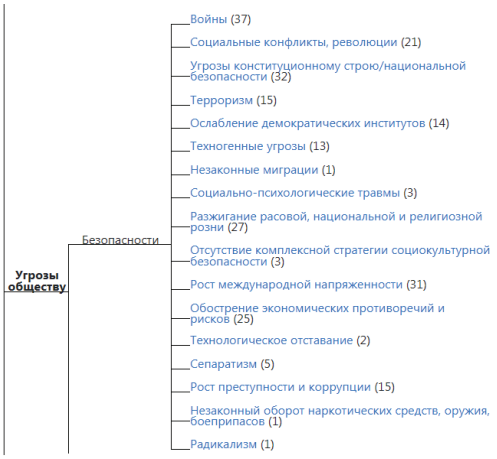


Fig. 2. “Sociocultural challenges” thesaurus (fragment)

References

Pozdnyakova E. M., Zabotkina V. I. 2018. Lingvo-cognitive approach to the study of sociocultural threats // SGEM 2018 Conference proceedings V. 5. № 3.1, pp. 291–298.

Blei David M., Ng A., Jordan M. 2003. Latent Dirichlet allocation // JMLR (3), pp. 993–1022.

Vorontsov K., Potapenko A. 2014. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // International Conference on Analysis of Images, Social Networks and Texts AIST 2014, pp. 29–46.

Daud A., Li J., Zhou L., Muhammad F. 2010. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China, vol. 4. No. 2, pp. 280–301.

SEMANTIC SPACES IN DICTIONARY AND IN THE BRAIN[1]

Liudmila Zaidelman
luda.zaidelman@yandex.ru

Artemiy Kotov
kotov@harpia.ru

Zakhar Nosovets
nosovets@phystech.edu

Russian State University for the Humanities, Kurchatov Institute

Vadim Ushakov
tiuq@yandex.ru
Institute for Advanced Brain Studies, Lomonosov Moscow State University
National Research Nuclear University MEPhI
Mental-Health Clinic No. 1 Named after N.A. Alexeev, Moscow Healthcare Department

Vera Zabotkina
zabotkina@rggu.ru
Russian State University for the Humanities

Boris Velichkovsky
boris.velichkovsky@tu-dresden.de
Russian State University for the Humanities, Kurchatov Institute

This report is a part of a large-scale brain mapping project aimed at finding the correspondence between the semantic categories in a presented oral text and the brain activity, registered in fMRI scanner (Velichkovsky et al. 2020). For the experiment we have prepared 15 texts in Russian. Professional audio recordings of these texts were presented to subjects in fMRI. Our hypothesis is that semantically related words and concepts could be revealed in closely located or even the same brain areas. Here, we describe the implemented methods for semantic markup of texts, based on different understandings of semantic proximity.

A word can designate an object (referent) or a general class (non-referential function). A concept may express or "profile" different features, depending on the situation (Yeh, Barsalou 2006). Furthermore, a word can invoke a visual representation of the object, or construct only "word" representation, if the task can be solved with the help of lexical associations (Barsalou 2003). In (Kobozeva 2004), the subject of the study of in semantics is discussed; whereby the author separates the concepts of *meaning* and *sense*: "The meaning of *X* is information associated with *X* conventionally <…>. The sense of *X* for *Y* in *T* is information associated with *X* in *Y*'s mind during the period *T* <…>". In the search for an ideal method of semantic markup, we appealed to both the *meanings* of words (see dictionary markup) and their *senses* (see vector markup).

Dictionary markup. The source for our first semantic markup was the Russian semantic dictionary (Shvedova 1998). In this approach, all words are divided into parts of speech, and further — into lexical classes (each class is a tree, the top of which is the general word of the whole class), then there are smaller units — lexical sets and subsets (tree branches) and finally — lexico-semantic ranks (tree leaves). Each word receives a representation of its semantics as a set of features that correspond to the word's position in each branch of the ontology — i.e. a word receives a marker for its semantic class and for each of its hypernyms. Depending on the position of the word within the ontological tree of the dictionary, from three to fifteen features were assigned to each word.

[1] Исследование выполнено при поддержке гранта РНФ 17-78-30029.

Such hierarchical representation of the data did not prove to be successful, as it did not allow us to compare words from different subtrees, not to mention the fact that each volume of the Dictionary provided an independent hierarchy, which was not compatible with the hierarchies in other volumes. The distance between two words from different branches would always be infinite since they lack even a single common semantic marker. In this representation words also lacked the situational information, e.g. all the words, relevant to the situation of a *tea party*, or the situation of *threat*.

Vector markup. Representing words as vectors in the space of their contexts is one of the main ideas of distributive semantics. One way to compute vector representations is *word2vec* neural network algorithms (see https://code.google.com/archive/p/word2vec/). For our work, we used a model trained on the Russian-language Wikipedia and texts from the National Russian Language Corpus (Kutuzov, Kuzmenko 2017). In this semantic markup each word is provided with a 300-dimensional vector characterizing its use in the corpus texts. This approach allows us to calculate semantic distances between words and thus construct the semantic space. At the same time, *word2vec* markup (a) does not distinguish homonymous and (b) separate features (vector values) turn out to be uninterpretable, which is unacceptable for the procedure of correlation to the BOLD-signal, registered in fMRI. The second drawback can be eliminated as proposed in (Huth, 2016). In the available vector space, we have calculated the distances between each text word and 997 most frequent Russian words, according to the frequency dictionary (Lyashevskaya, Sharov, 2009.). That is, we characterized each word from stimuli texts as a vector of distances to the words from the list of 997 *feature words*. In this way, text words can still be compared to each other, and each component (value) in the markup vector is interpretable.

Temporal/semantic markup. In addition to semantic markup, each word has a temporal markup — the start and end time of the sound of this word on the audio recording presented to subjects. The BOLD (blood-oxygen-level dependent) signal is monitored for test subjects while they are listening to audio fragments in the fMRI. Recorded BOLD-signal values are then used to judge the activity of the brain in each period. By comparing time data, we can evaluate the relationship between brain activity (based on BOLD data) and the presented features (matrix [features X activity]). Our semantic markup allows us to evaluate the relationship of features and stimulus words ([words X features]). The matrix recalculation procedure implies that features are independent and can be perceived individually, that is why it is important that in the final markup our features are not indivisible 300-dimensional vectors, but 997 independent features. Finally, we can evaluate the relationship between words and registered brain activity ([words X activity] matrix). For brain mapping, the stimulus words from the [words X activity] matrix for seven subjects were divided into 12 clusters. The words within each cluster cause a similar pattern of brain activity. Note that the clusters are composed not only of close words (synonyms, hyponyms), but also of words situationally related to each other. As, for example, in cluster 3: *aggressive, to compel, to call, enemy, explicit, force, hard, obliged, opponent, strong, to threaten, threat*.

Therefore, we suggest that semantic mapping of a lexicon, homeomorphic to the cerebral activity, should rely on the detailed distances between words (as in distributional semantics), but also should be represented as a list of categories (akin in traditional ontologies), in our case — by distances to the major language concepts.

References

Huth A. G. et al. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex // Nature, 532. №. 7600. Pp. 453–458.

Kan I. P., Barsalou L. W., Solomon K. O., Minor J. K., Thompson-Schill S. L. 2003. Role of mental imagery in a property verification task: fMRI evidence for perceptual representations of conceptual knowledge // Cogn. Neuropsychol., vol. 20, no. 2/4/5/6, pp. 525–540.

Kobozeva I. 2004. Linguistic semantics. Moscow: URSS.

Kutuzov A., Kuzmenko E. 2016. WebVectors: a toolkit for building web interfaces for vector semantic models // International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham. Pp. 155–161.

Lyashevskaya O., Sharov S. 2009. Frequency Dictionary of the Modern Russian Language. Moscow: Azbukovnik.

Shvedova N. Yu. 1998. Russian semantic dictionary. M.: Azbukovnik.

Velichkovsky B. M., Zabotkina V. I., Nosovets Z. A., Kotov A. A., Zaidelman L. Ya., Kartashov S. I., Korosteleva A. N., Malakhov D. G., Orlov V. A., Zinina A. A., Goldberg E., Ushakov V. L. 2020. Towards semantic brain mapping methodology based on a multidimensional markup of continuous Russian-language texts. STM. 12(2), pp. 14–25.

Yeh W., Barsalou L. W. 2006. The situated nature of concepts // Am. J. Psychol., vol. 119, no. 3, pp. 349–384.

ПОНИМАНИЕ РЕЧИ ЧЕЛОВЕКОМ: ДВОЙСТВЕННОСТЬ ОРГАНИЗАЦИИ

Н. Ш. Александрова
nina.alexandrova@gmx.net
Sprachbrücke e.V. Berlin

Одной из задач развития искусственного интеллекта является приближение к возможностям человека в решении тех или иных проблем. Несмотря на недоступную человеку скорость обработки информации и нечеловеческую память, искусственный интеллект не может соперничать с человеком во многих областях, в том числе и в распознавании речи. В этой связи актуальными становятся вопросы организации когнитивных функций человека: благодаря чему человек легко понимает речь?

Теоретической опорой работы является концепция *процедурной — декларативной памяти* (Paradis 2004): имплицитная лингвистическая компетенция и эксплицитное метаязыковое знание — два когнитивных механизма, совместное функционирование которых обеспечивает освоение языков и пользование ими в разных условиях и в разном возрасте. Понимание речи — основа вербальной коммуникации и этап освоения как первого, родного языка, так и последующих языков. Билингвизм и языковая патология — те проявления языковой системы, где каждый из когнитивных механизмов с присущей им способностью понимать речь можно наблюдать изолированно.

Билингвизм. Способность освоить язык *натуральным способом* в процессе общения (так осваивается первый родной язык и языки при раннем билингвизме; далее в течение жизни необходимость понимать незнакомый язык также может привести к освоению нового языка натуральным способом) и способность изучить язык *логическим способом* — когнитивные механизмы, которые у здоровых людей при освоении и пользовании языками находятся в динамическом взаимодействии. В то же время каждый механизм может функционировать самостоятельно: приблизительно до трех лет дети осваивают язык только в общении, логический путь изолированно можно наблюдать при изучении мертвых языков. Понимание речи при натуральном пути освоения языка непроизвольно, опережает навыки говорения и проходит путь *от общего к частному*, от прагматики — к лексике и грамматике. Для понимания незнакомого или малознакомого языка в процессе общения характерно схватывание общего смысла высказывания с опорой на ситуацию, мимику, жесты собеседника, часто без опоры на отдельные слова. При изучении языка логическим способом, напротив, понимание речи произвольно, отстает от навыков говорения, строится с опорой на выученные выражения, т. е. проходит путь *от частного к общему*, от лексики и грамматики к прагматике.

При языковой патологии может наблюдаться слабость того или другого когнитивного механизма, что вынуждает пациента опираться на сохранную стратегию. Так, при языковых нарушениях вследствие одномоментных поражений структуры мозга (инсульты, травмы) как у детей, так и у взрослых в большинстве случаев достаточно быстро восстанавливается понимание *общего смысла* высказываний, бытовой речи, но в дальнейшем можно наблюдать те или иные стойкие затруднения понимания *точного смысла* высказываний, н-р, понимания грамматических конструкций. Также в языковой патологии есть два синдрома, при которых, напротив, ведущим симптомом выступает стойкое нарушение понимания *общего смысла* обращенной речи, в том числе простейшей бытовой речи. Это импрессивная (сенсорная) алалия и синдром Ландау-Клеффнера (СЛК). Клиническая картина синдромов во многом схожа, но при импрессивной алалии ребенок не начинает в обычные сроки понимать речь, а при СЛК теряет способность понимать речь после периода нормального развития. Слух и невербальный интеллект при этом не страдают. Прогноз неблагоприятный. Причина стойкости симптоматики кроется, вероятно, в дисфункциональном, т. е. процессуальном характере

этих синдромов: СЛК связан с длительным патологическим изменением электрической активности мозга, этиология и патогенез импрессивной алалии неясны. При анализе симптоматики и течения неосложненной импрессивной алалии (Александрова, Александрова 2016; 2018) становится понятным, что у детей с данным синдромом нарушена способность к непроизвольному пониманию (схватыванию) *общего смысла* обращенной речи, которая является основой освоения родного языка. В то же время произвольная, осознанная способность заучивать отдельные слова и выражения, а также понимать их (важным при этом является «включение» произвольного внимания) остается сохранной. При правильно организованном обучении ребенок может приобрести обширный пассивный и активный словарь, но при этом не понимать общий смысл простейших реплик. Иными словами: при импрессивной алалии ребенок не может освоить язык как родной, но может учить его как иностранный. Известны случаи, когда пациенты, страдающие данным синдромом, осваивали несколько иностранных языков, но не могли пользоваться родным языком легко и свободно, как это свойственно здоровым людям. Обучение языку этих детей лишено опоры на родной язык и совпадает по используемым приемам с обучением робота.

Заключение: данные языковой патологии и билингвизма показывают, что понимание речи человеком обеспечивается совместной работой двух когнитивных механизмов, которые можно охарактеризовать как непроизвольное ситуативное понимание общего смысла обращенной речи без опоры на отдельные слова и понимание произвольное, осознанное, опирающееся на известные слова и требующее включения произвольного внимания. В норме параллельное во времени, но противоположно направленное влияние двух когнитивных механизмов создает эффект континуума.

Распознавание речи в искусственном интеллекте строится в направлении от частного к общему, от акустических сигналов к тексту. Возможно ли искусственное воссоздание противоположной составляющей человеческого понимания речи, в направлении от общего к частному, покажет время.

Литература

Paradis M. 2004. A Neurolinguistic Theory of Bilingualism. Amsterdam; Philadelphia: John Benjamins.

Александрова Н. Ш., Александрова О. А. 2016. Импрессивная (сенсорная) алалия // Журнал неврологии и психиатрии им. С. С. Корсакова. 2016; 11.

Александрова Н. Ш., Александрова О. А. 2018. Импрессивная (сенсорная) алалия: освоение языка как родного невозможно, изучение языка как иностранного доступно // VIII конф. по когн. науке: тезисы докладов, Светлогорск, 2018. С. 38–40.

УМНЫЙ (*SMART*) В СФЕРЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

З. Н. Афинская
МГУ имени М. В. Ломоносова
afin-zn@mail.ru

Новые интегративно-коммуникативные процессы в жизни общества основываются на взаимодействии различных областей знания, в том числе и когнитивной науки. Синтез наук, согласно Е. С. Кубряковой, «может способствовать пониманию сути когниции, а подобный синтез предполагает союз экспериментальной психологии, лингвистики и искусственного интеллекта» (Кубрякова 2012: 25). Междисциплинарные дисциплины находятся постоянно «в поиске новых терминов или в новом концептуальном осмыслении уже существующих терминов и понятий» (Зыкова 2018: 80). Взаимодействие становится одним из основных когнитивных смыслов термина.

Так *умный* (англ. *smart*) в качестве терминологического элемента, обозначая технологии, основанные на использовании искусственного интеллекта, оказался удачным фактом в терминологической картине научно-технического сообщества. Философский анализ новых коннотаций *ума,* связанных с проблемами *компьютеризации* и созданием «искусственного интеллекта», т. е. «системы сложных программ, воспроизводящих на ЭВМ ряд интеллектуальных процедур и операций (логический вывод, экспертные оценки, диагностика, математические вычисления, чтение текста, опознание объектов и т. д.)» (Философский словарь 1986: 493), показал, что все оттенки качества «умное» обладают, несомненно, позитивной коннотацией. Это качество в значительной степени обеспечивает его распространение в составе терминов, обозначающих новые технологические достижения (*умный транспорт, умное освещение, умный город, умные агротехнологии* и многое другое). В отличие от *ума* понятие *интеллект,* подразумевая высшую оценку когнитивной деятельности, предполагает, хотя и в скрытом виде, социальный аспект — работник не каждой профессии оценивается как обладатель интеллектуальных способностей.

Трансформация концептуального содержания терминов в процессе трансфера знаний приводит к расширению когнитивного пространства языка, стимулирует поиск и обретение смысла научного познания мира.

Литература

Зыкова И. В. 2018. К проблеме конвертируемости терминов при построении метаязыка междисциплинарной науки // Образы языка и зигзаги дискурса. Сб. научных статей к 70-летию В. З. Демьянкова / Отв. ред. В. В. Фещенко и др. М.: Культурная революция, с. 80–98.

Кубрякова Е. С. 2012. В поисках сущности языка. Когнитивные исследования. М.: Знак.

Философский словарь, под ред. И. Т. Фролова. 1986. Изд. пятое. М.: Изд-во Политической литературы.

НЕКОТОРЫЕ АСПЕКТЫ НЕВЕРБАЛЬНОЙ КОММУНИКАЦИИ В ПСИХОТЕРАПИИ

Е. А. Бауэр
rsmirr@mail.ru
Медицинский центр доктора Кауша, Кельн

Невербальная коммуникация играет важную роль в психотерапии, являясь важнейшим аспектом во взаимодействии между пациентом и психотерапевтом. Знание и учет невербальной коммуникации необходимы как для психологической и психиатрической диагностики, так и для самого психотерапевтического процесса. Любое взаимодействие между пациентом и психотерапевтом проявляется на невербальном уровне, отражающем приобретенный подсознательный опыт человека. Как отмечал Гарольд Сирлс (Harold Searles), вербальная интерпретация не может исказить ценности невербальной коммуникации (Searles 2008). Иногда, как, например, во время психоаналитических сеансов, клиент и терапевт могут молчать долгое время и это имеет определенные смысл и значение. Анализ невербальных сигналов в самом начале психотерапии помогает при диагностике психического расстройства и в прогнозировании дальнейшего психотерапевтического процесса (Krause 1997).

Невербальная коммуникация не только дает ценную информацию об эмоциональном состоянии пациента, а именно о таких аспектах, как настроение, страх, возбуждение, напряжение, активность и т. д., но и влияет на развитие психотерапевтического взаимодействия (Grawe 2000). Опытный психиатр и психотерапевт основываются в своем диагнозе не только на отдельных симптомах, но и на общем впечатлении. И нередко на основании этого они выдвигают гипотезу по поводу имеющегося заболевания. С другой стороны, чем опытнее специалист, тем больше отдельных аспектов он может вычленить из этого общего впечатления. «Чувство шизофрении» («praecox feeling»), введенное Х. С. Рюмке (Rümke 1942), помогает при постановке диагноза наряду с анализом симптоматики. Это так называемые тончайшие аспекты в поведении пациента, которые улавливает психиатр и психотерапевт, и которые дают им дополнительную информацию. При этом, как правило, способы выражения пациента, особенно мимика, производят впечатление чего-то неестественного и непластичного, что диссонирует с живым общением и взаимодействием (Fuchs 1996).

Пациенты гораздо более довольны терапией, если терапевт может чутко воспринимать их невербальные знаки, а также адекватно выражать свои чувства (DiMatteo et al. 1986). Психотерапевты, которые часто кивают, немного наклоняются к пациенту, и жесты которых показывают живой интерес и участие, показывают лучшие результаты эффективнности терапии, чем те психотерапевты, которые держатся закрыто и дистанцированно (Harrigan, Rosenthal 1986). Значение невербальной коммуникации для терапевтического успеха можно объяснить и на нейробиологическом уровне. Невербальная коммуникация связана с правым полушарием, ответственным за эмоциональные реакции пациента и в конечном итоге за изменения, которые происходят в ходе терапевтического процесса. Чтобы изменить дисфункциональные патерны поведения, необходимо выработать новые, используя при этом все имеющиеся каналы передачи информации. Поэтому очень важно в процессе психотерапии развивать способность улавливать нюансы невербальной коммуникации, дающей подчас больше информации, чем осознанно-вербальный уровень.

Знания и развитие способности к восприятию невербальных сигналов является важной составной частью обучения и работы психотерапевтов в европейских странах. Это также подтверждает наш опыт психотерапевтической работы в Германии, где большое внимание уделяется созданию психотерапевтического взаимодействия в ходе психотерапии (Бауэр, Кабардов 2014).

References

Бауэр Е. А., Кабардов М. К. 2014. Роль невербальной коммуникации в психотерапевтическом процессе // Психология общения и доверия: теория, практика сборник материалов Международной конференции. Под редакцией Т. П. Скрипкиной. С. 549–550.

DiMatteo M. R., Prince L. M., Hays R. 1986. Nonverbal communication in the medical context: The physician-patient-relationship // Blanck P. D., Buck R., Rosenthal R. (eds) Nonverbal communication in the clinical context. University Park and London: Pennsylvania State University Press.

Fuchs T. 1996. Leibliche Kommunikation und ihre Störungen // Z. Klein. Psychol. Psychopath. Psychother. 44, pp. 415–428.

Grawe K. 2000. Psychologische Therapie. 2. Auflage. Hogrefe, Göttingen.

Harrigan J. A., Rosenthal R. 1986. Nonverbal aspects of empathy and rapport in physician-patient interaction // Blanck P. D., Buck R., Rosenthal R. (eds.). Non 18 verbal communication in the clinical context. University Park and London: Pennsylvania State University Press.

Krause R. 1997. Allgemeine psychoanalytische Krankheitslehre. Band 1: Grundlagen.

Searles H. F. 2008. Der psychoanalytische Beitrag zur Schizophrenieforschung**.** Psychosozial-Verlag (Original: Collected Papers on Schizophrenia and Related Subjects 1965).

Rümke H. C. 1942. Kernsymptom der Schizophrenie und „Praecoxgefühl“ // Zentralblatt über die gesamte Neurologie und Psychiatrie, pp. 168–169.

К ПОСТРОЕНИЮ СУБЪЕКТИВНЫХ ТЕЗАУРУСОВ

В. П. Белянин
psyling@gmail.com
Registered Psychotherapist (Торонто)

В художественной литературе отражается не столько мир, сколько интерпретация мира автором. Л. С. Выготский писал: «Везде — в фонетике, в морфологии, в лексике и семантике, даже в ритмике, метрике и музыке — за грамматическими и формальными категориями скрываются психологические» (Выготский 1956: 334). При этом он, однако, полагал, что перейти к «психологии автора ... на основании толкования знаков нельзя» (Выготский 1987: 9).

Однако следы личности автора в тексте можно обнаружить, если применить методы формального направления, наработанные в литературоведении и семиотике (Б. Христиансен, В. В. Виноградов, Ю. Тынянов, Б. Эйхенбаум, В. Жирмунский, Р. Якобсон, В. Пропп и др.). Формальный анализ в сочетании с субъективно-психологическим подходом позволяет выявить в художественных текстах эмоционально-смысловую доминанту, которую мы понимаем как систему когнитивных и эмотивных эталонов, характерных для автора текста как личности и служащих психической основой метафоризации и вербализации картины мира в тексте (Белянин 2008: 49). Применение субъективно-психологического подхода к творчеству позволяет связать картину мира текста с индивидуальными особенностями личности автора.

В материалах «Клинического архива...» (Сегалин 1925–1930) содержатся медицинские свидетельства и генеалогические данные писателей, которые являются основой для постановки вероятностного ретроспективного дистанционного диагноза психопатологического и психиатрического характера при учёте содержательной стороны их речевых произведений. Эвропатологический подход и создание «патографий», с одной стороны, «снимает» «художественность» текста, оставляя это на усмотрение эстетики, с другой стороны, он позволяет поставить анализ текста как продукта речевой деятельности на естественно-научную основу.

На основе этого подхода в художественной литературе выделены ряд типов текстов по эмоционально-смысловой доминанте: «печальные» (основанные на депрессивности), «весёлые» (маниакальность), «тёмные» (дисфория), «светлые» (паранойяльность) и «красивые» (истероидность) (Белянин 1988, 1996, 2000, 2008). Задачей построения типологии являлась не постановка диагноза автору, а лишь определение его субъективного видения мира на основе выявления типологических особенностей его текстов в сопоставлении с другими типами текста.

Был создан шаблон для анализа текстов в форме функционального тезауруса — словаря, где слова представлены не по алфавиту, а в предметно-тематических группах. Объективные тезаурусы (П. Роже, О. С. Баранов) описывают действительность на уровне общезначимых связей и отношений, группируя слова по сходству. Функциональные же тезаурусы, в которых слова группируются по смежности, лучше подходят для описания текстов, где есть образность и впечатления (Гаспаров 1977). Этот же подход используется при определении тональности текста (sentiment analysis), где решается задача определения полярности того или иного текста на основании выраженного в нём мнения — положительного, негативного или нейтрального.

Этот шаблон — ПсихоТезаурус — содержит тематические группы, которые организованы на шести уровнях. Первый уровень — это тип текста («печальный», «веселый», «тёмный», «светлый», «красивый»). Второй уровень — это самые общие категории: человек, качества, эмоции, отношения, деятельность, природа, мир. На третьем уровне, к примеру, в категории «эмоции» имеются следующие рубрики: положительные, отрицательные, нейтральные.

Дальнейшее дробление выглядит так: отрицательные — печаль, гнев, отвращение, страх; положительные — радость — веселье, улыбка, удовольствие и др.

Каждая рубрика наполнена лексическими элементами, которые организованы в семантические поля (совокупность слов, покрывающих определённую область значений), которые объединены исследователем на основании их совместной встречаемости в том или ином типе текстов. Тезаурусы пересекаются, но лишь частично: рубрики для разных типов текстов могут быть разными. По сути дела, субъективные тезаурусы описывают разные семантические пространства как системы категорий индивидуального сознания, при помощи которых происходит оценка и классификация различных объектов. На момент написания тезисов ПсихоТезаурус содержит порядка 20 000 русских и 10 000 английских слов и коллокаций; он постоянно пополняется.

Представляется, что ПсихоТезаурус как метод анализа текста может быть улучшен благодаря применению машинного обучения, когда будут учитываться те параметры текстов, которые не были обнаружены при «ручном» анализе. Предполагается, что анализ текстов разных жанров с помощью ПсихоТезауруса позволит определять актуальное или возможное психическое состояние автора, что может быть использовано не только для речевой идентификации автора, но и, допустим, для целей определения его работоспособности или предотвращения неблагоприятного развития у него ментальных проблем.

References

Белянин В. П. Введение в психиатрическое литературоведение. München: Otto Sagner, 1996.

Белянин В. П. Основы психолингвистической диагностики: (Модели мира в литературе). М.: Тривола, 2000.

Белянин В. П. Психолингвистические аспекты художественного текста. М.: МГУ, 1988.

Белянин В. П. Психологическое литературоведение. М.: Генезис, 2006.

Выготский Л. С. Мышление и речь.М., 1956.

Выготский Л. С. Психология искусства. М., 1987.

Гаспаров М. Л. Художественный мир М. Кузмина: тезаурус формальный и тезаурус функциональный // Избранные статьи. М., 1977.

Сегалин Г. В. Клинический архив гениальности и одарённости. Свердловск: 1925–1930.

ВОСПРИЯТИЕ НАЗВАНИЯ ПОЛИКОДОВОГО ТЕКСТА В АСПЕКТЕ ПЕРЕВОДА

Л. Б. Бойко

boyko14@gmail.com

Балтийский федеральный университет им. И. Канта

Неукротимое и стремительное расширение диапазона средств коммуникации сделало особенно актуальным современное понятие мультимодальности (поликодовости), получившее развитие в научных изысканиях последних лет. Одно из важных наблюдений свидетельствует о том, что информационные потоки не только устремляются по множественным каналам, но и распределяется сложным образом; при этом вклад, который они вносят в процесс создания нового знания и в процесс коммуникации, не одинаков (Кибрик 2010, Молчанова 2014). Особый интерес с этой точки зрения представляет собой арт-дискурс, в создании которого участвуют разнообразные формы передачи знаний. Восприятие произведения искусства, так же как и его создание, представляет собой креативный процесс, связанный с образным кодированием информации, однако этот процесс не ограничен участием двоих — создателя произведения и адресата, и допускает вовлеченность третьих лиц в концептуализации объекта. Поскольку современные музеи стремятся сделать свои коллекции доступными для широкой публики, они сопровождают произведения искусства интерпретирующими вербальными текстами, как устными, так и письменными, пытаясь добавить «нечто большее» (Blunden 2020) к той информации, которую зритель получил по невербальным перцептивным каналам. Процесс словесного описания любых невербальных систем — экфразис — подразумевает и деятельность, и событие; его конечной целью является оказание дополнительного воздействия на получателя информации (Schneck 1999). Несомненно, необходимо изучать такие поликодовые формы в их семиотической целостности (Молчанова 2015; Денисова, Петухова 2018).

Вербальное сопровождение произведения искусства не только дает в пользование реципиенту еще один информационный канал, но и создает полифонический эффект при восприятии целостного дискурса, при этом автором вербальной составляющей может быть как сам творец (например, художник, дающий название картине), так и другой участник создания арт-дискурса. В рассмотрении взаимосвязи языка и искусства на первый план выходит многообразие ролей, которые играет язык в создании современного произведения искусства (Фещенко 2016) — зачастую вербальный информационный поток оказывается главным в создании новой когнитивной структуры. В любом случае совмещение комплекса визуальных и вербальных знаков делают их практически неразделимыми, и в целостном арт-дискурсе можно наблюдать наложение, напоминающее палимпсест.

Сложность значений и отношений, связанных с взаимодействием словесного текста, предмета искусства и адресата раскрывается в новом ракурсе в процессе переводе, обнажающем когнитивную природу категориальных несоответствий между символическими формами культуры и вербальными средствами их репрезентации. В ракурсе межкодовых взаимодействий предметом рассмотрения, в частности, могут послужить названия произведений искусства и их перевод. Название картины рассматривается здесь как знак и минитекст, обладающий специфическим набором функций, с одной стороны; с другой стороны, оно предстает как особый вид имени собственного. Благодаря гетерогенной природе задействованных семиотических кодов (визуальный + языковой) восприятие и интерпретация (а следовательно, и перевод) названий требуют иных стратегий, чем, например, при раскрытии смыслов во взаимодействии заглавия и текста внутри одной семиотической системы. Название и корпус визуального текста (собственно объект искусства) выступают как два взаимодействующих текста, создающих новое знание. В искусстве мы имеем дело с вер-

бальным кодом заголовка и визуальным кодом произведения искусства, что кардинального меняет последовательность восприятия информации по сравнению с монокодовым дискурсом. Если разумным алгоритмом прочтения монокодового текста будет «название — текст — (название)» то в случае с произведением изобразительного искусства, которое по природе своей есть «objet-pour-la-vue» (Bosredon 1997), зритель начнет с изображения, обращаясь к названию только для того, чтобы получить подтверждение своим предположениям. (Bosredon 2002). Переводчик имеет дело с вербальным компонентом арт-дискурса и, более того, зачастую лишен возможности ознакомиться с визуальной составляющей, что разрушает цельность арт-дискурса и лишает переводчика возможности успешно концептуализировать содержание и принять решение о вербализации. В докладе будет показано, как направление, в котором идет восприятие поликодовой информации, радикально сказывается на выборе переводческих стратегий и тактик.

Литература

Денисова Н. В. Петухова Т. И. 2018. Вербальный код в поликодовом пространстве произведения живописи: когнитивно-коммуникативный и переводческий аспекты // Когнитивные исследования языка. № 35. С. 86–93.

Кибрик А. А. 2010. Мультимодальная лингвистика // Когнитивные исследования IV. М.: ИП РАН, С. 134–152.

Молчанова Г. Г. 2014. Когнитивная невербалика как поликодовое средство межкультурной коммуникации: кинесика. // Вестн. Моск. ун-та. Сер. 19. Лингвистика и межкультурная коммуникация. 2014. No 2. С.13–30.

Молчанова Г. Г. 2015. Синергия визуального и вербального в хроматике (живописи) как семиотический код коммуникации //Вестник МГУ Сер. 19. Лингвистика и межкультурная коммуникация. 2015. №2.

Фещенко В. В. 2016. Образы и роли языка в визуальном искусстве // Критика и семиотика. 2016. No 2. С. 67–82.

Blunden J. 2020. Adding ‘something more’ to looking: the interaction of artefact, verbiage and visitor in museum exhibitions. Visual Communication, 19(1), pp. 45–71. https://doi.org/10.1177/1470357217741938

Bosredon B. 1997. Les de titres de tableaux. Une pragmatique de l’identification. Paris: Presses universitaires de France.

Bosredon B. 2002. Les titres de Magritte: surprise et convenance discursive // Linx: Revue des linguistes de l’Université Paris Ouest Nanterre la Defence. 47. https://doi.org/10.4000/linx.577.

Schneck E.-P. 1999. Pictorial desires and textual anxieties: Modes of ekphrastic discourse in nineteen-century American culture // World & Image, vol. 15 /1, pp. 54–62.

МОДЕЛЬ ДЛЯ ОБРАБОТКИ НЕДИСКРЕТНЫХ ДАННЫХ В ЛИНГВИСТИЧЕСКОМ ОПИСАНИИ

Е. Г. Борисова

egbor@mail.ru

Московский государственный педагогический университет

Задачи компьютерной лингвистики в течение многих лет заставляли стремиться к дискретности описания языка, которое надо было перевести на язык логики, а в дальнейшем и дискретной математики. До появления компьютерной лингвистики эта задача тоже стояла, т.к. средства выражения имеют преимущественно дискретный характер: буквы, фонемы, ЛСВ, однако она преодолевалась благодаря тому, что методология допускала нечеткость описания. Исследователи постоянно сталкивались с расплывчатостью многих единиц, отсутствием границ между ними. Это очень заметно при описании ЛСВ (значений) слов, когда пришлось вводить новое понятие «оттенок» — промежуточное между двумя значениями. В еще большей степени это касается значений грамматических категорий, когда для глагольного вида пришлось вводить не только частные значения, но и «трудные случаи».

Нечеткость заметна при установлении приемлемости того или иного предложения, словосочетания или словоформы: разные информанты могут давать различные, иногда прямо противоположные заключения. Это, в свою очередь, связано с различиями в подъязыках, которые тоже представляют собой нарушение дискретности: отдельные лексемы и даже грамматические явления (например, причастия на *-вши*) в чем-то есть, в чем-то отсутствуют. Степень (не)приемлемости тоже оказывается недискретной, что проявляется в способе ее маркировки. Микродиахрония знаменует собой еще один аспект нарушения дискретности: языковая система «размывается» за счет сосуществования двух несколько различающихся грамматических систем.

Необходимость учета недискретных явлений в языке в явном виде отмечалась Э. Кассирером, А. Е. Кибриком (Кибрик 1995), в наши дни — А. А. Кибриком. Для описания таких явлений привлекались математические модели с использованием понятия размытые (fuzzy) множества. Другим инструментом может служить понятие прототипа: система «ядер» образует дискретную систему, а их периферии допускают взаимопересечения, размытость. Есть возможность также описать их с опорой на правила функционирования языка, учитывающие прагматические принципы, постулаты Грайса, в первую очередь, принцип релевантности.

Если учесть, что действия говорящего включают выбор между конкурирующими лексемами и граммемами, то логично предположить, что эти действия определяются основной функцией коммуникации: стремлением наиболее адекватно передать информацию. А адекватность передачи, в свою очередь, определяется постулатами речевого общения, предписывающими говорящему быть максимально понятным и релевантным. Задача выбора средств выражения осложняется неконгруэнтностью континуальной реальности и дискретных языковых единиц. Ни одно лексическое значение не совпадает с отражаемой реальностью на 100%. Поэтому модель говорения должна предусматривать возможность сравнения нескольких средств выражения для одной ситуации. Следует обратиться к представлению (смыслу) ситуации, соответствующему намерению говорящего отразить его определенные стороны. Окажется, что для восстановления замысла приходится обращаться к импликатурам, позволяющим слушающему дополнить содержание сообщения до передаваемого смысла. Например, *Он был первый россиянин, ступивший на эту землю* предполагает, что человек приехал на чем-то, потом оказался на этом месте, находился там еще какое-то время, т. е. что этот имплицитный смысл при необходимости может быть восстановлен. Причем

импликатуры могут иметь разную вероятность в зависимости от представлений слушающего (Имплицитность 1999).

Выбор подходящего средства выражения заключается в сравнении двух или больше кандидатов, имеющих близкие значения, с выражаемым замыслом (с учетом предполагаемых импликатур). Выбирается та единица, от которой до замысла ведут более вероятные и короткие импликатуры. Например, для описания замысла «я был в институте» при указании времени начала появления обычно выбирается глагол «войти»: *Я вошел в институт ровно в девять*. Глагол *прийти* тоже возможен: из его значения «начать находиться» следует вывод: «пересекалась граница пространства». Но глагол *войти*, чье значение «пересечь границу пространства по направлению внутрь», подойдет больше, потому что в контексте, где содержится указание на время пересечения, этот глагол в явном виде передает требуемый смысл, а *пришел* требует импликатуры. Впрочем, она очень вероятна, поэтому и пример с *пришел* вполне может быть допустим. Но если разница в выводах, их вероятности, более заметна, то остается только один допустимый вариант. В частности, ?*я дошел до института ровно в восемь*» сомнительно, а **ступил в институт ровно в восемь* и вовсе невозможно, хотя действие, описываемое глаголом *ступить*, при этом совершалось, однако «нахождение в институте» для этой фразы является смыслом, передаваемым имплицитно и с не очень большой вероятностью (Борисова 2006).

Введение в моделирование речевой деятельности процедуры выбора с опорой на имплицитные смыслы позволяет отразить возможную недискретность в оценках допустимости примера: импликатуры восстанавливаются участниками общения исходя из различных представлений о реальности, и оценка приемлемости оказывается разной. Эта же процедура объясняет и наличие «оттенков» между значениями: говорящий, опираясь на зафиксированные в языке лексические значения слов (ЛСВ), рассчитывает на определенные выводы. Если эти выводы повторяются, они тоже начинают закрепляться в лексической системе. Так, слова *серебро* и *золото* в значении «посуда из данных металлов» понятны при выводе смысла с учетом контекста. Для слова *серебро* оно частично закрепилось и отражено в словаре в виде оттенка, однако вряд ли оно будет активно использоваться в речи: фраза *Я подарил им на свадьбу серебро* не будет понятна. А *золото* и вовсе не узуализовалось в этом значении, возможно, из-за редкости ситуации.

В целом введение процедуры выбора отражает противоречие между недискретностью описываемой ситуации и дискретностью (хотя бы желательной) используемых средств.

Литература

Борисова Е. Г. 2006. Интерактивный подход в лингвистике: пределы применимости. Компьютерная лингвистика и интеллектуальные технологии // Труды Международной конференции «Диалог». С. 84–88.

Имплицитность в языке и речи. 1999. Кол. монография под ред. Е. Г. Борисовой и Ю. С. Мартемьянова. М.: ЯРК.

Кибрик А. Е. 1995. Куда идет современная лингвистика // Лингвистика на исходе XX века. М.: Русская филология. С. 33–52.

МЕТОД КОРПУСНОГО ИССЛЕДОВАНИЯ ДИСТРИБУЦИИ И СЕМАНТИКИ НЕОЛОГИЗМОВ

В. В. Бочкарев
vbochkarev@mail.ru

А. В. Шевлякова
anna_ling@mail.ru

Казанский федеральный университет

Появление больших корпусов текстов на рубеже XX–XXI вв. открыло большие перспективы для неологии. Например, корпусное исследование неологизмов, в частности возможности их образования с помощью префиксоидов, рассматривается в работе (Захаров 2016). Кроме этого, корпуса текстов являются ценным языковым ресурсом для лексикографов. Большое количество новых слов, вошедших в русский язык на рубеже XX–XXI вв. описывается, например, в словаре Е. А. Гришиной, О. Н. Ляшевской, созданном на основе НКРЯ (http://dict.ruslang.ru/gram.php).

Создание сверхбольших корпусов диктует разработку и использование новых методов исследований языка. В данной работе предложен метод, позволяющий автоматически выявлять, каким образом неологизм встраивается в систему языковых отношений.

Анализ дистрибуции и семантики слов строится на основе векторной репрезентации высокой размерности. Слову ставится в соответствие вектор частот всех возможных его сочетаний с 20-ю тысячами наиболее частотных (за период 1920–2009 гг.) слов русского языка. Учитывая, что возможны словосочетания вида C+X и X+C, где C — исследуемое слово, а X — одно из списка наиболее частотных слов, размерность векторного представления составила 40 тысяч. Информация о частотах тех или иных биграмм извлекалась из русского подкорпуса Google Books Ngram (Michel et al. 2011). Могут быть использованы также представления, использующие частоты синтаксических биграмм, которые тоже могут быть извлечены из корпуса Google Books Ngram. Синтаксические биграммы — базовые единицы синтаксической структуры, включающие два слова — главное и зависимое, которые не обязательно стоят рядом друг с другом в предложении (Sidorov et al. 2012). Используя данное представление, оценка степени семантического сходства/различия слов выполняется с помощью различных векторных метрик: дивергенции Йенсена-Шеннона, расстояния Зергеля (Soergel), манхеттенского расстояние и некоторых других.

Предлагаемая схема анализа для неологизмов выглядит следующим образом. На первом этапе, используя описанное выше векторное представление слов, выбирается ограниченный круг слов, наиболее сходных с исследуемым словом по дистрибуции. Далее для каждого слова из отобранного списка определяются наборы биграмм (либо синтаксических биграмм) вида C+X и X+C, включающих это слово. После этого находим: а) слово, для которого список сочетающихся с ним слов в наибольшей степени перекрывается с аналогичным списком для исследуемого слова; б) слово, для которого наибольший процент сочетающихся с ним слов сочетается также и с исследуемым словом. Таким образом, мы находим слова, которые стали широко использоваться раньше исследуемого слова и имеющие наиболее сходную сочетаемость.

Для тестирования предлагаемого подхода было проанализировано 21 слово, для которых известно время появления и которые могли считаться неологизмами в момент первого появления в корпусе.

Рассмотрим пример со словом *самолет*, для которого на первом этапе исследования был выделен список семантически близких слов (использовалась дивергенция Йенсена-Шеннона). В это список попали слова, обозначающие транспортные средства (поезд, аэроплан, вертолет и другие). Далее из полученного списка отбирались слова, сочетаемость которых наиболее полно совпадала с сочетаемостью слова *самолет* (анализ проводился по

биграммам и по синтаксическим биграммам). Слово *самолет* входит в 1362 биграммы, представленные в Google Books Ngram. Словом, для которого список биграмм имеет наибольшую степень совпадения со списком биграмм слова *самолет*, является *корабль*. Слово *корабль* входит в 1495 биграмм, в 962 из которых (70.6%) встречается также слово *самолет*. Это слово появилось значительно раньше слова *самолет*, то есть, согласно дистрибуции, оно служит определенным прообразом слова *самолет*, так как последнее имеет ряд сходных сем и синтаксических связей. Близкое по смыслу слово *судно* входит в 1099 биграмм, из которых 744 общих со словом самолет. Слова *корабль*, *судно* являются в некотором смысле родовыми для слова самолет (см. выражения *воздушное судно*, *командир воздушного корабля*), и от них *самолет* в значительной степени унаследовал дистрибутивные возможности. Однако, наследование произошло только в 70,6% случаев в силу специфики описываемого новым словом объекта. В отобранный список также вошло слово *аэроплан*, которое было заимствовано и вошло в широкое употребление немного раньше слова *самолет*. Слово *аэроплан* встречается в составе 107 биграмм, из которых 105 случаев (или 98.1%) общие со словом *самолет*. Таким образом, мы автоматически выявляем также близкий синоним и непосредственного предшественника исследуемого слова. Отметим еще слова *вертолет* (входит в 263 биграммы, из которых 258 общих с *самолет*) и *бомбардировщик* (входит в 102 биграммы, из которых 98 общих), которые вошли в употребление позже слова *самолет*, унаследовав от последнего ряд значений и связей.

С помощью предложенного метода можно путем компьютерной обработки определять, в какую семантическую группу входит новое слово и выявлять его синонимы, что может найти применение в лексикографии. Кроме того, подобный метод может использоваться в исследованиях эволюции языка и для выявления неочевидных семантических связей между словами. Более подробно примеры будут представлены в докладе.

References

Захаров В. П. 2016. Словообразовательные неологизмы в русском языке (корпусное исследование) // Т. Н. Буцева (ред.). Неология и неография: современное состояние и перспективы (к 50-летию научного направления): Сборник научных статей. СПб.: Нестор-История, с. 57–63.

Michel J.-B., Shen Y. K., Aiden A. P., Veres A., Gray M. K. et al. 2011. Quantitative analysis of culture using millions of digitized books // Science, 331, pp. 176–182.

Sidorov G., Velasquez F., Stamatatos E., Gelbukh A. et al. 2012. Syntactic dependency-based N-grams as classification features // LNAI, vol. 7630, pp. 1–11.

МИКРОДИАХРОНИЧЕСКОЕ ИССЛЕДОВАНИЕ ТЕРМИНОВ *БУРЫЙ* И *КОРИЧНЕВЫЙ* НА МАТЕРИАЛЕ КОРПУСА GOOGLE BOOKS NGRAM

В. В. Бочкарев
vbochkarev@mail.ru

А. В. Шевлякова
anna_ling@mail.ru

Казанский федеральный университет

Г. В. Парамей
parameg@hope.ac.uk

Ливерпуль Хоуп Университет

Е. В. Рахилина
rakhilina@gmail.com

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

Национальный исследовательский университет «Высшая школа экономики»

Бурное развитие компьютерных технологий в конце XX века способствовало созданию больших корпусов текстов, которые в настоящее время представляют обширные эмпирические данные и инструментарий для изучения лингвистических феноменов. Для проверки гипотез с использованием корпусных данных разрабатываются и применяются квантитативные методы. Эти методы играют особенно важную роль при работе с диахроническими корпусами текстов для изучения языковой вариативности и эволюции языка в разные периоды времени (Hilpert, Gries 2016). Целью данной работы является микродиахроническое исследование цветовых терминов *бурый* и *коричневый* на материале русскоязычного подкорпуса Google Books Ngram (Michel et al. 2011). Лексика цветообозначений уже довольно долго и широко изучается в отечественной и зарубежной лингвистике, в том числе с исторической и эволюционной точек зрения (Бахилина 1975; Biggam 2012). Однако детальный корпусный анализ функционирования конкурирующих названий цветов в диахронии на относительно малых промежутках времени не проводился.

В задачи настоящего исследования входило проверить компьютерными методами гипотезу о соотношении близких по значению цветовых терминов *бурый* и *коричневый*, представленную в (Rakhilina 2007). Суть гипотезы заключается в том, что появившееся в языке новое цветообозначение (в нашем случае *коричневый*) проходит определенные этапы адаптации к уже имеющейся системе прежде, чем войти в число базовых. Сначала оно сочетается с ограниченным кругом артефактов, который со временем расширяется, а затем и с со словами, обозначающими природные объекты, тем самым составляя конкуренцию исторически более раннему базовому термину цвета, схожему по значению (в нашем случае *бурый*). В свою очередь, сочетаемость более раннего цветового термина все больше сокращается, хотя он и употребляется с рядом природных объектов, — прежде всего в устойчивых словосочетаниях.

Для того чтобы количественно проверить эту гипотезу и определить, происходит ли вытеснение более старого термина (*бурый*) более новым (*коричневый*), из корпуса Google Books Ngram (GBN) сначала автоматически извлекались, а затем анализировались соответствующие атрибутивные конструкции. Корпус GBN содержит тексты, наиболее ранние из которых датируются 1607 годом. Несмотря на это, фактически изучаемый период составляет около 200 лет: данных для статистического анализа в корпусе становится достаточно лишь с начала XIX века.

Для анализа изменений в семантике цветовых терминов был использован недавно предложенный метод обнаружения новых значений слов (Bochkarev et al. 2020). Данный метод основан на векторной репрезентации высокой размерности и позволяет получить набор

коллокаций, характеризующих употребление слова в новых значениях, а также датировать время появления новых значений. В ходе анализа было выявлено, что для *бурого* и *коричневого* в корпусе встречаются биграммы с 796-ю существительными, из которых 133 образуют сочетания только с *бурым*, 404 — только с *коричневым*, а 259 имен сочетаются с обоими цветами. Количественный анализ показывает, что для существительных, употребляющихся и с *бурым*, и с *коричневым*, в среднем наблюдается рост доли случаев употребления с термином *коричневый*. Средняя по 259 исследованным существительным доля употребления вариантов с *коричневым* увеличилась от приблизительно 30% во второй половине XIX — начале XX века до 70% в начале XXI века. Было выявлено две волны, когда происходила смена *бурого* на *коричневый*: 1) с начала 20-х годов XX века; 2) с начала 50-х.

Если рассматривать полученные списки биграмм с качественной точки зрения, то можно заметить следующее. Только с *бурым* употребляются преимущественно названия природных объектов (более 95%): названия животных, растений, ископаемых, особенностей ландшафта и др. Только с *коричневым* сочетаются артефакты и малая доля (около 3%) природных объектов. В конкурирующих парах также наблюдается переход от *бурого* к *коричневому*: становятся *коричневыми* изначально бурые природные объекты, как *коршун, жук* или *камень*.

Таким образом, в целом выдвинутая гипотеза подтверждается, однако наше исследование показывает, что смена цветовых терминов протекает по более сложным правилам, «дробящим» контексты употребления природных и искусственных объектов на более мелкие группы, которым свойственна разная скорость изменений. В докладе будет приведен подробный анализ типов таких контекстов.

Работа выполнена при поддержке Российского научного фонда, грант № 20-18-00206.

Литература

Бахилина Н. Б. 1975. История цветообозначений в русском языке. М.: Наука.

Biggam C. P. 2012. The Semantics of Colour: A Historical Approach. Cambridge, UK: Cambridge University Press.

Bochkarev V., Shevlyakova A., Solovyev V. A. 2020. Method of semantic change detection using diachronic corpora data // W. van der Aalst et al. (eds.). Analysis of Images, Social Networks and Texts. AIST 2019. Communications in Computer and Information Science, Vol. 1086, pp. 94–106).

Hilpert M., Gries S. 2016. Quantitative approaches to diachronic corpus linguistics // M. Kytö, P. Pahta (eds.). The Cambridge Handbook of English Historical Linguistics. Cambridge, UK: Cambridge University Press, pp. 36–53.

Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K. et al. Quantitative analysis of culture using millions of digitized books. Science, 331, 176-182, 2011.

Rakhilina E. 2007. Linguistic construal of colors: The case of Russian // R. E. MacLaury, G. V. Paramei, D. Dedrick (eds.). Anthropology of color: Interdisciplinary multilevel modeling. Amsterdam/Philadelphia: John Benjamins, pp. 363–377.

ДИНАМИКА КОНЦЕПТУАЛЬНОЙ СТРУКТУРЫ АКСИОЛОГИЧЕСКОЙ КАТЕГОРИИ[1]

Е. Л. Боярская

depti313@gmail.com, eboyarskaya@kantiana.ru

Балтийский федеральный университет им. И. Канта, РГГУ

Е. В. Шевченко

eliza_veta@mail.ru, eshevchenko@kantiana.ru

И. В. Томашевская

itomashevskaya@kantiana.ru

Балтийский федеральный университет им. И. Канта

Структура аксиологической категории *истина vs ложь (truth vs lie)* исследуется с позиций кросс-дисциплинарности: в рамках классической и когнитивной лингвистики, философии и теории информации. Как известно, аксиологические категории используются для характеристики абсолютной или сравнительной ценности объекта. Существует ряд дихотомий, которые по своим базовым признакам прототипа могут быть причислены к основным аксиологическим категориям: например, *добро vs зло*, *хорошо vs плохо*, *правда vs ложь*. Настоящее исследование посвящено анализу структуры концептов, входящих в аксиологическую дихотомию *истина vs ложь* (*truth vs lie)*, на основе принципов динамической концептуальной семантики, основным постулатом которой является понятие процесса, в том числе процесса приращения и обмена информацией (Scott 2016, Stanford Encyclopedia of Philosophy 2016). Современные информационные технологии расширили мировую информационную среду, именно поэтому применение принципов динамической концептуальной семантики открывает новые перспективы анализа феноменов языка. Постоянный приток концептуальной информации извне приводит к изменениям концептуальной картины мира, структуры концептуальных категорий, характеристик их прототипов, а также отдельно взятых концептов (Болдырев 2018, Заботкина 2017), именно эти процессы наглядно демонстрирует динамика концептуальной дихотомии *истина vs ложь (truth vs lie)*.

Как показало проведенное исследование, актуализация данной концептуальной дихотомии обусловлена гносеологическими, эпистемиологическими (соответствие действительности) и аксиологическими факторами (Боярская, Шевченко, Томашевская 2020). В рамках настоящего исследования было осуществлено моделирование структуры концептов *истина — ложь* (*truth — lie*), которое потребовало анализа максимально развернутых лексикографических дефиниций, вербализующих данные концепты (WordNet3.1; CoBuild English Language Dictionary). На основании анализа дефиниций представляется возможным осуществить моделирование структуры фрейма TRUTH (ИСТИНА), которая выглядит следующим образом: 1) факт(ы); 2) степень верификации факта(ов); 3) цель предоставления факта(ов); 4) манера предоставления факта(ов); 5) пост-эффекты факта(ов); 6) контрагенсы предоставления факта(ов). Структура фрейма LIE (ЛОЖЬ) по перечню слотов совпадает с предыдущей, однако концептуальное наполнение слотов частично отличается: 1) факт(ы); 2) степень верификации — нет; 3) цель предоставления факта (ов) — манипуляция, достижение негативного эффекта; 4) манера предоставления факта (ов); 5) пост-эффекты факта(ов) — отрицательные; 6) контрагенсы — агенсы, ставящие целью донесение истины и препятствие манипуляции (в случае с фреймом TRUTH (ИСТИНА) — цель — манипуляция) (Боярская, Шевченко, Томашевская 2020). В английском языке лексикализованный концепт *truth* вербализуется одной лексической единицей, в то время как в русском языке существуют, по крайней мере, две лексические единицы, фиксирующие максимально схожий объем концептуальной информации, — *истина* и *правда*.

[1] Данное исследование подготовлено в рамках проекта РНФ № 17-78-30029

Как показал анализ материала, почерпнутого из корпусов английского и русского языков (British National Corpus, iWeb, NOW и др.), в структуре прототипа Истина\Truth появляются новые признаки, которые под действием концептуальной информации, поступающей извне, меняют свой иерархический статус с периферийного на более центральный, при этом наследуя ряд характеристик прототипа, что ведет к их выделению в отдельный концепт. Так, в современном английском языке в структуре концептуальной категории lie первоначально выделился концепт *fake news*, который с течением времени под действием активных информационных процессов приобрел новые концептуальные характеристики. Намеренная фильтрация информационного контента привела к расширению объема концептуальной информации, кодируемой данной лексической единицей, что, в конечном итоге, привело к формированию и закреплению поразительных с аксиологической точки зрения концептов: *alternative truth*, *alternative facts*, *half-truth* и *post-truth*. Данные концепты, по нашему мнению, принадлежат к пограничной концептуальной зоне континуума *truth vs lie* и формируют зону бленда. В результате размытия границ концептуальных категорий и появления новых концептов появились и альтернативные точки референции (альтернативная правда, полуправда).

Таким образом, проведенное исследование подтверждает целесообразность применения принципов динамической концептуальной семантики в ходе анализа характера изменений концептуальной картины мира, структуры концептуальных категорий и статуса отдельных концептов. Аксиологическая категория *истина vs ложь (truth vs lie)* демонстрирует тенденцию подвижности характеристик своих прототипов, что приводит к изменению их статуса в составе прототипа исходной категории и дальнейшему формированию нового концепта *post-truth*, который может стать прототипом новой субординантной категории, инкорпорирующей несколько точек альтернативной аксиологической референции.

References

Болдырев Н. Н. 2018. Язык и система знаний. Когнитивная теория языка. М.: ЯСК.

Боярская Е. Л., Шевченко Е. В., Томашевская И. В. forthcoming. Изменения концептуальной структуры аксиологической категории: эволюция или деволюция// CAICS 2020 (National Congress on Cognitive Research, Artificial Intelligence and Neuroinformatics).

Заботкина В. И. 2017. Репрезентация событий: интегрированный подход с позиций когнитивных наук // Коллективная монография / Языки славянской культуры. М.

Stanford Encyclopedia of Philosophy 2016. Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2016/entries/dynamic-semantics .

Scott M. 2016. Supplemental Update, Semantics and Pragmatics, 9(5). doi:10.3765/sp.9.5

Michel J.-B., Shen Y. K., Aiden A. P., Veres A., Gray M. K. et al. 2011. Quantitative analysis of culture using millions of digitized books // Science, 331, pp. 176–182.

ЭВИДЕНЦИАЛЬНЫЕ МАРКЕРЫ В КИНЕТИЧЕСКИХ ПАТТЕРНАХ ГОВОРЯЩЕГО

Е. В. Буденная
jane.sdrv@gmail.com
НИУ ВШЭ / ИЯз РАН

Ю. В. Николаева
julianikk@gmail.com
МГУ им. Ломоносова

А. А. Евдокимова
arochka@gmail.com
ИЯз РАН

Данная работа посвящена исследованию эвиденциальных явлений в рамках мультимодальной коммуникации. Категории эвиденциальности, указывающая на источник сведений говорящего о сообщаемой ситуации, посвящена обширная литература (см., например, Chafe, Nickols 1986; Aikhenvald 2004, русский обзор см. Плунгян 2011: 340 и след., применительно к дискурсивному уровню см. Макарцев 2014), однако в невербальных каналах эвиденциальность практически не изучалась.

Традиционно считается, что в русском языке эвиденциальность грамматически не выражена. Однако ряд явлений позволяет говорить об эвиденциальном компоненте в кинетических каналах говорящего — прежде всего, в движениях рук и головы.

Материалом исследования послужил корпус RUPEX («Рассказы и разговоры о грушах», www.multidiscourse.ru). В контексте данного исследования интерес представляли мануальные и цефалические паттерны рассказчика (N), посмотревшего «Фильм о грушах» (Chafe 1980), в сопоставлении с пересказчиком (R), который данного фильма не видел.

Методы разметки жестов рук базировались на правилах, описанных в Литвиненко и др. 2017. Для указательных и изобразительных жестов рук была размечена точка зрения (viewpoint): жесты персонажа vs. жесты наблюдателя (McNeill 1992). Так, если рассказчик показывал, как мальчик берет корзину, при этом обеими руками держа воображаемую корзину перед собой, это был жест персонажа (C–VPT); когда рассказчик показывал, что мальчик и девочка едут на велосипедах навстречу друг другу, при этом правая и левая рука двигались навстречу, изображая героев истории, это был жест наблюдателя (O-VPT).

Методы разметки жестов головы основывались на принципах, обозначенных в Kibrik, Fedorova 2020. С точки зрения эвиденциального компонента рассматривались изобразительные, указательные, а также регуляторные жесты, задействованные в переключении зон коммуникации (Гришина 2017: 129). Для всех трех типов жестов размечалась точка зрения. У изобразительных жестов она маркировалась аналогично жестам рук; другие типы жестов оценивались в контексте изменения зоны видимости (ibid.): если она менялась, жест относился к O-VPT, если же нет, в фокусе внимания оказывалось взаимоотношение с жестами других каналов. В том случае, когда голова синхронизировалась ритмически, семантически и по направлению с жестами рук, соответствующий жест маркировался той же точкой зрения, что и параллельный ручной жест, в остальных случаях выбирался дефолтный тип O-VPT.

Для пилотного анализа была выполнена мануальная разметка для записей #04 и #22 и цефалическая разметка записи #22 (длительность каждой из записей около 25 минут). Обе записи составляют часть так называемого эталонного корпуса (Kibrik, Fedorova 2020 in print).

Результаты анализа отразили следующие закономерности:

1. Для мануального канала:

- Рассказчик значимо чаще использовал иллюстративные жесты персонажа, по сравнению с пересказчиком: 82 из 164 у N, vs. 74 из 214 у R (χ-квадрат, p-value=0.003).
- Внутри каждого из 9 содержательных дискурсивных эпизодов рассказчик использовал C–VPT-жесты на большем количестве последовательных ЭДЕ, по сравнению с пересказчиком (t-критерий Стьюдента и далее χ-квадрат, p-value<0.01).
- Рассказчик использовал цепочки последовательных кореферентных ЭДЕ, покрытых C–VPT-жестами, для большего числа персонажей, по сравнению с пересказчиком

(6 персонажей у N против 4 персонажей у R в #04, 4 персонажа у N против 2 персонажей у R в #22).

2. Для цефалического канала (запись #22):

- Дистрибуция C–VPT-жестов головы у рассказчика и пересказчика отражает ту же тенденцию, которая ранее была обнаружена в мануальном канале: N значимо чаще использовал C–VPT-жесты, по сравнению с R (37 из 234 у N vs. 4 из 179 у R, χ-квадрат, p-value<0.001). При этом, как и в руках, N чаще совершает изобразительные жесты, по сравнению с R (45 из 391 у N против 10 из 484 у R, χ-квадрат, p-value<0.001). Данный факт объясняется тем, что рассказчик, видевший фильм, сильнее вовлечен в излагаемый сюжет, по сравнению с пересказчиком, который фильма не видел и лишь соединяет факты, рассказанные ему ранее двумя другими участниками.
- Пересказчик существенно чаще использовал регуляторные жесты, по сравнению с рассказчиком (112 из 484 у R, 23 из 391 у N, χ-квадрат, p-value<0.001). Данный факт связан с положением пересказчика по отношению к слушателю и тем, что он, пересказывая, проверяет свой рассказ, поворачиваясь к рассказчику и комментатору, чтобы посмотреть на их невербальную реакцию.

Литература

Гришина Е. А. 2017. Русская жестикуляция с лингвистической точки зрения. М.: ЯСК.

Кибрик А. А. 2018. Русский мультиканальный дискурс. Часть II. Разработка корпуса и направления исследований // Психологический журнал 39 (2). С. 78–89.

Литвиненко А. О., Николаева Ю. В., Кибрик А. А. 2017. Аннотирование мультиканального дискурса: мануальные жесты // Е. В. Печенкова М. В. Фаликман (ред.) Когнитивная наука в Москве: новые исследования. Материалы конференции 15 июня 2017 г. М.: БукиВеди, ИППиП. С. 516–520.

Макарцев М. М. 2014. Эвиденциальность в пространстве балканского текста. М.; СПб.: Нестор-История.

Плунгян В. А. 2011. В. А. Плунгян. Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира. М.

Aikhenvald A. Y. 2004. Evidentiality. Oxford: Oxford University Press.

Chafe W. (ed.) 1980. The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Ablex, Norwood.

Chafe W., Nichols J. (eds.) 1986. Evidentiality: The linguistic coding of epistemology, Ablex, Norwood, NJ.

Kibrik A. A., Fedorova O. V. (eds.) 2020. The MCD handbook: A practical guide to annotating multichannel discourse, Moscow (in print).

McNeill D. 1992. Hand and mind: What gestures reveal about thought. University of Chicago Press.

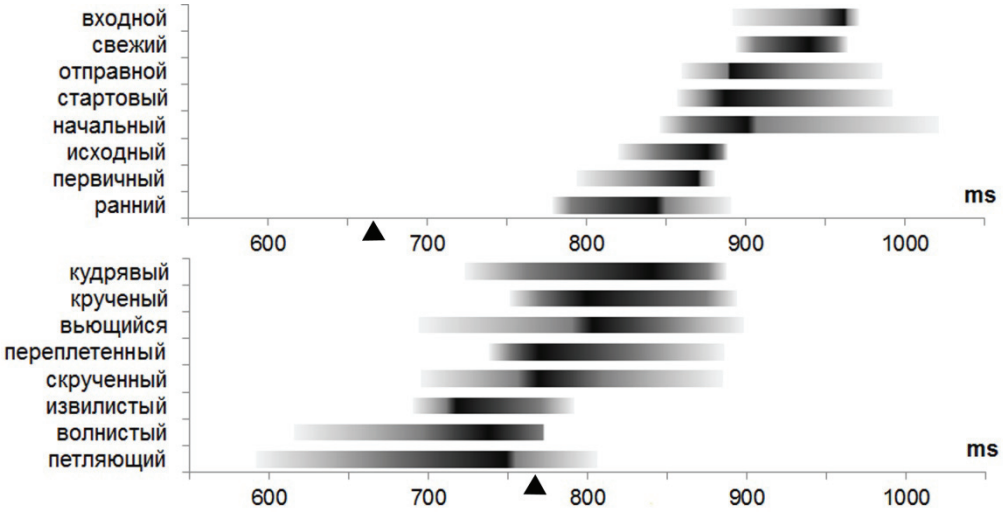
СТРУКТУРА ТЕЗАУРУСА РУССКИХ ПРИЛАГАТЕЛЬНЫХ, ОПРЕДЕЛЕННАЯ ПО ИЗМЕРЕНИЯМ ВРЕМЕНИ РАСПОЗНАВАНИЯ СЛОВ НА СЛУХ

В. Л. Введенский
victorlvo@yandex.ru

К. Г. Гуртовой
gurtoko@hotmail.com

НИЦ «Курчатовский институт»

Мы продолжаем исследование процесса восприятия услышанных слов, описанное в работе Vvedensky et al. 2020. В эксперименте испытуемый нажимал клавишу клавиатуры компьютера, как только узнавал услышанное слово, и повторял его вслух для контроля правильности распознавания. Компьютер произносил слова в случайном порядке в темпе реальной речи. Для эксперимента брались группы из 8 прилагательных, близких по смыслу. Мы исследовали около 20 групп семантически близких прилагательных, содержащих от 20 до 60 слов каждая. Это в значительной мере покрывает множество эмоционально нейтральных прилагательных, имеющихся в русском языке (Ляшевская, Шаров, 2020).



В сотнях экспериментов мы неизменно получаем результаты, показанные на рисунке. Это данные двух разных слушателей. Полосы отражают распределение времени распознавания конкретного слова, произнесенного пять раз. Время отложено от начала звучания слова, а стрелка показывает момент окончания звучания. Длительность всех слов в группе одинакова, несмотря на значительный разброс в количестве используемых для написания букв. Испытуемые сильно различаются по скорости распознавания — одни нажимают клавишу за 100 мс до окончания звучания, другие запаздывают до 800 мс. Различие между «самыми быстрыми» и «самыми медленными» словами также сильно варьирует — между 200 и 600 мс. Несмотря на такую нестабильность, неизменно наблюдается надежное упорядочение слов в группах, предъявляемых слушателям. Для любой группы слов мы видим, что испытуемые неосознанно упорядочивают их по скорости восприятия.

Мы полагаем, что в основе этого явления лежит характер представления слов в коре головного мозга человека. Визуализация процессов в мозге методом фМРТ показала, что реакции на слова распределены по всей поверхности обоих полушарий (Huth et al. 2016). При этом семантически близкие слова компактно группируются в определенных местах. Следует вскрыть внутреннюю структуру этих групп слов, объединенных по смыслу. В наших экспериментах постоянно наблюдается процесс типа сканирования, с помощью которого слова извлекаются из памяти в коре головного мозга. Измерения на поверхности коры, проводимые

при нейрохирургических вмешательствах, неизменно показывают присутствие в коре бегущих волн возбуждения нейронов (Martinet et al. 2017). Такие бегущие волны наблюдаются и с помощью магнитной энцефалографии МЭГ (Gurtovoy et al. 1989). Вполне возможно, что именно эти волны участвуют в механизме распознавания услышанных слов. Роль бегущих волн в обработке информации корой мозга детально обсуждена в статье (Muller 2018).

Общий результат наших измерений указывает на то, что слова определенного типа образуют структуру на поверхности коры, подобную ряду параллельных дюн, где обычные слова лежат в долинах, а «перевалы» контролируются многозначными словами, связывающими различные семантические группы. Эксперимент показывает, где находятся «точки входа» для волн, считывающих информацию, то есть какие слова распознаются быстрее всего. Как правило, это не самые распространенные слова, которые находятся на удалении от входа. Для картирования на коре распределения слов, установленного по времени распознавания, лучше всего подходит МЭГ (Vvedensky, Nikolayeva 2016).

Литература

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. http://dict.ruslang.ru/freq.php

Gurtovoy K. G., Kajola M. J., Vvedensky V. L. 1989. Distribution of phase for alpha-rhythm measured by multichannel magnetometer over occipital head // Williamson S. J., Hoke M., Stroink G., Kotani M. (eds.). Advances in Biomagnetism. Boston: Springer, pp. 249–252.

Huth A. G., de Heer W. A., Griffiths T. L., Theunissen F. E., Gallant J. L. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex // Nature, 532. Pp. 453–458.

Martinet L.-E., Fiddyment G., Madsen J. R., Eskandar E. N., Truccolo W., Eden U. T., Cash S. S., Kramer M. A. 2017. Human seizures couple across spatial scales through travelling wave dynamics // Nature Communications, vol. 8, article number 14896.

Muller L., Chavane F., Reynolds J., Sejnowski T. J. 2018. Cortical travelling waves: mechanisms and computational principles // Nature Reviews Neuroscience 19 (5). Pp 255–268. doi: 10.1038/nrn.2018.20 2018.

Vvedensky V. L., Nikolayeva A. Yu. 2016. Detailed structure of the cortical magnetic response to words // International Symposium on Neural Networks: Advances in Neural Networks, pp. 691–697.

Vvedensky V., Gurtovoy K., Sokolov M., Matveev M. Ordering of words by the spoken word recognition time // Advances in Neural Computation, Machine Learning, and Cognitive Research III. Studies in Computational Intelligence, vol 856, Springer, 2020, pp. 151–156.,

СЕМИОТИЧЕСКОЕ (ВОЗ)ДЕЙСТВИЕ И ВОЗМОЖНОСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: УСЛОВНЫЙ «ТЕСТ ТЬЮРИНГА» КАК ТЕСТИРОВАНИЕ ИСПЫТАТЕЛЯ

А. В. Вдовиченко

an1vdo@mail.ru

Институт языкознания РАН

Понятие *семиотического, или коммуникативного, (воз)действия*, сформулированное в рамках коммуникативной модели (Вдовиченко 2008), вносит существенные коррективы в понимание *знаковых процессов*. В докладе понятие семиотического (воз)действия проецируется на отношения искусственного интеллекта и испытателя (наблюдателя), проводящего условный «тест Тьюринга» для оценки возможностей и свойств искусственного интеллекта.

Парадокс отношений ИИ и наблюдателя заключается в том, что, с одной стороны, искусственный интеллект создан естественным интеллектом, с полным пониманием — со стороны последнего — механизмов функционирования первого: при решении вводных задач любой сложности в ИИ работают *заранее заложенные* в него *алгоритмы,* включая случаи более сложных генетических алгоритмов, используемых в т.н. самообучающихся нейронных сетях (Столяр, Владыкин 2011). С другой стороны, несмотря на неавтономность ИИ, его очевидную зависимость от вложенных в него алгоритмов, тем не менее, зачастую становится актуальным («парадоксообразующим») вопрос, *способна ли машина* (ИИ) *мыслить*, как это делает обладатель естественного интеллекта (ср. Soares 2015).

Вопрос о способности ИИ мыслить, как человек, фактически предполагает следующую формулировку: *является ли алгоритмическим само человеческое мышление?* Мыслит ли алгоритмами сам создатель алгоритмов? Если это так, то мышление человека и мышление машины принципиально едины. И, наоборот, если естественное мышление не алгоритмично, то ИИ и ЕИ — принципиально различны, и в чем, в таком случае, следует *видеть специфику человеческого мышления*, на фоне заведомо алгоритмического машинного мышления? Этот вопрос имеет и этическую (философскую) составляющую: является ли человеческое мышление столь же механистичным и заданным, как машинное?

В докладе приводится *краткий эксперимент-наблюдение*, позволяющий наглядно представить специфику мышления человека (эксперимент «Одно слово»). Но прежде приводятся аргументы в пользу возможности знакового (лингвистического) подхода к разрешению этого вопроса.

Сама, казалось бы, парадоксальная возможность сравнивать ИИ и ЕИ возникает, во-первых, вследствие недостаточной ясности того, *что* представляет собой мышление (его, например, можно редуцировать до способности констатировать имеющиеся автономные объекты, свойства, признаки, и совершать над ними «объективные» счетно-логические операции). А во-вторых, вследствие того, что, ввиду скрытости механизмов естественного мышления, *о нем можно судить только по внешним (в том числе знаковым) признакам — поведению* (действиям, проявлениям, результатам) обладателя мышления. Именно здесь актуализуется *знаковый процесс,* производимый как ЕИ, так и ИИ, предполагающий интерпретацию знака в различных теоретических моделях, и, соответственно, открывается возможность рассматривать *лингвистические знаки* при решении вопроса о соотношении ИИ и ЕИ.

Оставляя нейробиологам содержимое закрытого «черного ящика» сознания (синапсы и пр.), обратимся к *знаковому поведению* обладателей ЕИ и ИИ. На основании производимых знаков (в данном случае вербальных) машину на какой-то стадии знаковой деятельности можно признать *успешно прошедшей* условный тест Тьюринга (вернее, одну из известных ныне многочисленных вариаций этого теста). Для наблюдателя это может быть даже личный вариант теста Тьюринга.

Например, когда водитель по безупречно вежливой подсказке навигатора [поверните направо] поворачивает направо, он признает, что поведение искусственного интеллекта разумно, ему можно в данном случае довериться и исполнить его инструкцию. Самим фактом поворачивания направо обладатель ЕИ признает способность ИИ мыслить адекватно и правильно, как это делает человек. Машина, таким образом, справилась с тестом Тьюринга: она говорит разумно, показывает разумно, разумно взаимодействует с ЕИ. Наблюдатель получает основания признать, что на данном ограниченном отрезке когнитивно-коммуникативной деятельности знак [поверните] для машины и для наблюдателя означает одно и то же, знак [направо] — одно и то же, и т. д. Значит, в этом простом случае интеллекты — искусственный и естественный — вполне подобны? Значит, зная про алгоритмическое устроение искусственного интеллекта, можно заключить, что естественный интеллект тоже алгоритмический в своем устроении (по меньшей мере, при анализе и интерпретации данной знаковой ситуации, от которой недалеко и до всех остальных)? Значит, машина может мыслить? Значит, человек мыслит так же, как машина?

Интерпретация данного общения ИИ и ЕИ, вероятно, несколько смущающая своей простотой, все же имеет под собой некоторые основания. Дело в том, что *попытки навязать единообразный знаковый язык сознанию обладателя ЕИ*, встречающиеся в некоторых версиях когнитивных и других исследований, *по сути представляют собой внедрение в человеческое сознание языкового алгоритма* (формулы, правила, модули, единицы, единообразие, стандартные последовательности и пр.; напр., об этом: Бабкин, Козырев, Куркина 2006: 8–10). Если признавать, что язык *напрямую* связан с когнитивными процессами (что тоже бывает в некоторых версиях лингвистической теории, напр., Соссюр 1977: 144; Сепир (1921) 1993: 36, 193; Хомский 1972; Jackendoff 2007; Chafe 2013 et al.) и др., то ситуация усугубляется: *мышление (сознание) становится неизбежно алгоритмичным*, без всякой надежды на то, чтобы отличаться от машины.

Чтобы элиминировать это странное заключение, нужно признать, что оно возникает из некорректного классического понимания знака как модуля «форма-значение»: судить о сознании — значит, судить по знаковому поведению; по знаковому поведению, значит, судить по знакам; знаки — это носители смысла (значения); то, что в сознании, — передается знаками; значит, тот, кто пользуется знаками правильно — мыслит правильно и имеет человеческое сознание. В результате машина, обученная говорить словами время от времени уместно и правильно, становится обладателем сознания, идентичного сознанию человека.

Модульность знака (смыслоформальное единство «форма-значение») отрицается в *коммуникативной модели*. Тем самым преодолевается цепь рассуждения, приводящего к признанию тождества ЕИ и ИИ, возникающего на основании знаковой алгоритмизации.

Необходимым условием смыслопорождения в любой знаковой структуре постулируется не единообразный знак, а *акциональный* («воздейственный») *режим сознания* семиотического актора, или мыслимое им опосредованное (воз)действие: кто полагается в фокусе его сознания как адресат привносимых изменений, какие объекты (действия, признаки, связи, ценности, эмоции и пр.) формируются в сознании актора, какие обстоятельства (наблюдаемые, вспоминаемые) сопутствуют (воз)действию, какие «прогнозы» когнитивных состояний адресата в каждый отдельный момент последовательности производимых воздействий присутствуют в фокусе внимания семиотического актора, какие каналы и знаковые клише признаются доступными и эффективными для данного воздействия и пр. (Так, понять языковое высказывание [Он прилетает в два по местному] невозможно: непонятно, кто или что, где, когда, какое «два» и пр. Зато вполне возможно понять то же самое как элемент семиотического действия конкретного актора — вернее, понять *режим сознания актора*, который включил в структуру своего действия данные клише [он прилетает в два по местному]. Понимаются, таким образом, не слова языка, а режим сознания конкретного актора в параметрированной ситуации) (ср. Вдовиченко, Тарасов 2017).

Поскольку никакой знак, в том числе вербальный, не обладает собственной референцией, *понимать автономную знаковую (в том числе вербальную) структуру невозможно* ввиду отсутствия конкретного акционального режима сознания актора во «всеобщем (вербальном) знаке и языке». Место «означающего» в семиозисе может занимать только обладатель активного сознания, совершающий попытку произвести изменения в ином когнитивном состоянии опосредованными способами (что верно для устной и письменной формы вербального процесса как интегрируемой части «воздейственного» семиотического процесса) (Вдовиченко 2016). Указание (референция) в естественном семиотическом действии производится *на режимы сознания актора*, что зачастую некорректно интерпретируется как «независимая» референция на объекты (связи, признаки, действия и пр.). Любое «означаемое» (*ср.* стоический *лектон*) создается семиотическим актором в *процедуре назначения актуальных множеств*, в ходе которой выделяются адресаты, объекты, единства, связи, способности реципиента «вместить» производимое опосредованное (знаковое) воздействие, и пр. (Вдовиченко 2020). Режим конкретного сознания в семиотической процедуре эксплицируется намеренно и целенаправленно комплексом средств, в том числе среди прочих и вербальными клише, в параметрированной ситуации воздействия, в отличие от незнакового (внутреннего) когнитивного процесса, в котором отсутствует акциональность (мыслимая «воздейственность» на постороннее сознание) и, соответственно, отсутствует знаковость.

С этой точки зрения, любой условно выделяемый знак (слово, сочетание слов, предложение, жест, остенсив, демонстративное поведение и пр.) интерпретируется как *намек на режим конкретного сознания семиотического актора* и не может восприниматься как модуль «форма-значение»: для получения «означаемого» необходимо воссоздать акциональный режим сознания семиотического актора, на которое условно выделяемые знаки *только намекают* (в угадывании *намека* и состоит процедура интерпретации знаковой структуры) (Вдовиченко 2018).

Таким образом, «мысли» машины, или искусственного интеллекта, — это, в самом деле, *алгоритмы*, или исполняемые команды (созданные инструкции), которые сами представляют собой знаки-модули, и на выходе выдают столь же модульные образования (слова, фразы, показания и пр.), стоящие в конце цепочки инструкций как результат их выполнения. Машина не намекает на внутренние состояния, не интерпретирует знаки, она работает знаками-модулями. По степени неопосредованности знак машины и его «понимание» полностью подобны механистическим процессам физического мира, совершающимся без участия «когнитивной призмы» (*ср*. удар по неподвижному мячу > движение мяча в заданном направлении).

В отличие от машины, в естественном интеллекте знаковость играет *совсем иную роль*. *Вербальные знаки*, претендующие на роль *алгоритма* в естественном сознании, *не составляют сути мышления* (сознания, интеллекта; *ср.* глухонемые люди, и др. примеры). Сознание свободно от вербальной (знаковой) составляющей, испытывая в ней необходимость *только при осуществлении попыток опосредованных коммуникативных воздействий* на представимое сознание для *намекания на собственные акциональные режимы сознания* (воздействие, к тому же, зачастую производится и без участия вербальных знаков; вербальные знаки никогда не изолированы от других каналов воздействия). Язык и вербальные знаки не являются алгоритмами сознания.

Краткий эксперимент «Одно слово» (проводится во время доклада) позволяет продемонстрировать *свободу естественного интеллекта от знаков* и *процесс создания актуальных множеств*, что недоступно искусственному интеллекту.

Исследователи знакового (вербального) процесса, проводя условный тест Тьюринга для ИИ, могут одновременно протестировать себя и выбрать между двумя возможностями:

1) признание знаков-*модулей* и состоящего из них *языка* означает *алгоритмичность естественного мышления*, тождество человеческого сознания машинному. Наоборот,

2) признание условно выделяемых знаков-*намеков* (отсылающих к невербальным режимам сознания семиотического актора) и *упразднение языка* как неэффективного теоретического инструмента открывает для естественного интеллекта пространство свободы, которой он де-факто пользуется. *Алгоритмизации* ЕИ удается избежать.

Каждая из выбираемых возможностей влечет за собой масштабные теоретические и практические следствия.

Литература

Бабкин Э., Козырев О., Куркина И. 2006. Принципы и алгоритмы искусственного интеллекта. Н. Новгород: Нижегород. гос. техн. ун-т.

Вдовиченко А., Тарасов Е., 2017. Вербальные данные в составе коммуникативного действия: язык, текст, автор, интерпретатор // Вопросы психолингвистики. Вып. 4 (34). С. 22–39.

Вдовиченко А. 2008. Расставание с «языком». Критическая ретроспектива лингвистического знания. М.: ПСТГУ.

Вдовиченко А. 2016. О несамотождественности языкового знака. Причины и следствия «лингвистического имяславия» // Вопросы философии № 6. С. 164–175.

Вдовиченко А. 2018. С возвращением, автор, но где же твой текст и язык? О вербальных данных в статике и динамике. Часть I // Вопросы философии № 6. С. 156–167; Часть II // Вопросы философии № 7. С. 57–69.

Вдовиченко А. 2020. Смыслообразование в логических парадоксах: принцип коммуникативной определенности // Часть 1. Вопросы философии № 2. С. 71–85; Часть 2. Вопросы философии № 3. С. 107–118; Часть 3. Вопросы философии № 4. С. 143–160.

Сепир Э., 1993. Язык. Введение в изучение речи // Сепир Э. Избранные труды по языкознанию и культурологии. М.: Прогресс.

Соссюр Ф. 1977. Труды по языкознанию. М.: Прогресс.

Столяр С., Владыкин А., 2011. Информатика: Представление данных и алгоритмы. СПб.: Невский Диалект; М.: БИНОМ. Лаборатория знаний.

Хомский Н. 1972. Язык и мышление. М.: МГУ.

Chafe W. 2013. Toward a thought-based linguistics // Bischoff S. T., Jarty C. (eds). Functional approaches to language. Berlin: De Gruyter Mouton. P. 107–130.

Jackendoff R. 2007. Language, consciousness, culture: Essays on mental structure. Cambridge (MA): MIT Press.

Soares N. 2015. The Value Learning Problem. Technical report 2015–4. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/ ValueLearningProblem.pdf

НЕЙРОСЕМАНТИКА: ЛАТЕРАЛИЗАЦИЯ РЕПРЕЗЕНТАЦИЙ МЕНТАЛИСТСКИХ ПОНЯТИЙ[1]

Б. М. Величковский
Boris.Velichkovsky@tu-dresden.de
НИЦ «Курчатовский институт»
РГГУ

В. И. Заботкина
zabotkina@rggu.ru
РГГУ

Л. Я. Зайдельман
luda.zaidelman@yandex.ru
НИЦ «Курчатовский институт»
РГГУ

А. А. Котов
Kotov_aa@nrcki.ru
НИЦ «Курчатовский институт»
РГГУ

З. А. Носовец
nosovets@phystech.edu
НИЦ «Курчатовский институт»
РГГУ

В. Л. Ушаков
tiuq@yandex.ru
Институт перспективных
исследований мозга
МГУ имени М. В. Ломоносова

Примеры семантической селективности отдельных областей мозга известны из клинических наблюдений в течение десятилетий, но вопрос о принципах и механизмах репрезентации семантики речи в головном мозге был впервые поставлен и изучен лишь недавно, в работе Александра Хута и его коллег (Huth et al. 2016). Новая, нейросемантическая парадигма изучения значений существенно дополняет начатую в необихевиоризме и продолженную в психосемантике парадигму исследований. Хут и коллеги изучали репрезентацию понятий английского языка, используя моделирование BOLD (Blood-Oxygene-Level-Dependent) ответов мозга 7 носителей этого языка в ответ на продолжительное акустическое предъявление осмысленного нарратива. Картирование семантических категорий естественного языка осуществлялось этой группой исследователей на идеализированную двумерную поверхность коры головного мозга. Авторы отмечают две особенности полученных ими результатов. Во-первых, семантические карты языка довольно точно повторяют очертания т. н. дефолтной системы мозга, т. е. системы мозга, которая увеличивает свою активность в состоянии покоя. Во-вторых, не наблюдалось ожидавшейся преимущественной локализации семантических категорий в левом полушарии, со второй половины XIX века считающегося собственно «речевым» и «доминантным» в отношении произвольной активности. Более того, складывается впечатление более активной работы правого полушария, но данные, которые позволили бы проверить это предположение, в публикации отсутствуют.

Нейрофизиологические объяснения в области семантики ранее искались при помощи методологии когнитивного вычитания, когда из экспериментального условия вычитается контрольное (baseline) условие. В качестве контрольного обычно бралось состояние покоя, когда мозг предположительно находится на минимуме своих энергетических и когнитивных возможностей. Это предположение оказалось ошибочным: покой связан, как отмечено выше, с интенсивной работой ряда структур мозга, которые и обеспечивают функционирование внутреннего мира человека. В когнитивных исследованиях работы мозга вычитание возможно, если мы имеем дело с модулярными архитектурами.

Нами (Velichkovsky et al. 2020) показана возможность нейросематического картирования по данным функциональной магнито-резонансной томографии в ответ на предъявление русскоязычных текстов. В данной работе стимульный материал включал тексты, отражающие опыт преодоления социокультурных угроз из перспективы первого лица. Каждое слово было представлено в виде 997-мерного вектора, измерения которого были оценкой близости этого слова к наиболее частотным существительным и глаголам. Было оценено влияние каждого из измерений этих векторов на каждый из 10000 вокселей серого вещества головного мозга

[1] Исследование выполнено за счет гранта Российского научного фонда № 17-78-30029.

25 испытуемых. Для локализации зон мозга, связанных с восприятием потока речи, была зарегистрирована BOLD-активность при прослушивании текстов, что позволило построить матрицу преобразования векторов лингвистической разметки в вектора активности мозга.

В этом сообщении нами учитывались данные 7 испытуемых, наиболее вовлеченных в прослушивание нарративных текстов по результатам субъективного опроса и объективным показателям. С помощью метода главных компонент, как и в референтной работе (Huth et al. 2016), было построено четырехмерное пространство, для кластеризации элементов которого были выбраны точки, наиболее удаленные от центра множества и имеющие максимальную нагрузку на компоненты. Это позволило выделить 12 кластеров, которые допускали довольно простую интерпретацию. Так, кластеры 2, 9 и 10 были связаны в основном с пространственно-временной организацией описываемых в нарративах действий, а кластеры 1, 3 и 11 — с угрозой, лишениями и конфликтом. Категории долженствования, возможности и целеполагания явно концентрировались в кластере 5, тогда как понятия, связанные с субъективным опытом, такие как, например, «сознание», «чувство» и «интеллект», были однозначно отнесены кластерным анализом семантики и активности мозга к кластеру 8. Именно этот «менталистский кластер» оказался единственным, который обнаружил тенденцию к локализации числа статистически достоверно активированных вокселей мозга в правом полушарии. Во всех остальных кластерах асимметрия либо отсутствовала, либо была левосторонней.

Нейросемантический подход позволяет описывать индивидуальные различия в восприятии информации, содержащейся в относительно небольших по объему текстах. Интересной перспективой является возможность определения путем подсчета глобальной корреляции, находятся ли индивиды и целые социальные группы «на одной волне», или, другими словами, для оценки возможностей и пределов взаимопонимания. Полученные нами данные подтверждают более ранние предположения о распределенном характере репрезентации семантики в мозге человека. Результаты исследования объясняют трудности применения методологии когнитивного вычитания и свидетельствуют также о массивном участии правополушарных механизмов в обработке менталистской части внутреннего лексикона, как минимум, у носителей русского языка.

Работа поддержана грантом РНФ 17-78-30029.

Литература

Huth A. G., de Heer W. A., Griffiths T. L., Theunissen F. E., Gallant J. L. 1916. Natural speech reveals the semantic maps that tile human cerebral cortex // Nature 532, pp. 453–458.

Velichkovsky B. M., Zabotkina V. I., Nosovets Z. A., Kotov A. A., Zaidelman L. Ya., Kartashov S. I., Korosteleva A. N., Malakhov D. G., Orlov V. A., Zinina A. A., Goldberg E., Ushakov V. L. 2020. Towards semantic brain mapping methodology based on a multidimensional markup of continuous Russian-language texts // STM 12(2), pp. 14–25.

О РАСШИРЕНИИ ЗАПРОСОВ ДЛЯ ПОИСКА ФАКТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ[1]

Л. Л. Волкова
liliya@bmstu.ru
РГГУ, МГТУ им. Н. Э. Баумана

Д. Ю. Мальцева
d.y.maltseva@gmail.com
МГТУ им. Н. Э. Баумана

Эта работа посвящена решению задачи семантического поиска по базе фактов, выделенных семантическим анализатором. Факт представлен как фрагмент дерева зависимостей, где каждой лексеме присвоена семантическая валентность (в данной работе — агенс (субъект), пациенс (объект), предикат (действие)), по валентностям распределены семантические признаки (согласно Wierzbicka 1980, около 800 признаков). Будет решена задача снижения числа отказов при запросах к базе фактов путём переформулирования запросов.

Ключевые семантические анализаторы и вопросно-ответные системы для русского языка — ABBYY Compreno (Anisimovich et al. 2012), IBM Watson (Когнитивная система…), Яндекс Алиса (Как устроена Алиса…), Ф-2 (Volkova et al. 2017) (робот не отвечает на вопросы) — часто ограничивают домен анализируемой информации (IBM Watson адаптирована для медицинской рекомендательной системы, Яндекс Алиса ограничена заложенными в неё сценариями взаимодействия) либо проводят семантический анализ на основании грамматик и/или словарей семантических маркеров, что служит естественным ограничением области их знаний. Для таких систем нужны эксперты для составления грамматик и сценариев, наполнения словарей.

Обработка естественного языка предполагает работу с лексемами и семантическими признаками, что, в свою очередь, предполагает работу с лексическими объединениями слов: синонимами, квазисинонимами (частичными синонимами), гипонимами (также контекстно зависимы и не универсальны в поставленной задаче), гиперонимами, омонимами и др. В работе использован словарь лексем: для лексемы даны семантические признаки и ближайшие связанные синонимы, квазисинонимы, гиперонимы (до трёх каждой категории). База фактов содержит миллион строк. Факт представлен несколькими лексемами (каждая стоит в валентности), лексеме сопоставлено не менее одного признака (строки). Вопрос формулируется как факт, состоящий из трёх указанных валентностей, в одной из которых вопросное слово маркирует искомое, например, что(pat)-едят(p)-коровы(ag) (с лемматизацией). На выходе ожидается набор фактов, удовлетворяющих запросу. В случае отсутствия точного соответствия исходный запрос необходимо переформулировать. Для этого предусмотрены следующие замены (в порядке приоритета): синонимы, квазисинонимы, гиперонимы. При неудаче проводится замена семантических признаков на близкие на основании того, что ряд признаков имеют подчинение по дереву. Пример: семантический признак ‘поглощать’ (пищу) является предком признаков ‘есть’ и ‘пить’, все три признака можно заменять. Так, пример будет переформулирован: что-{едят, кушают, пьют}-{коровы, травоядные}, и будут найдены 2 ответа — траву и воду (слабый ответ, который при отсутствии первого всё же будет ответом: у дилеммы студента — не дать никакого ответа или ответить на смежный вопрос — нет очевидного решения).

С помощью предложенного метода проведён ряд изысканий для конкретной базы знаний (фактов), они демонстрируют механизм изучения этой базы, приводящий к возникновению сценариев поиска, свойственных когнитивному процессу человека. Также это позволяет сделать выводы о наполнении словарей замен. Рассмотрим выборку фактов из базы с глаголами, схожими со ‘смотреть’. Здесь встречаются валентности траектории, цели, инструментатива, локатива. Употреблены глаголы, связанные со зрением: смотреть, посмотреть, взгля-

[1] Исследование выполнено за счет гранта Российского научного фонда (проект № 19-18-00547).

нуть, уставиться, покоситься, видеть. Схожее значение имеет ‘обернуться’, однако это нетривиальный случай, не обязательно подразумевающий взгляд. Также возможны следующие конструкции: поднять глаза, остановить, перевести, бросить, метнуть взгляд. Таким образом, для поиска релевантно использование таких глаголов в сочетании со словами ‘взгляд’, ‘взор’, ‘глаза’ в валентности пациенс, а также использование глагола ‘обводить’ в сочетании с инструментативом ‘взгляд’, тогда цель взгляда будет стоять в валентности пациенс. Не все факты в выборке имеют цель взгляда, у некоторых есть только траектория. Встречается исключение ‘смотреть в упор’: цель взгляда следует извлекать из дерева зависимостей при помощи дополнительного правила.

На основании анализа наборов примеров составляются правила, позволяющие проводить поиск для набора случаев либо предметной области. Отдельно следует рассматривать вопросы, подразумевающие либо все возможные случаи (глагол в прошедшем времени или с модальным словом: ‘на что можно смотреть?’), либо конкретную ситуацию, ограниченную контекстом беседы.

Проанализировав результаты выдачи, можно сделать следующие выводы. Предложенный метод снижает число случаев пустой поисковой выдачи за счёт проводимых в запросе замен, однако не исключает этих случаев. На объём выдачи по запросам, для которых в базе нет точных соответствий, критически влияет полнота словарей замены лексем. В результатах семантического поиска допускаются такие же ошибки, какие допускает человек в процессе обучения. Так, не все глаголы взаимозаменяемы, и требуется вводить дополнительные правила для уточнения корректности замен, в том числе в зависимости от контекста (‘играть’ в музыке или в спорте). Правила для замены слов по семантическим признакам требуют ограничения для обеспечения корректности замен. В задаче семантического поиска требуется экспертная работа по написанию правил и составлению, корректировке словарей. В перспективе возможно применять словарь гипонимов для ответов на специальные виды вопросов, подразумевающих анализ отношения часть-целое (‘что люди умеют делать с инструментами?’ влечёт замену лексем ‘люди’, ‘инструменты’). Возможно выделять в запросе прямое указание на семантический признак (в примере выше — семантические признаки из группы инструментативов).

References

Как устроена Алиса. Лекция Яндекса [эл. ресурс]. Режим доступа: https://m.habr.com/en/company/yandex/blog/349372/ [Дата обращения 17.06.2019].

Когнитивная система IBM Watson [эл. ресурс]. Режим доступа: https://m.habr.com/en/company/ibm/blog/266015/ [Дата обращения 17.06.2019].

Anisimovich K. V. et al. 2012. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11(18): В 2 т. Т.2: Доклады специальных секций. М.: Изд-во РГГУ. С. 91–103.

Volkova L., Kotov A., Klyshinsky E., Arinkin N. A Robot Commenting Texts in an Emotional Way // CCIS 754, pp. 256–266. Heidelberg: Springer. 2017.

Wierzbicka A. 1980. Lingua Mentalis: The semantics of natural language. New York: Academic Press.

ЭМОЦИОНАЛЬНЫЕ РЕАКЦИИ РОБОТА-КОМПАНЬОНА НА ФРАЗЫ НА РУССКОМ ЯЗЫКЕ[1]

Л. Л. Волкова
liliya@bmstu.ru
РГГУ, МГТУ им. Н. Э. Баумана

А. А. Котов
kotov_aa@nrcki.ru
РГГУ, НИЦ «Курчатовский институт»

Данная работа находится на пересечении двух перспективных областей — моделирования аффективного поведения роботов и естественно-языковых интерфейсов (ЕЯИ). При беседе мимика, движения рук и головы несут дополнительную информацию. Хотя при мультимодальной коммуникации невербальный канал не является главным (хотя может выражать до 60 % передаваемой информации (Engleberg, Wynn 2006)), он важен для формирования личностного восприятия как предмета разговора, так и носителя ЕЯИ. Аффективные ИИ и роботы призваны увлечь человека (Russell 2003; Andre et al. (eds.) 2004; Breazeal 2002; Cassel et al. (eds.) 2000; Paiva et al. (eds.) 2007; Han et al. 2012) и могут иметь важное социальное применение: в больницах и у выздоравливающих дома, в спецшколах и детских домах они могут развлечь человека беседой, увлечь его совместным решением головоломок (Zinina et al. 2020), справиться о здоровье и при необходимости вызвать врача. Аффективная составляющая поведения и вариативность реакций снимают проблему привыкания и потери интереса, типичную для робо-игрушек и голосовых помощников с ограниченным, узким набором сценариев (Jokinen 2015).

Для моделирования аффективных реакций выделяют различные базовые эмоции (Ekman 1999), используют параметрическую модель эмоций Валентность-Возбуждение (Tao et al. (eds.) 2005), комбинации эмоций (Ochs et al. 2005; Kirby 2010). ИИ нашего робота оперирует коммуникативными функциями, или целями (выразить согласие или неодобрение, смущение или сочувствие, ожидание обратной связи, дистанцирование, привлечение внимания и др.).

Мы используем синтактико-семантический анализатор, преобразующий предложения на русском языке в деревья зависимостей, аннотированные семантическими маркерами (используется собственный LR-парсер, использующий более 600 правил и словарь из более 100000 лемм с более чем 800 семантическими признаками, согласно (Wierzbicka 1980), и около 120000 употреблений маркеров в словаре), и извлекающий из них факты (поддеревья, в которых семантические маркеры распределены по вершинам, соответствующим валентностям: предикат, объект, субъект, причина, локация и др.). Пример факта из 2 валентностей p (предикат; лексема «дать») и pat (дополнение, пациенс; лексема «слово») с перечислением семантических маркеров: {p дать {п-отдавать-терять, п-иметь}, pat слово {о-текст, о-абстр, о-коммуникация, о-ком-действие, о-слово, о-соц-ритуал, о-обещание}}. В фактах распознаются входящие эмоциональные паттерны (смысл), в зависимости от структуры поддерева и конкретных семантических маркеров, например: неадекватность (событий, действий, восприятия), обман, опасность, воодушевление, внимание, вкус-запах (должна последовать реакция на описанный хороший вкус или запах) — всего 34 паттерна (13 негативных и 21 позитивный). Под пониманием смысла здесь можно понимать выбор наиболее адекватной реакции.

На смысл фактов формируется реакция с маркером «эмоции» (коммуникативной функции), они могут конкурировать. Для функции, набравшей наибольшее число очков активации, выбирается одна из множества предусмотренных реакций (так, радость можно проявлять по-разному). Мультимодальная реакция формируется в виде кадра поведения, включающего текст, жесты рук и головы, мимику. Текст озвучивается посредством Яндекс SpeechKit,

[1] Исследование выполнено за счет гранта Российского научного фонда (проект № 19-18-00547).

жесты исполняются на модели робота: у робота 2 руки по 2 степени свободы и голова-монитор на шее с 2 степенями свободы. Мы приняли решение не идти по пути создания андроидов, чтобы обойти проблему «зловещей долины» восприятия робота человеком (Mori 1970/2012).

Робот подчас понимает факты буквально. Например, фразу из текста «они [механические игрушки] толкали вперед воображение изобретателей» робот понял в ключе опасности и субъективности (командование: толкать вперёд). В подобных случаях, основанных на двусмысленности слов, робот фактически демонстрирует юмор. Приведём ещё два примера в адаптации:

«Ему [первобытному человеку] часто приходилось вступать в бой и с огромным косматым мамонтом, и с могучим пещерным медведем».

Робот: Бой с мамонтом — это огромный мир необычных вещей и событий!

«Мы все говорим по привычке, что электрическая лампочка “горит”».

Робот: То есть всё-таки обманываете!

Таким образом, моделирование аффективных реакций робота на смысл фраз позволил нечаянно добиться воспроизведения юмора. Наш робот не контролирует этот процесс, но сам факт свидетельствует в пользу правильности выбранного подхода к моделированию мышления.

References

Andre E., Dybkjær L., Minker W. et al. (eds.). 2004. Affective Dialogue Systems. Berlin: Springer.

Breazeal C. 2002. Designing Sociable Robots. Cambridge: MIT Press.

Cassel J., Sullivan J., Prevost S. et al. (eds.). 2000. Embodied Conversational Agents. Cambridge, London: MIT Press.

Ekman P. 1999. Basic Emotions // T. Dalgleish, M. Power (eds.). Handbook of Cognition and Emotion. Sussex: John Wiley & Sons.

Engleberg I. N., Wynn D. R. 2006. Communication Principles and Strategies // My Communication Kit Series. Boston: Allyn & Bacon, p. 133.

Han J. G. et al. 2012. Investigating the use of non-verbal cues in human-robot interaction with a Nao robot // Proc. 3rd IEEE Int. Conf. on Cognitive Infocommunications (CogInfoCom 2012). Kosice, pp. 679–683.

Jokinen K. 2015. Communicative engagement and authonomous robots: social interaction in WikiTalk software // Int. Symposium «Research for gesturing in human-computer interaction: robots and virtual agents», MSLU, 26 october 2015.

Kirby R., Forlizzi J., Simmons R. 2010. Affective social robots // Robotics and Autonomous Systems, vol. 58. Amsterdam: Elsevier, pp. 322–332.

Ochs M. et al. 2005. Intelligent Expressions of Emotions // Proc. ACII 2005, pp. 707–714.

Paiva A., Prada R., Picard R. W. (eds.) 2007. Affective Computing and Intelligent Interaction // LNCS, vol. 4738. Berlin, Heidelberg: Springer.

Russell J. 2003. Core affect and the psychological construction of emotion // Psychological Review, vol. 110 (1). American Psychological Association, Washington, pp. 145–172.

Tao J., Tan T., Picard R. (eds.) 2005. Affective Computing and Intelligent Interaction // LNCS, vol. 3784. Berlin, Heidelberg: Springer.

Mori M. 1970/2012. The uncanny valley (K. F. MacDorman & N. Kageki, Trans.) // IEEE Robotics & Automation Magazine, vol. 19 (2). IEEE, Piscataway, pp. 98–100.

Wierzbicka A. 1980. Lingua Mentalis: The semantics of natural language. New York: Academic Press.

Zinina A. et al. 2020. Emotional behavior of robots during the spatial puzzle solving with children // Computer linguistics and intellectual technologies.

ПРИЕМЫ МНЕМОТЕХНИКИ В РИГВЕДЕ (О СПОСОБАХ ЗАУЧИВАНИЯ, ХРАНЕНИЯ В ПАМЯТИ И ФУНКЦИОНИРОВАНИИ ДРЕВНЕИНДИЙСКИХ ГИМНОВ)

О. А. Волошина

oxanav2005@mail.ru

МГУ им. М. В. Ломоносова

Поскольку древнеиндийские тексты имели хождение исключительно в устной форме, проблема заучивания и хранения в памяти объемного текста приобретала огромное значение для жрецов-брахманов. В докладе обсуждаются разработанные в ведийскую эпоху мнемонические техники, используемые в процессе заучивания, хранения в памяти, трансляции и функционирования в ритуале священных гимнов Ригведы.

С одной стороны, сложнейший суггестивный текст Ригведы разворачивается линейно в процессе произнесения текста, в этом смысле можно говорить о связности текста — о принципах его линейного развертывания. С другой стороны, речь идет о сложности ведийского гимна, о семантической насыщенности текста, в котором каждое слово связано ассоциативными связями с устойчивыми эпитетами, с синонимами, словами близкими по значению и т. п. Таким образом, каждое слово вступает в парадигматические отношения, позволяя припоминать, во-первых, все контексты Ригведы, в которых встречается конкретное слово (сочетания слов), во-вторых, подбирать синонимы, перифрастические конструкции, метафорические обозначения, позволяющие формировать семантическое поле, основанное на общем значении.

Для успешного заучивания Ригведы были разработаны специальные техники, основанные на сегментации слитного текста и способов аранжировки сегментов, то есть слитный текст членился на отрывки (например, на слова), которые затем комбинировались и переплетались часто весьма причудливым образом. Например, kramapāṭha (букв. *чтение шаг за шагом*), jaṭāpāṭha (букв. *заплетенное чтение*) или ghanapāṭha (букв. *густое или прочное чтение*). Такие сложные комбинации часто затемняли смысл текста (или вообще лишали его смысла), но древние индийцы вовсе не стремились сохранить значение сакральных гимнов. Лишение текста смысла, заучивание бессмысленных сочетаний слогов, как представлялось брахманам, обеспечивало трансляцию и функционирование текста без искажений, без вариантов, без возможностей субъективной интерпретации.

Однако попытки осмыслить древнейшие гимны, понять содержание молитв решали задачу интерпретации смысла ведийских гимнов. Для успешного выполнения процедуры семантического анализа к гимнам Ригведы был составлен древнейший глоссарий Нигханту (Nighaṇṭu). Этот небольшой по объему список опорных слов (чуть более тысячи слов), очень важных для понимания текста Ригведы, был создан как мнемонический словарь, позволяющий вспоминать и объединять фрагменты сакрального текста, посвященные определенной теме.

В докладе приводятся примеры функционирования списка ключевых слов к Ригведе (глоссарий Нигханту) в качестве своеобразных мнемонических ориентиров, опираясь на которые, рассказчик мог излагать или комментировать текст. Классы слов в Нигханту отражали семантическую структуру текста Ригведы в целом, являясь семантическими опорами толкования и интерпретации.

Ключевые слова — эти смысловые вехи текста — отражали основное содержание Ригведы как бы в сжатом, концентрированном виде. Как известно, набор ключевых слов есть, с одной стороны, механизм компрессии текста, а с другой стороны, по ключевым словам можно воссоздать (реконструировать) текст. Однако списки слов в Нигханту выстраиваются

не по принципу последовательного вычленения опорных слов из следующих в определенном порядке гимнов, а представляют собой набор ключевых слов ко всему тексту в целом, являясь основой реконструкции не текста как такового, а совокупности текстов, мифологической картины мира, воплощенной в Ригведе.

Литература

The Nighaṇṭu and the Nirukta. The oldest indian treatise on etymology, philology and semantics. 1967. Text and translation by Lakshman Sarup. Delhi, Varanasi, Patna.

Волошина О. А. 2016. Магический мир звучащего слова (об устной форме научного текста в древней Индии) // К 90-летию Ю. В. Рождественского. М.: Добросвет. С. 31–42.

ОНОМАСИОЛОГИЧЕСКИЙ ПОДХОД К ПРОБЛЕМЕ ЛЕКСИЧЕСКИХ И СЕМАНТИЧЕСКИХ ИЗМЕНЕНИЙ

В. В. Глебкин

gleb1514@gmail.com

РАНХиГС

Ономасиологический подход, развиваемый в последние десятилетия рядом авторов (в первую очередь Дирком Герарцем и его группой (Geeraerts et al. 1994; Geeraerts 1997 2006, 2018)), не привлек еще, на мой взгляд, должного внимания исследователей в области когнитивной лингвистики. В отличие от базового для лингвистики семасиологического подхода, изучающего семантическую структуру отдельных слов и лексических конструкций и ее эволюцию со временем, ономасиологический подход исходит из определенного концептуального содержания и исследует спектр слов и лексических конструкций, это содержание обозначающих (определения двух указанных подходов см., напр., в Baldinger 1980, 278).

Так, в работе Geeraerts et al. 1994 описывается лексическая конкуренция между словами *caleçon* и *legging(s)* для обозначения брюк определенного типа в бельгийском голландском и нидерландском голландском в период с 1988 по 1992 г. В исследовании показано, что если в нидерландском голландском выбор *legging(s)* явно доминирует на всем протяжении указанного периода, то в бельгийском голландском сначала единственно возможной формой является *caleçon*, но к 1992 г. в одном случае из трех носители языка выбирают *legging(s)*. Авторы также указывают на социокультурные факторы, которые могли привести к указанным изменениям.

Предложенный в работах Герарца и коллег подход допускает распространение и на ситуации более разветвленного и/или менее определенного лексического выбора. Одной из таких ситуаций является лексический выбор при обращениях. Так, в русском языке возможен широкий спектр обращений к незнакомому человеку: обращения, опирающиеся на универсальные модели (*мил человек (милый человек), добрый человек, нехороший человек*); формальные социальные модели (*товарищ, гражданин, господин, товарищ начальник, гражданин начальник*); неформальные социальные модели (*друг*, *мой друг* (*друг мой*), *милый друг*, *дорогой друг*, *подруга*, *дорогая подруга*, *милая подруга*); гендерно-возрастные (*мальчик*, *девочка*, *юноша*, *молодой человек*, *девушка*, *мужчина*, *женщина*, *дядя*, *дяденька*, *дед*); «семейные» модели (*сынок*, *браток*, *братишка*, *доченька*, *дочка*, *мамаша*, *папаша*). В одних случаях выбор обращения остается за говорящим и носит вероятностный характер, в других он полностью детерминирован социокультурной ситуацией. Интересно в данном контексте сопоставление обращений *гражданин начальник* и *товарищ начальник* и изменение их частотности со временем.

Особую сложность и интерес представляют ситуации, когда формируется новый, еще анонимный фрагмент социокультурной реальности, требующий своего именования. Разные носители культуры предлагают свои варианты имен, либо вводя неологизмы, либо трансформируя семантическую структуру уже известных слов и выражений. Эти варианты вступают между собой в конкуренцию. История такой конкуренции составляет важное дополнение к традиционному диахронному анализу отдельных лексем и лексических конструкций. Так, результатом социокультурных трансформаций, вызванных «Манифестом о вольности дворянства» Петра III (1762 г.) и «Жалованной грамоты дворянству» Екатерины II (1785 г.), освобождавших дворян от обязательной службы, стало появление социальной группы, жизнь которой оказалась «замкнутой на себя», лишенной внешнего регулятора (служение императору, Отечеству, религиозное служение и т. д.), что породило проблему «пустого времени» и необходимости его заполнения. Это, в свою очередь, привело к особому психологическому пе-

реживанию «пустого времени» и поиску слов, его маркирующих. Разные авторы использовали для такой маркировки понятия *скука*, *тоска*, *уныние*, *хандра, сплин.* В итоге победило слово *скука*, заметно изменившее при этом свою семантическую структуру (Глебкин 2018).

Другим подобным примером являются культурные трансформации второй половины XIX века, в результате которых наряду с дворянством, отвечавшим ранее за производство и трансляцию образцов «высокой» культуры и противостоящим по этому критерию народу, в производство и трансляцию культуры включились представители размытой по формальным признакам социокультурной группы, маркируемой как разночинцы. Произошедший сдвиг вызвал потребность в поиске слова, объединяющего дворянство и разночинцев по указанному критерию, что породило ситуацию лексической конкуренции. В качестве конкурирующих конструкций выступали выражения *образованный класс*, *привилегированный класс* и др., но победителем в этой борьбе вышло слово *интеллигенция*, ставшее в дальнейшем системообразующим для русской и советской культуры.

Интересен вопрос о факторах, обеспечивающих победу в подобной конкурентной борьбе. Среди таких факторов можно выделить внутренние (простота слова, удобство в произношении, степень корреляции между базовой семантической интуицией и новыми смыслами, которые слово должно обрести и т. д.) и внешними (статус и авторитет человека / субгруппы, выдвигающей эти слова в качестве претендента для фиксации совершившихся изменений).

References

Глебкин В. В. 2018. Категории русской культуры XVIII–XX вв. Скука. М.; СПб.: Центр гуманитарных инициатив.

Baldinger K. 1980. Semantic Theory. Oxford: Basil Blackwell.

Geeraerts D. 1997. Diachronic prototype semantics: a contribution to historical lexicology. Oxford: Clarendon Press; New York: Oxford University Press.

Geeraerts D. 2006. Words and other wonders: papers on lexical and semantic topics. Berlin; New York: Mouton de Gruyter.

Geeraerts D. 2018. Ten Lectures on Cognitive Sociolinguistics. Leiden; Boston: Brill.

Geeraerts D., Grondelaers S., Bakema P. 1994. The structure of lexical variation: meaning, naming, and context. Berlin; New York: M. de Gruyter.

ВОЗРАСТНЫЕ РАЗЛИЧИЯ РЕПЕРТУАРА ЦВЕТОНАИМЕНОВАНИЙ У СОВРЕМЕННЫХ НОСИТЕЛЕЙ РУССКОГО ЯЗЫКА[1]

Ю. А. Грибер
Смоленский государственный университет
y.griber@gmail.com

Г. В. Парамей
Ливерпуль Хоуп Университет
parameg@hope.ac.uk

Словарь цветообозначений в русском языке исключительно богат по репетуару и лингвистическим формам. Целью настоящей работы было изучение лексического состава и форм цветонаименований, а также структуры и свойств системы цветовых категорий в различных возрастных группах носителей русского языка с использованием психолингвистического подхода (Paramei, Griber, & Mylonas 2018).

Данные получены в 2018–2019 гг. с помощью онлайн эксперимента (http://colournaming.com), в ходе которого участникам — носителям русского языка — в случайном порядке предъявлялись виртуальные цветные карточки, отобранные компьютерной программой из 606 образцов эксперимента. Координаты цветовых образцов задавались в пространстве CIELAB. Респондентам предлагалось назвать каждый цвет, используя наиболее подходящее цветообозначение — простые и составные слова и словосочетания.

Выборочная совокупность респондентов формировалась с помощью комбинированной стратегии. На первом этапе (n<1000) использовалась простая случайная безповторная выборка. Далее применялась стратегия типической выборки, стратифицированной по возрасту, полу и месту проживания респондентов и осуществлялся направленный рекрутинг участников до выполнения всех необходимых условий репрезентативности выборки.

Очищенная база данных содержала 55515 ответов, полученных от 2457 респондентов (1402 женщины, 1055 мужчин), принадлежащих к разным возрастным группам в диапазоне от 16 до 98 лет. Для анализа возрастных различий полученные данные были стратифицированы по семи возрастным группам: 16–19 лет (средний возраст 17,59 ± 1,19), 20–29 (средний возраст 24,86 ± 3,11), 30–39 (средний возраст 35,04 ± 2,89), 40–49 (средний возраст 45,18 ± 2,86), 50–59 (средний возраст 55,63 ± 2,90), 60–69 (средний возраст 65,05 ± 2,87) и ≥70 лет (средний возраст 78,20 ± 4,97).

Для межгруппового сравнения использовались следующие показатели: (1) частота встречаемости основных (N=12) и неосновных цветовых терминов; (2) словообразовательные модели и количество слов в цветовых дескрипторах; (3) перечень объектов-референтов при описании цвета; (4) положение (координаты) центроидов основных цветовых категорий в пространстве CIELAB.

Лингвистический анализ данных выполнялся методом агломеративной иерархической кластеризации с помощью программы на языке R в среде Microsoft R Open 3.5. Расстояние между возрастными кластерами вычислялось по алгоритму минимума дисперсии Уорда (Ward). Для визуализации использовался пакет ggplot2 (рис. 1).

Сравнительный анализ цветонаименований в активном словаре респондентов разного возраста позволил сделать выводы о наличии ядра и периферии в системе цветонаименований. Ядро включает основные цветонаименования (напр.: *белый, красный, голубой* и т. д.) и традиционные цветонаименования с высокой частотностью и деривационной продуктивностью (напр.: *бирюзовый, сиреневый, бордовый*). Периферия организована более свободно и постоянно пополняется за счет включения в качестве новых референтов «рыночных» цветов,

[1] Публикация подготовлена в рамках поддержанного РФФИ научного проекта № 17-29-09145 «Картография цвета: диагностика развития цветонаименований русского языка с использованием естественно-научных, историографических, социологических и психологических методов».

образованных с использованием эталонов, сравнений с новыми продуктами питания, «заморскими» цветами и животными, пряностями и фруктами, строительными материалами (напр.: *тиффани*, *тауп*, *карри*).

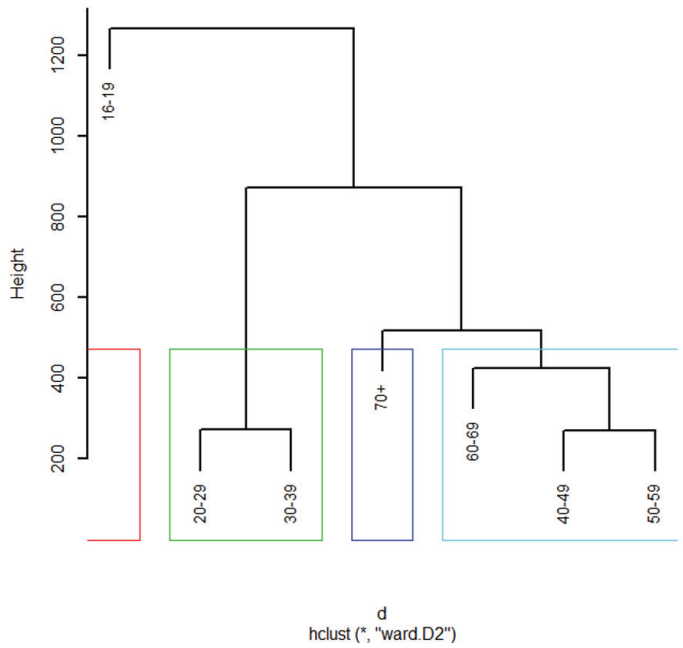


Рис. 1. Кластерный анализ возрастных различий

Репертуар цветонаименований похож у респондентов-группах 40-, 50- и 60-летних (рис. 1). Он, однако, заметно отличается от цветового словаря более молодых, 20- и 30-летних респондентов, сформировавшегося в постсоветский период.

Наибольшая лексическая динамика наблюдается в обозначении оттенков синей и зеленой части спектра, в то время как наименование красных и желтых оттенков остается относительно стабильным независимо от возраста.

С возрастом заметно снижается частота использования необычных цветонаименований и, наоборот, увеличивается частотность традиционных и основных цветонаименований, доля которых возрастает до 69% цветового словаря в группе 60–69 лет и до 72% у респондентов старше 70 лет, что почти вдвое больше, чем у молодых участников. В «зрелой» возрастной группе (70 лет и старше) цветовой словарь также заметно уступает в богатстве: из лексикона исчезают низкочастотные цветовые термины. Вероятно, поскольку с возрастом ослабляются семантические связи в лингвистических репрезентациях (Burke, Shafto 2004).

Литература

Burke D. M., Shafto M. A. 2004. Aging and language production // Current Directions in Psychological Science. Vol. 13. P. 21–24.

Paramei G. V., Griber Y. A., & Mylonas D. 2018. An online color naming experiment in Russian using Munsell color samples // Color Research & Application. Vol. 43. P. 358–374.

ДЕФИЦИТ СВЯЗНОЙ РЕЧИ У ПАЦИЕНТОВ С ОПУХОЛЯМИ ГОЛОВНОГО МОЗГА В РАННЕМ ПОСЛЕОПЕРАЦИОННОМ ПЕРИОДЕ

Н. Э. Гронская
ngronskaya@hse.ru
НИУ ВШЭ — Нижний Новгород

О. В. Драгой
odragoy@hse.ru
НИУ ВШЭ

Н. Ю. Антонова
natalie.eskadron@gmail.com
НИУ ВШЭ — Нижний Новгород

М. В. Шибанова
mshibanova4@gmail.com
НИУ ВШЭ — Нижний Новгород

М. С. Нелюбина
marnelyubina@gmail.com
НИУ ВШЭ — Нижний Новгород

К. С. Яшин
jashinmed@gmail.com

И. А. Медяник
med_neuro@inbox.ru

Приволжский исследовательский медицинский университет

Введение

Речевые нарушения (афазия) у пациентов с поражениями мозга являются крайне инвалидизирующими (Glozman 1980). Существует категория пациентов, у которых патология мозга развивается в течение продолжительного времени, предоставляя языковой функции возможность реорганизоваться за счет сохраняющейся в любом возрасте пластичности мозга. Наиболее многочисленной группой таких пациентов являются люди с опухолями головного мозга. Однако самой частой субъективной жалобой пациентов после нейрохирургической операции являются трудности повседневной коммуникации, значительно снижающие качество жизни (Papagno et al. 2012). Кроме того, в работе (McCarron et al. 2017) отмечается, что в раннем послеоперационном периоде (2–3 дня после операции) речь пациентов отличается в худшую сторону по сравнению с их же речью до операции. Цель настоящего исследования состояла в том, чтобы изучить лингвистическую природу дискурсивных трудностей у пациентов с опухолями головного мозга в раннем послеоперационном периоде и, в частности, выяснить, являются ли возникающие трудности следствием нарушения более микроуровневого лингвистического дефицита.

Метод

В исследовании приняли участие 22 русскоговорящих пациента (13 женщин; возрастной диапазон — от 18 до 62 лет, средний возраст — 43 года) с опухолями левого полушария головного мозга, значительно вовлеченного в языковые процессы у большинства людей. Только один пациент имел опухоль в правом полушарии. Все, кроме одного пациента, были правшами, один — переученный левша. Один билингвальный пациент являлся носителем русского и мордовского языков, остальные 21 были монолингвальны. Выборку составили пациенты с первичными и вторичными нейрохирургическими вмешательствами. Все пациенты были прооперированы в Приволжском исследовательском медицинском университете в Нижнем Новгороде. Все пациенты были протестированы при помощи Русского афазиологического теста (РАТ; Ivanova et al. 2016) до и через 1–8 дней после операции. Данная батарея тестов включает в себя субтесты на порождение и понимание речи на разных языковых уровнях. В данном исследовании нас, прежде всего, интересовал дискурсивный уровень, ко-

торый оценивался по результатам выполнения задания на порождение нарратива по сюжетному рисунку. Четыре параметра — общая беглость речи, грамматическая правильность, парафазии (нерелевантные лексические замены) и информативность оценивались по шкале от 1 до 5 баллов. Статистический анализ проводился в среде R (https://www.r-project.org/) с использованием парного t-теста и корреляции Пирсона.

Результаты

До операции большинство пациентов продемонстрировали нормальные или близкие к ним результаты — по данным сравнения с нормативами РАТ. Послеоперационное тестирование показало, что только у 5 пациентов наблюдаются такие же результаты, как до операции, или лучше, остальные 17 человек продемонстрировали ухудшение показателей связной речи в задании рассказа по рисунку по одной или нескольким шкалам оценки и как группа — по всем четырем шкалам ($p < 0.05$). Максимальное ухудшение наблюдалось по критериям беглости речи (в среднем, на 15%), грамматической правильности (16%) и информативности (13%); в меньшей степени — по критерию наличия парафазий (9%). Наконец, выявленные нарушения порождения связной речи значимо не коррелировали с более низкоуровневыми показателями нарушения речи на фонологическом, лексическом и синтаксическом уровнях по данным соответствующих субтестов РАТ.

Выводы

В результате исследования было выявлено снижение способности к порождению дискурса в остром послеоперационном периоде у большинства протестированных нейрохирургических пациентов по четырем критериям: беглость речи, грамматическая правильность, парафазии и информативность. При этом в случае возникновении послеоперационных нарушений связной речи они наблюдались на всех четырех уровнях. Чрезвычайно интересно, что выявленные дискурсивные нарушения не коррелируют с более микроуровневыми нарушениями (на уровне фонологии, лексики и синтаксиса), то есть являются независимыми от них в данной популяции пациентов.

Литература

Ivanova M., Dragoy O., Akinina J., Soloukhina O., Iskra E., Khudyakova M., & Akhutina T. 2016. AutoRAT at your fingertips: Introducing the new Russian Aphasia Test on a tablet. Frontiers in Psychology Conference Abstract: 54th Annual Academy of Aphasia Meeting.

Glozman J. M., Kalita N. G., Tsyganok A. A. 1980. On the system of methods of aphasics group rehabilitation. International Journal of Rehabilitation Research, 3(4), pp. 519–526.

McCarron A., Chavez A., Babiak M., Berger M. S., Chang E. F., Wilson S. M. 2017. Connected speech in transient aphasias after left hemisphere resective surgery. Aphasiology, 31(11), pp. 1266–1281.

Papagno C., Casarotti A., Comi A., & Bello L. 2012. Measuring clinical outcomes in neuro-oncology. A battery to evaluate low-grade gliomas (LGG). Journal of NeuroOncology, 108(2), pp. 269–275.

АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ТОНАЛЬНОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

А. А. Гурин

anatoly196674@gmail.com

РЭУ им. Г. В. Плеханова

Семантический анализ данных имеет огромный потенциал применения в различных областях науки и бизнеса. В настоящее время он используется в мониторинговых, аналитических, сигнальных системах, в системах документооборота, рекламных платформах и многом другом. Анализ тональности текста заключается в извлечении из текста мнений и эмоций, а также их последующей обработки, относится к методам контент-анализа и является одним из средств изучения субъективности в естественном языке.

Большая часть информации в мире не структурирована, поэтому возникает проблема обработки и извлечении полезных данных. Существует множество методов извлечения и обработки текстовой информации. Это может быть использование лексико-грамматических конструкций наряду с оценочным лексиконом.

Анализ позитивной и негативной тональности является одним из направлений в автоматизированном анализе естественно-языковых текстов. Под тональностью текста понимается позитивное или негативное отношение его автора к объекту (персоне, организации, товару), фигурирующему в тексте. Любой пользователь сети интернет может оставлять свои отзывы о товаре или услуге, высказывать мнение о людях и событиях. В связи с этим возникает необходимость оперативно и качественно отслеживать информационное поле и выявлять нужные данные (в зависимости от сферы применения).

Сложности при решении задачи анализа тональности текста связаны с использованием метафор, устойчивых выражений. Зависимость тональности от контекста, например, комментарий «прочитай книгу» может быть расценен как позитивный на сайте отзывов о книгах и как негативный на сайте отзывов о фильмах. При анализе отзывов существует важная проблема отделения фактов от мнений, а также учет точки зрения. Также обработка тональности зависит от предметной области текста.

За основу разработанной модели был взят подход «Мешок слов» (Bag of words) (Brownie 2017), реализованный с использованием словарей тональности русского языка. Один из словарей, участвующий в определении тональности был собран на онлайн-краудфандинг-сервисе «карта слов» и доступен всем желающим для проведения научных исследований, остальные словари собирались и размечались самостоятельно. Разработанный алгоритм и его реализация позволяют наиболее точно определить тональность высказывания или текста, классифицируя текст на 5 основных типов оценок: сильно негативная, негативная, нейтральная, положительная и сильно положительная. Вдобавок к определению общей тональности текста в алгоритме реализовано распознавание эмоциональной составляющей с использованием модели OCC (Ortony, Clore, Collins) (Steunebrink et al. 2008), описывающей качественную и количественную сторону 22 видов эмоций. Таким образом, анализируемому тексту присваивается эмоциональная оценка с выделением преобладающих эмоций.

Инструментом реализации данного алгоритма стал язык программирования Python, который включает в себя высокопроизводительные библиотеки, для работы с текстовыми данными, такие как NLTK, Numpy, Pymorpher2, и другие. Данные библиотеки позволяют быстро и эффективно проводить подготовку данных на предварительном этапе, разбивая тексты на предложения, слова и очищая их от лишних символов.

После предварительного этапа обрабатываемому тексту необходимо пройти нормализацию. Обработанные данные лемматизируются и сопоставляются с тональными словарями. После присвоения весов их сумма считается и дается оценка всего текста. На данном этапе

немаловажным фактором является скорость обработки, программная реализация алгоритма должна быть оптимальной и производить расчеты за адекватное время. Таким образом, в данной реализации для тональных словарей используется конструкция «dict». Данная конструкция является самой быстродоступной для перебора, а также позволяет хранить данные типа ключ/значение, что упрощает работу со словарями, с одной стороны, и ускоряет работу по перебору значений, с другой.

Для демонстрации работы алгоритма было взято произведение Л. Н. Толстого «Война и мир», том I и II. Особый интерес представляет общая тональность каждого из томов, какие эмоции преобладают в данном произведении и время, за которое будет получена оценка тональности. Ниже на рисунке 1 представлен снимок экрана — результата тональности после обработки текста первого тома.

```
Очки тональности текста: 1.0779765908785147
Текст имеет положительную окраску
Содержание эмоции радости в тексте: 0.4669900369436992
Содержание эмоции горя в тексте: 0.1960093578661175
Содержание эмоции гнева в тексте: 0.2736904858683576
Содержание эмоции страха в тексте: 0.213171467541031
Содержание эмоции надежды в тексте: 0.18607339963327282
Содержание эмоции возмущения в тексте: 0.2718839480078404
Содержание эмоции жалости в тексте: 0.3089179741484432
Содержание эмоции облегчения в тексте: 0.16258840744654907
Содержание эмоции удовлетворения в тексте: 0.09574650660741223
Содержание эмоции разочарования в тексте: 0.0009032689302586059
Содержание эмоции радости за другого в тексте: 0.4669900369436992
Содержание эмоции гордости в тексте: 0.16258840744654907
Содержание эмоции восхищения в тексте: 0.807856633155392
Содержание эмоции стыда в тексте: 0.12916745702698065
Содержание эмоции подтвердившегося страха в тексте: 0.213171467541031
Содержание эмоции упрека в тексте: 0.05238959795499914
Содержание эмоции любви в тексте: 0.31253104986947766
Содержание эмоции ненависти в тексте: 0.38132762467369416
Содержание эмоции злорадства в тексте: 0.27289199613400822
Содержание эмоции благодарности в тексте: 0.82748466700990896
Содержание эмоции вознаграждения в тексте: 0.62957844439024831
Содержание эмоции раскаяния в тексте: 0.3251768148930981
Всего слов в тексте: 110709
Всего предложений в тексте:7719
Время работы алгоритма составляет: 75.0640115737915 секунд
```

Рис. 1. Вывод результата оценки тональности текста первого тома

Первый том был определен как положительный. В нем присутствуют все эмоции, особенно выражены радость, гнев, возмущение, жалость, восхищение, любовь, ненависть, благодарность и вознаграждение. Действия первого тома описывают события войны, однако в данном тексте содержится большое количество прилагательных и наречий, которые положительным образом сказываются на общей тональности.

Следующая итерация — это определение эмоциональной тональности текста у второго тома. Данный том освещает события общественной жизни. В томе автор описывает личные отношения героев и их переживания, затрагивает темы отцов и детей, дружбы, любви и поиска смысла жизни. На рисунке 2 представлен снимок экрана с определенной тональностью второго тома.

Данный том алгоритм классифицировал как сильно положительный. Здесь также преобладают эмоции радости, гнева, возмущения, любви, благодарности и вознаграждения.

```
Очки тональности текста: 1.313246819992291
Текст имеет сильно положительную окраску
Содержание эмоции радости в тексте: 0.5476884452354545
Содержание эмоции горя в тексте: 0.16594014117808903
Содержание эмоции гнева в тексте: 0.31726379323686454
Содержание эмоции страха в тексте: 0.22784527156576992
Содержание эмоции надежды в тексте: 0.2820122606549907
Содержание эмоции возмущения в тексте: 0.3207029671472912
Содержание эмоции жалости в тексте: 0.41699983663923923
Содержание эмоции облегчения в тексте: 0.14186592380510202
Содержание эмоции удовлетворения в тексте: 0.13068860859621517
Содержание эмоции разочарования в тексте: 0.0025793804328200364
Содержание эмоции радости за другого в тексте: 0.5476884452354545
Содержание эмоции гордости в тексте: 0.21236898896884968
Содержание эмоции восхищения в тексте: 0.5768010523872165
Содержание эмоции стыда в тексте: 0.09801645644716139
Содержание эмоции подтвердившегося страха в тексте: 0.22784527156576992
Содержание эмоции упрека в тексте: 0.07136285864135435
Содержание эмоции любви в тексте: 0.4986802170118737
Содержание эмоции ненависти в тексте: 0.1793804328200367
Содержание эмоции злорадства в тексте: 0.112912378446697
Содержание эмоции благодарности в тексте: 0.6124489497622671
Содержание эмоции вознаграждения в тексте: 0.67574342043042
Содержание эмоции раскаяния в тексте: 0.20395659762525042
Всего слов в тексте: 116307
Всего предложений в тексте:7994
Время работы алгоритма составляет: 79.61076784133911 секунд
```

Рис. 2. Вывод результата оценки тональности текста второго тома.

Текущая реализация алгоритма позволяет классифицировать текст, определяя общую оценку, кроме того, предоставляется возможность детализировать её и посмотреть вектор эмоций, которые преобладают в тексте. Скорость алгоритма является адекватной, что позволит проводить анализ различных сообщений и отзывов в режиме реального времени. Первый том был классифицирован за 1 минуту и 15 секунд, второй том за 1 минуту и 19 секунд. Время работы алгоритма начинается с инициализации всех переменных вместе с загрузкой данных и заканчивается с выводом отчета на экран.

Список источников

Brownlee J. 2017. A Gentle Introduction to the Bag-of-Words Model // Deep Learning for Natural Language Processing Tweet Share.

Steunebrink B. R., Dastani M. M. & Meyer, J-J. Ch. 2008. A Formal Model of Emotions: Integrating Qualitative and Quantitative Aspects // G. Mali, C. D. Spyropoulos, N. Fakotakis, N. Avouris (eds.), Proc. 18th European Conference on Artificial Intelligence (ECAI'08) (pp. 256–260). Greece/Amsterdam: Patras / IOS Press.

Интернет источник: Онлайн-сервис карта слов русского языка и выражений. https://kartaslov.ru/

Интернет источник: Морфологический анализатор pymorphy2 (библиотека для языка python). https://pymorphy2.readthedocs.io/en/latest/

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ИЗУЧЕНИЯ ФОНЕТИЧЕСКОЙ РЕДУКЦИИ СЛОВ В РУССКОЙ УСТНОЙ РЕЧИ[1]

М. В. Дайтер
mvdayter@gmail.com

Е. И. Риехакайнен
e.riehakajnen@spbu.ru

Санкт-Петербургский государственный университет

Фонетическая редукция слов (т. е. выпадение одного или нескольких звуков в составе слова) является характерной особенностью любой устной речи, но чаще всего встречается в наименее подготовленной ее разновидности — в спонтанной речи. Описанию редукции и выявлению причин ее возникновения в речи посвящены многочисленные исследования на материале различных языков (Земская 1973: 40–124; Ernestus 2000; Jurafsky et al. 2000; Spilková 2014; Стойка 2016 и др.); некоторые сформулированные в этих и других работах принципы учитываются при разработке автоматических синтезаторов речи по тексту (см, например, (Лобанов, Цирульник 2008)). Однако по-прежнему отсутствуют данные о том, в равной ли степени различные факторы влияют на появление редукции, вследствие чего предсказать, в каком случае слово будет редуцировано в речи, практически невозможно. Использование методов машинного обучения, на наш взгляд, может помочь получить ответы на данные вопросы, что в перспективе позволит усовершенствовать автоматические синтезаторы и системы распознавания устной речи.

Наше исследование, направленное на выявление факторов, способствующих фонетической редукции слов в русской устной речи, проводится на материале Корпуса русской устной речи (http://russpeech.spbu.ru/). Все записи в корпусе снабжены орфографической расшифровкой и акустико-фонетической транскрипцией (описание принципов транскрипции: http://russpeech.spbu.ru/transkrip.htm). Из этого корпуса для тестирования алгоритма были выбраны записи одной радиопередачи и двух телепередач общей продолжительностью звучания 93 минуты. С помощью программного обеспечения, реализованного на языке Python, был составлен перечень всех встретившихся в данном материале словоупотреблений. В ряде случаев за отдельное словоупотребление принимались т.н. неоднословные единицы (*потому что*, *то есть* и др.; список определялся по (Венцов, Грудева 2008)).

Набор данных для алгоритма был организован в таблицу, состоящую из 9648 строк (по количеству словоупотреблений). Первый столбец таблицы — это перечень словоупотреблений в орфографии, во втором столбце указана транскрипция, соответствующая тому, как это слово было произнесено в тексте. В остальных столбцах таблицы указывались факторы, которые могут оказывать влияние на возникновение редукции; за основу перечня были взяты признаки, рассмотренные в (Riekhakaynen 2013): количество слогов; частотность в русской устной речи (определялась по устному подкорпусу Национального корпуса русского языка (http://ruscorpora.ru/new/)); частотность словоформы в том тексте, в котором она встретилась в материале; часть речи; наличие/отсутствие фонетической редукции в предшествующем слове; наличие/отсутствие редукции самого словоупотребления (целевая переменная).

В ходе работы решалась задача бинарной классификации, были задействованы следующие алгоритмы: случайные леса, сверхслучайные деревья, логистическая регрессия. Сверхслучайные деревья, в отличие от случайных лесов, выбирают точку разделения случайно. Для реализации алгоритмов использовалась библиотека scikit-learn языка Python. Случайные леса и сверхслучайные деревья были использованы для отбора признаков словоформ, оказывающих наибольшее влияние на факт редукции, логистическая регрессия, построенная на

[1] Исследование выполняется при поддержке гранта РФФИ № 19-012-00629 «Алгоритмы восстановления редуцированных форм: роль системы языка».

основе выделенных признаков, — для выявления зависимости между этими признаками и целевой переменной. Признаки отбирались методом рекурсивного исключения признаков, оптимизируемой метрикой была сбалансированная точность в силу несбалансированности классов в наборе данных (32% редуцированных словоформ). Качество работы алгоритмов оценивалось с помощью таких метрик, как точность и сбалансированная точность.

При работе с полным набором данных с помощью сверхслучайных деревьев были выявлены следующие признаки: количество слогов, частотность и такие части речи, как прилагательное, предлог и вводное слово (части речи были оформлены в виде бинарных переменных). Чем больше количество слогов и чем выше частотность, тем выше вероятность редукции. Если слово является прилагательным или вводным словом, оно скорее редуцируется. С предлогами ситуация обратная. Список, выделенный с помощью алгоритма случайного леса, отличается: количество слогов, частотность, прилагательное и вводное слово. Стоит отметить, что при рекурсивном исключении признаков признак, является ли слово предлогом, получил ранг 2. Точность и сбалансированная точность при округлении до двух знаков после запятой при этом равны у обоих алгоритмов: 80% и 75% соответственно.

Литература

Венцов А. В., Грудева Е. В. 2008. Частотный словарь словоформ русского языка (проект). Череповец: Изд-во ЧГУ.

Земская Е. А. (ред.). 1973. Русская разговорная речь. М.: Наука.

Лобанов Б. М., Цирульник Л. И. 2008. Моделирование внутрисловных и межсловных фонетико-акустических явлений в синтезаторе русской речи по тексту // Идеи и методы экспериментального изучения речи: сб. ст. памяти проф. Л. А. Чистович и проф. В. А. Кожевникова / отв. ред. Л. В. Бондарко, В. В. Люблинская. СПб.: Ф-т филологии и искусств. С. 47–63.

Стойка Д. А. 2016. Редуцированные формы русской речи: лингвистический и экстралингвистический аспекты: дис. … канд. филол. н. СПб.: Санкт-Петербургский государственный университет.

Ernestus M. 2000. Voice Assimilation and Segment Reduction in Casual Dutch. A Corpus-Based Study of the Phonology-Phonetics Interface. Utrecht: Landelijke Onderzoekschool Taalwetenschap.

Jurafsky D., Bell A., Gregory M., Raymond W. D. 2000. Probabilistic Relations between Words: Evidence from Reduction in Lexical Production // Frequency and the Emergence of Linguistic Structure / J. Bybee, P. Hopper (eds.). Philadelphia PA: John Benjamins, pp. 229–254.

Riekhakaynen E. 2013. Reduction in Spontaneous Speech: How to Survive // New Perspectives on Speech in Action: Proceedings of the 2nd SJUSK Conference on Contemporary Speech Habits / J. Heegård, P. J. Henrichsen (eds.). Frederiksberg: Samfundslitteratur Press, pp. 153–165.

Spilková H. 2014. Phonetic Reduction in Spontaneous Speech: An Investigation of Native and Non-Native Production. Trondheim: Norwegian University of Science and Technology.

АВТОМАТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ МЕНТАЛЬНЫХ ДЕЙСТВИЙ, КОНСТИТУИРУЮЩИХ РЕЧЕВЫЕ ЖАНРЫ НАУЧНОГО ТЕОРЕТИЧЕСКОГО ТЕКСТА[1]

Д. А. Девяткин
ФИЦ «Информатика и управление» РАН
lonelykeen@gmail.com

Л. А. Каджая
kadzhaya@psu.ru

В. А. Мишланов
vmishlanov@yandex.ru

В. А. Салимовский
salimovskt@rambler.ru

Пермский государственный национальный исследовательский университет

Под речевыми жанрами (РЖ) мы понимаем относительно устойчивые формы ментальной деятельности на ступени ее реализации посредством системы речевых действий в тексте как единице общения. (Ср. во многом близкие трактовки жанра в работах В. В. Дементьева, К. А. Седова, К. А. Долинина.) Рассмотрение РЖ в таком аспекте предполагает применение специального — лингвокогнитивного — метода анализа, который заключается в проецировании структуры определенного вида ментальной деятельности на массив воплощающих ее текстов и последующем соотнесении основной целеустановки отдельного текста с тем или иным участком данной структуры. Это позволяет выделить тексты, реализующие один и тот же набор типовых ментальных действий и, следовательно, представляющие определенный РЖ.

Как известно, научное познание имеет эмпирический и теоретический уровни. Эмпирическая деятельность заключается в применении концептуального аппарата науки к изучаемым объектам в ходе наблюдения и эксперимента, а теоретическая — в преобразовании, развитии самого этого концептуального аппарата (В. С. Швырев). В структуре теоретического исследования принято различать три основные фазы: 1) создание отправной теоретической онтологии (картины действительности), 2) построение теории на уже найденном основании, 3) применение теории для объяснения определенной группы явлений (И. В. Кузнецов, М. В. Мостепаненко и др.). Эти фазы, представленные относительно устойчивыми системами познавательных действий, объективируясь в научно-речевых произведениях, определяют их жанровое своеобразие. Так, первой из указанных фаз соответствует постановочный текст, второй фазе — жанр «экспликация научного понятия» (развитие основного понятия теории есть ее построение), третьей фазе — верификационный текст (проверка теории экспериментальным методом).

Основными познавательно-коммуникативными действиями, конституирующими постановочный текст, являются: анализ наличного знания (идей, теорий, результатов исследований), обнаружение его неполноты или рассогласованности, постановка проблемы, формулирование новой идеи, раскрытие ее эвристического потенциала. Эти действия осуществляются с использованием определенного набора языковых и речевых единиц. Например, анализ научных теорий предполагает широкое использование конструкций прямой и косвенной речи, вводных словосочетаний и предложений со значением отнесенности к автору речи, других авторизующих средств; интерпретация тех или иных положений теории проявляется на поверхностно-речевом уровне в активизации средств текстовой категории оценочности. Для речевого жанра «экспликация научного понятия» характерны познавательно-коммуникативные действия, раскрывающие существенные свойства явлений в раз-

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 17-29-07049 офи_м.

личных планах — структурно-морфологическом, типологическом, функциональном и др. При совершении этих действий реализуются модели предложений с предикатами бытийной, квалифицирующей, классификационной, причинно-следственной, компаративной семантики. Продуцирование верификационного текста предполагает включение в свой состав таких познавательно-коммуникативных действий, как формулирование проверяемой гипотезы, описание методики эксперимента, анализ опытных данных для подтверждения или опровержения гипотезы. Каждое из этих действий совершается с использованием своего набора языковых единиц.

Таким образом, жанр как форма текстовой деятельности характеризуется особой речевой системностью (Кожина 2014), создающей лингвистическую базу для компьютерного анализа текста. В отличие от исследований, решающих задачи автоматического реферирования научных статей и обычно охватывающих лишь вводную их часть (см., напр., многочисленные работы, ориентирующиеся на жанроведческую теорию Дж. Свейлза, Swales 1990), наш анализ распространяется на все текстовое пространство изучаемых РЖ.

Метод распознавания воплощаемых в речи ментальных действий основывается на концепции реляционно-ситуационного анализа (РСА) текста (Осипов 2011). В работе с РСА-моделью используются шаблоны, задающие множество неоднородных семантических сетей (НСС), характерных для текстов с определенным общим свойством. При создании шаблона лингвист, опираясь на логико-психологическую характеристику выявляемых ментальных действий, по возможности полно и точно описывает языковые средства их воплощения в тексте, после чего программист формализует лингвистическое описание.

Применяемый метод распознавания когнитивных действий является трехэтапным. На первом этапе выполняется автоматический анализ текста с использованием морфологического анализатора MyStem (Segalovich, Titov 1997–2014) и синтаксического анализатора U-Pipe (Straka et al. 2016). При построении НСС используется семантический анализатор (Shelmanov et al. 2017). На втором этапе НСС высказываний соотносятся с контекстно-свободными шаблонами. На третьем этапе для уточнения результатов применяется метод классификации последовательностей: используются композиции деревьев решений со скользящим окном, классификатор, основанный на условных случайных полях, и рекуррентные нейронные сети.

Осуществленное нами ранее автоматическое распознавание ментальных действий, объективируемых в научных эмпирических текстах, имеет средний показатель оценки качества 0,84 (от 0,73 до 0,96 для разных действий). В докладе будут представлены результаты идентификации ментальных действий, конституирующих речевые жанры теоретического текста.

Литература

Кожина М. Н. 2014. Речеведение. Теория функциональной стилистики. М.: Флинта: Наука.

Осипов Г. С. 2011. Методы искусственного интеллекта. М.: Физматлит.

Segalovich I., Titov V. 1997–2014. MyStem [Software].

Shelmanov A. et al. 2017. Semantic Role Labeling with Neural Networks for Texts in Russian // Annual International Conference «Dialog». Moscow: Russian State University for Humanities, vol. 1, pp. 245–257.

Straka M. et al. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing // Proceedings of the tenth international conference on language resources and evaluation (LREC 2016), pp. 4290–4297.

Swales J. M. 1990. Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press.

ВОЗМОЖНОСТЬ ИСПОЛЬЗОВАНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОМАТИЧЕСКОГО ВЫЯВЛЕНИЯ СТЕРТЫХ МЕТАФОР

О. В. Донина
Воронежский государственный университет
olga-donina@mail.ru

Д. С. Дмитриев
DSR Corporation
biomans@mail.ru

Для апробации возможности использования нейронных сетей при выявлении стертых метафор был выбран один из детально разработанных в теоретической литературе подходов, а именно — криптоклассный анализ. Наш выбор был обусловлен наличием формализованного алгоритма выявления метафоры, описанного в Борискина 2011. В рамках данного подхода под криптоклассом понимается «скрытая языковая категория, характеризующаяся тождеством семантического признака составляющих ее слов и проявляющаяся в особенностях лексико-синтаксической сочетаемости» (Борискина 2003: 14–15), а под метафоронимом — абстрактные сущности, у которых признак класса нельзя выявить явно, а только посредством анализа его сочетаемости, которая в данном случае метафорична. Так, в случае с классом Res Acutae («Острое») мы можем говорить о том, что метафоронимами здесь выступают такие имена, как «язык» и «ум», так как их сочетаемость со словом «острый»: «острый язык», «острый ум» — указывает на способность этих имен проявлять свойство остроты, т. е. проявлять категориальный признак указанного класса, и именно подобные сочетания мы рассматриваем как примеры стертой метафоры.

Изначально для решения поставленной задачи было определено несколько возможных вариантов создания автоматической бинарной классификации, а именно:

1. Поиск по шаблонам,
2. Использование грамматик,
3. Машинное обучение (и возможное последующее использование XAI — explainable AI),
4. Кластеризация имеющихся примеров и выявление отличительных характеристик для последующего использования их классификатором,
5. Синтаксический парсинг и классификация с учетом его данных.

В рамках выступления будут рассмотрены причины выбора машинного обучения для создания классификатора.

Нами был подготовлен датасет, насчитывающий 389 857 размеченных вручную примеров. Первым шагом стало определение BaseLine — случайного стратифицированного (с учётом представительности классов) выбора метки, который составил 0,73.

Далее мы провели ряд экспериментов и сравнили полученные результаты. Были использованы такие методы, как наивный Байес (по отдельным словам), логистическая регрессия (хорошо работающая при решении задач выявления бинарной классификации), а также сверточные нейронные сети.

При использовании логистической регрессии было протестировано влияние на выявление стертой метафоры изменения следующих параметров: 1) наличие/отсутствие лемматизации, 2) наличие/отсутствие стемминга, 3) наличие/отсутствие стоп-слов в корпусе, 4) разбиение мешка слов (bag of words) на слова: униграммы (liblinear-unigram) / биграммы, 5) разбиение мешка слов на символы: 2-символьное разбиение (2-char-gram) и 3-символьное разбиение (3-char-gram). После всех экспериментов с логистической регрессией было показано, что лучшие результаты в рамках этой модели на наших данных были получены при 3-символьном разбиении мешка слов, с учетом стоп-слов и без лемматизации/стемминга (f1-score (avg) = 0,85).

Следующим шагом стало использование нейронных сетей. Мы использовали фреймворк Keras с Tensorflow в качестве бэкэнда. На вход нейронной сети поступала матрица подсчета

токенов, в которую была преобразована коллекция текстовых документов при помощи CountVectorizer. Хотя для логистической регрессии мы пробовали разные варианты мешков слов (bag of words), для искусственных нейронных сетей мы решили остановиться на 2-символьном разбиении. Опишем строение сверточной нейронной сети, которое мы взяли за базовое: она состоит из шести слоев, два из которых — dropout (исключения), позволяющие избежать переобучения модели; в качестве функции активации используются ReLU и Sigmoid; в качестве оптимизатора используется Adam, а в качестве loss (потери) — binary_crossentropy.

После ряда экспериментов с различными признаками последовательной сверточной нейронной сети удалось достичь результата f1-score (avg) = 0,88. Проведенный анализ публикаций в рамках аналогичной проблематики показал, что достигнутые нами результаты не уступают state-of-the-art исследованиям в данной области.

Таким образом, на примере выявления стертой метафоры была показана возможность реализации современного теоретического лингвистического подхода методами машинного обучения.

References

Борискина О. О. 2011. Криптоклассы английского языка. Воронеж: Истоки.

Борискина О. О. 2003. Теория языковой категоризации: национальное языкознание сквозь призму криптокласса. Воронеж: Воронежский государственный университет.

О НЕКОТОРЫХ КОМПОНЕНТАХ «КАРТИНЫ МИРА» СОВРЕМЕННОГО ПОДРОСТКА

К. А. Елистратова

xenos88@mail.ru

МОБУ СОШ «Муринский ЦО №4»

Становление личности в современных условиях можно охарактеризовать как процесс, находящийся под влиянием ведущих тенденций — глобализации и модернизации (Горская 2005), что закономерно находит отражение в картине мира современного школьника. **Предметом настоящего исследования** является картина мира современного старшеклассника как целостная система, а также ее отдельные структурные компоненты с учетом когнитивных различий в зависимости от профиля образования. В качестве **объекта исследования** (участники ассоциативного эксперимента) выступали две экспериментальные группы — обучающиеся 11 классов с а) углубленным изучением естественно-научных дисциплин; б) дисциплин гуманитарного профиля.

Нас интересовали сферы, определяющие социализацию и самоидентификацию современного молодого человека («Интернет», «Политика», «Мораль»), а также сферы на базе образовательного профиля, определяющего когнитивные составляющие картины мира обучающихся («Физика/Математика», «Литература»). Кроме того, в случаях со сферами-стимулами «Политика» и «Мораль» важно было выяснить, в какой мере для информантов «осмысленны или, напротив, пусты слова, обозначающие ценностно нагруженные понятия» (Фрумкина 2006: 195–196).

Остановимся подробнее на нескольких стимулах.

Стимул сферы «Политика» — **«Закон»** у обучающихся естественно-научного класса обнаружил следующие устойчивые ассоциации: «порядок» — в данном случае имеется отсылка к известному детективному сериалу федерального канала (синтагматический характер механизма ассоциаций: сочинительная связь); «статья» — ассоциация по смежности (часть — целое).

Абсолютная доминанта ответов обучающихся из класса гуманитарного профиля — ассоциат «право» (14). В средней общеобразовательной школе «Муринский центр образования №4» большое внимание уделяется гражданско-правовому образованию — в частности предмету «Право»; в центре образования продуктивно работает ученический правовой клуб «Фемида». В этом смысле цепь «стимул — реакция» логически оправданна. Стимул **«государство»** существует в сознании школьников не как абстрактное, философское понятие, а как представление о российском государстве. Ассоциации к слову «государство» по сути представляют собой ассоциации к словосочетанию «российское государство», то есть концепт «государство» не существует как некий абстрактный феномен, но осуществляет себя через пример родного национального государства. Данный тезис основывается на том очевидном наблюдении, что в процессе ассоциирования на тему «государство» практически всегда происходит смещение сознания в сторону тех фактов и реалий, которые связаны с политикой, историей, личностями именно российского государства. Государство подразумевает в первую очередь наличие властной структуры, организующей жизнь людей государства: власть (4) и представлением о личностном характере государственной власти, которую реализует президент: в нашем случае по-прежнему срабатывает устойчивая связь государство — Путин. Государство включает в себя представление о территориальном разделении мира: реакции «страна», «граница». Слово «государство» этимологически подразумевает личностную власть государя на территории своей страны: отсюда прецедентная реакция «Макиа-

велли» (1): ср. Н. Макиавелли «Государь» (гражданско-правовое образование в гимназии). Общая негативная маркировка реакций сферы «Политика» позволяет интерпретировать данные эксперимента следующим образом: недовольство обучающихся существующим политическим строем: единичный ассоциат прецедентного толка «Платон» (1) с его идеалистической концепцией «Государства» можно рассматривать как реакцию, данную на основе антонимических механизмов ассоциирования.

В сфере **«Интернет»** нас будут интересовать исключительно индивидуально-личностные реакции, позволяющие выявить оценочное наполнение стимулов в сознании современной молодежи, ибо абсолютное большинство полученных реакций стандартны и устойчивы. При общей позитивной тональности всего, что имеет непосредственное отношение к сфере «Интернет», выделяется ряд единичных реакций, коннотируемых отрицательно: все — в гуманитарном классе: не жизнь (1), иллюзия (1), пустота (1), грезить (1).

Различия в уровне профильного образования со всей определенностью проявили себя в реакциях на стимулы сфер **«Литература» и «Физика / Математика»**. В естественно-научном классе почти половина информантов либо не дала ответа на стимул «**Бродский**», либо интерпретировала его как героя романа (вполне возможно, что по фонетическому / словообразовательному сходству звукообраза фамилии: Бродский — ср. Ленский): наиболее частотный ответ из полученных «Нобелевская премия» (3): знание факта биографии (парадигматический тип ассоциирования). По тому же принципу фонетического сходства дана единичная реакция «Троцкий» (в оригинале: Тротский). В гуманитарном классе — набор реакций весьма разнообразен и обширен (что подтверждает знание учащимися текстов поэта): абсолютное большинство ответов — оценочные реакции.

Представляются весьма любопытными результаты сравнения ассоциаций, данных на стимулы сферы **«Физика / Математика»**: для обучающихся естественно-научного класса характерна абсолютная доминанта устойчивых реакций. На весь реестр приходится лишь единичная личностная реакция — Шелдон Купер (1) — сериал «Теория большого взрыва». В ответах обучающихся гуманитарного класса, помимо основных устойчивых, наблюдается целый ряд единичных, но оценочных, личностно окрашенных. Среди ответов на алгебраический «интеграл» выделяется единичный литературный ассоциат — Замятин «Мы» (1): отсылка к космическому кораблю «Интегралу» в романе.

Среди реакций на стимул «жизнь» следует отметить ряд ассоциатов интертекстуального толка: кальдероновский сон (1), блоковский фонарь (1), шекспировский театр (1) вкупе с самим Шекспиром (1), арбенинское зву-чать (1).

Резюмируя все вышесказанное, отметим следующее: в когнитивных и ценностно-смысловых компонентах субъективных картин мира обучающихся средней общеобразовательной школы «Муринский центр образования №4» гуманитарного и естественно-научного классов существуют вполне определенные различия: а) когнитивные и ценностно-смысловые компоненты субъективной картины мира учащихся гуманитарной гимназии обладают большей сложностью, многокомпонентностью, многоаспектностыо, ассоциативной сложностью, расщеплением на реальный и идеальный планы; б) в картине мира обучающихся гуманитарного класса мир характеризуется большей эмоциональной притягательностью. Специфика ответов обучающихся гуманитарного класса — абсолютное большинство несистемных, а индивидуально-личностных ассоциаций. Кругозор и общая осведомленность учащихся-гуманитариев в событиях политической жизни значительно выше. В их ответах значительна доля реакций интертекстуального толка; в) общая особенность ответов современных старшеклассников — ряд ответов, вписывающихся со всей определенностью в тенденции современной массовой культуры: единичный ответ учащейся ЖГГ: как пример — следующая пара: «жизнь» — футбол (1), в котором усматривается вполне определенная отсылка к телевизионной рекламе: «Живи — футболом, дыши футболом, пей Coca-cola».

Литература

1. Горская Р. С. 2005. Влияние углубленного изучения английского языка на когнитивные и эмоциональные компоненты картины мира: Автореф. дис. ... канд. психол. наук. СПб.

2. Фрумкина Р. М. 2006. Психолингвистика. М. С. 195—196.

УРОВЕНЬ АКТИВАЦИИ VS. СПЕЦИАЛЬНЫЕ ОГРАНИЧЕНИЯ ПРИ МАТЕМАТИЧЕСКОМ МОДЕЛИРОВАНИИ РЕФЕРЕНЦИАЛЬНОГО ВЫБОРА

Д. А. Залманов
dm.zalmanov@gmail.com
Институт языкознания РАН

А. А. Кибрик
aakibrik@gmail.com
Институт языкознания РАН
МГУ им. М. В. Ломоносова

Уровень активации (УА) референта в рабочей памяти говорящего — важнейший предиктор того, какие типы референциальных выражений приемлемы для упоминания референта в текущей позиции в дискурсе. Есть и другие факторы, влияющие на референциальный выбор. Например, хорошо известна роль фильтра референциального конфликта, который отбраковывает редуцированные упоминания даже при высоком УА, если такое упоминание может вызвать референциальную неоднозначность. В данной работе мы исследуем еще одну группу ограничений, влияющих на референциальный выбор и не имеющих прямого отношения к УА.

Рассмотрим пример такого ограничения:

(1) Completion of the acquisition is subject to execution of a definitive agreement, approval by all three companies' boards and the approval of [Applied Solar's]$_{i1}$ shareholders. An [Applied Solar]$_{i2}$ spokesman said completion is expected at the end of the year…

В (1) расстояние между анафором *i2* и антецедентом *i1* мало, активация референта высока, и при другом построении предложения такой УА с большой вероятностью создал бы условия для того, чтобы анафор был выражен местоимением (ср. *«One of [its]$_{i2}$ spokespeople…»*). Однако в данном примере синтаксическая роль препозитивного субстантивного модификатора (*noun premodifier* в терминологии Biber et al. 1999: 588) исключает возможность использовать местоимение:

(1') ^{OK}A [company]$_{i2}$ spokesman…

(1'') *A [it]$_{i2}$ spokesman…

Мы предполагаем, что математическая модель референциального выбора, которая на основе признаков, наблюдаемых для некоторого упоминания референта в дискурсе, оценивает УА референта и далее предсказывает тип референциального выражения, должна разрабатываться на данных, очищенных от эффектов, подобных эффекту в (1). Если подобные ограничения не устраняются из набора данных, а учитываются в модели как факторы, то это должно приводить к ложным выводам о взаимосвязи между факторами, УА и типами референциальных выражений.

На материале корпуса WSJ MoRA 2020 мы сравнили результат предсказания формы референциального выражения алгоритмами машинного обучения в 2-классовой задаче (местоимение vs. полная ИГ) для двух наборов данных:

Набор 1: Анафорами являются все ИГ, для которых в тексте можно определить эксплицитно выраженный антецедент и которые имеют в качестве вершины существительное или личное или посессивное местоимение (схема аннотации описана в Kibrik et al. 2016).

Набор 2: Из набора 1 исключены следующие классы употреблений (используемые примеры, как и пример (1) выше, также получены из корпуса WSJ MoRA 2020):

1. К репрезентации референта в текущем дискурсе добавляется новая информация. В этом случае можно предполагать, что употребление именно дескрипции (а не местоимения) может быть обусловлено не столько степенью активации референта, сколько необходимостью сообщить новую информацию о референте (*the company* vs. ***the Minneapolis glass products and aluminum window maker***).

2. Референт упоминается дважды в пределах одной клаузы в разных ролях, при этом рефлексив неграмматичен: *[First Tennessee]$_i$, assisted by IBM, said [it]$_i$ will attempt to place the employees* ***[within the company]$_i$****, IBM or other companies in Memphis*.
3. Референт в принципе не может быть прономинализован (по крайней мере, не может быть обозначен через личные и посессивные местоимения, употребление которых предсказывает модель): ***from Bataan province*** → ** from it*
4. Анафор находится в позиции препозитивного субстантивного модификатора (см. (1) выше).
5. Анафор находится внутри цитируемой прямой речи: выбор выражения не определяется предшествующим дискурсом.

Количество примеров категорий, перечисленных выше, совокупно составило 684 при общем количестве примеров в корпусе — 2391 (набор 1). Можно видеть, что это существенная часть данных (28,6%), следовательно, она создает погрешность при построении модели.

В таблице 1 показано сравнение результатов выполнения задач классификации тремя разными алгоритмами для наборов 1 и 2.

Таблица 1. Сравнение результатов классификации: аккуратность предсказания, %

	Логистическая регрессия	Бэггинг	Бустинг
Набор 1 (2391 единиц)	88,04	86,50	87,70
Набор 2 (1707 единиц)	88,67	88,10	87,81

Результаты классификации для набора 2 оказались лучше, чем для набора 1, для всех трех используемых алгоритмов (разница составляет до 1,6%). Очевидно, исключение из обучающей выборки категорий, перечисленных выше, повышает качество модели. Таким образом, в данном исследовании мы указываем на существенный дополнительный класс специальных ограничений референциального выбора, недостаточно изученный в предшествующих исследованиях.

Литература

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. Longman grammar of spoken and written English. Harlow: Pearson Education limited.

Kibrik A. A., Khudyakova M. V., Dobrov G. B., Linnik A., Zalmanov D. A. 2016. Referential choice: Predictability and its limits // Frontiers in psychology, Frontiers Research Foundation (Switzerland), Vol. 7, № 1429.

ЦИФРОВОЕ ИЗМЕРЕНИЕ «НАИВНОЙ» ГЕОГРАФИИ[1]

Н. Л. Зелянская — zelyanskaya@gmail.com
К. И. Белоусов — belousovki@gmail.com
Р. В. Чумаков — chumakoff.r.v@gmail.com
Пермский государственный национальный исследовательский университет

«Наивная» география — междисциплинарная область исследований, посвященных анализу геопространственных представлений информантов (Saarinen 1988, 1990; Didelon 2011; Zelyanskaya 2017). В качестве основного метода исследования выступает эксперимент, направленный на визуализацию ментальных репрезентаций геопространства. В ходе эксперимента респонденты на листе формата А3 рисовали карту России, на которой отмечали важные для них географические объекты России и страны, с которыми граничит Россия. Возле всех нанесенных на карту географических объектов нужно было записать связанные с ними ассоциации (примеры «наивных» карт доступны на сайте https://naivemaps.ru/ — раздел «Материал»). Результатом визуализации становится «наивная» карта («sketch map»).

Наши исследования в области «наивной» географии, осуществляемые с 2008 г., прошли три этапа:

1. Работа с «бумажными» картами; описание общих закономерностей организации «наивных» карт; частотный анализ геообъектов, наносимых на карты информантами (далее — Ии.); анализ кодов (пространственных, графических, языковых). Общий объем материала составил около 500 карт, собранных в Москве, Санкт-Петербурге, Сибири (Барнауле, Бийске), Южном, Среднем и Северном Урале (Оренбурге, Уфе, Перми) и Калининграде.

2. Разметка «бумажных» карт в специально созданном приложении (разработчик Д.А. Баранов) с целью сопоставления расположения объектов на «наивных» картах и сравнения их с эталоном (реальной картой). Для исследования были отобраны десять наиболее частотных российских городов, представленных на картах, собранных в трех регионах: Алтайском крае, Оренбургской области и Пермском крае. С помощью разработанной методики (Зелянская 2016: 130–134) выявлялись пропорции между геообъектами на «наивных» картах, представляющих три региональные «точки зрения» на пространство России. Результатом вычислений стали три региональные матрицы расстояний между геообъектами. Данный подход позволил создать визуальные модели структуры пространства, реконструированного из каждой его точки в контексте сжатия/расширения относительно эталона (примеры см. https://naivemaps.ru/?cat=13). Однако упомянутый редактор имеет ограничения: релевантная разметка может проводиться только с «точечными» геообъектами (в первую очередь, городами), что исключает из объектов моделирования и интерпретации «протяженные» геообъекты (реки, горы, моря и др.) и такие характеристики, как цвет, размер и т. п. любых, в т.ч. «точечных», объектов.

3. Современный уровень развития компьютерных технологий позволяет осуществлять исследования в области «наивной» географии в цифровой среде на всех этапах работы. Для сбора «наивных» карт разработан веб-редактор карт (https://creativemaps.studio/), в котором есть необходимый набор инструментов и иконок в векторном формате для рисования цифровой карты в браузере.

Сама карта экспортируется из приложения в формате .csv. Данные содержат поля: uuid (универсальный идентификатор фигуры), x (координаты фигуры по оси X), y (координаты фигуры по оси Y), size (диаметр описанной окружности, вмещающей фигуру), tool (название фигуры), color (цвет фигуры), font_color (цвет текста фигуры), emoji (список эмотиконов, прикрепленных к фигуре), font_size (размер шрифта названия фигуры), transparency (прозрач-

[1] Исследование выполнено при финансовой поддержке РНФ (проект № 20-18-00336).

ность фигуры), name (отображаемое на холсте название фигуры), description (описание фигуры — т. е. мнение пользователя об этом объекте), layout (слой, на котором размещается фигура).

Данные о каждом объекте на цифровых «наивных» картах позволяют программно анализировать сложные мультимодальные объекты, сочетающие пространственно-графические коды (размер, цвет, координаты, расположение), языковые коды (вербальные ассоциации, наименования геообъектов), временные параметры (последовательность нанесения объектов на карту, интервалы времени между нанесением объектов) и г) частотные параметры (встречаемость объектов в выборках «наивных» карт). Включение в модель географических и социально-демографических данных об объектах (координаты географических объектов, численность населения) позволяет выявлять скрытые зависимости и структуры «наивных» карт.

Исследования в области «наивной» географии, проводимые в настоящее время на материале цифровых «наивных» карт, позволяют перейти от сугубо качественных описаний и методов описательной статистики к моделированию, основанному на методах многомерной статистики, геопространственного анализа, машинного обучения и визуальной аналитики. Появившиеся технологии позволяют значительно расширить географию сбора данных (цифровых «наивных» карт), что необходимо как для использования методов машинного обучения, так и для моделирования пространства страны из разных ее регионов.

Литература

Зелянская Н. Л., Баранов Д. А., Белоусов К. И. 2016. Наивная география и топология геоментальных карт // Социо- и психолингвистические исследования. Вып. 4. С. 126–136.

Didelon C. et al. 2011. A world of interstices: A fuzzy logic approach to the analysis of interpretative maps / C. Didelon S. de Ruffray M. Boquet N. Lambert // The Cartographic Journal. Vol. 48, pp. 100–107.

Saarinen T. F. 1988. Centering of Mental Maps of the World. National Geographic Research. № 4(1), pp. 112–127.

Saarinen T. F., MacCabe C. 1990. The world image of Germany Erdkunde. № 44, pp. 260–267.

Zelyanskaya N., Belousov K., Ichkineeva D. 2017. Naive geography and geopolitical semiotics: The semiotic analysis of geomental maps of Russians // Semiotica. Vol. 2017. Iss. 215, pp. 235–253. DOI: https://doi.org/10.1515/sem-2016–0231.

МЕТОД ОЦЕНКИ СЛОЖНОСТИ ТЕКСТА, ОСНОВАННЫЙ НА ОНТОЛОГИЯХ[1]

В. В. Иванов
nomemm@gmail.com
Университет Иннополис

М. И. Солнышкина
mesoln@yandex.ru
Казанский федеральный университет

Основной на сегодняшний день подход к оценке сложности текстов — использование формулы Флеша-Кинкейда (Flesch 1948), учитывающей всего два параметра — среднюю длину предложений и среднюю длину слов (в слогах) текста. Этот подход привлекателен своей простотой, однако он связан лишь с поверхностной сложностью текста. В ряде работ рассматривались и другие параметры, например, доля абстрактной лексики, частотность слов. Но и эти параметры не репрезентируют глубинную концептуальную сложность текстов. Нами предложен подход, основанный на использовании тезаурусов и онтологий и претендующий на учет того, насколько сложными являются взаимоотношения концептов, встречающихся в тексте.

Общая идея состоит в следующем. Берется некоторая лексическая онтология (это может быть тезаурус или база знаний, такая как Википедия в формате DBpedia), и находятся все слова, присутствующие и в тексте, и в онтологии. В данной работе мы используем тезаурус RuThes-Lite (Loukachevitch et al. 2014), хорошо известный открытый ресурс, содержащий более 100 000 слов русского языка. Слова текста предварительно приводятся в основную форму с помощью программы TreeTagger (TreeTagger). Таким образом осуществляется отображение текста на онтологию. Далее, множество полученных в итоге отображения узлов онтологии вместе со связями между ними образует граф. Мы рассчитываем ряд стандартных характеристик полученного графа, отражающих его сложность. Утверждается, что эти характеристики отражают и сложность текста.

Приведем перечень характеристик: число вершин в графе (Co); число связных компонент (NC); средний размер компоненты (CS); максимальна степень вершин (MND); общее число связанных вершин (TCN); доля связанных вершин в общем числе вершин в графе (TCN/Co); длина среднего кратчайшего пути (ASP).

Описанный метод реализуется для русского языка впервые. Для поверки предложенного подхода нами взяты тексты заведомо разной концептуальной сложности — учебники для разных классов средней школы по одному предмету и одного и того же автора. Это две линейки учебников по обществознанию для 5 — 11 классов под редакцией А.Ф. Никитина и Л.Н. Боголюбова. Подробно корпус учебников описан в Ivanov et al. 2018. Для получения статистически значимых результатов мы делим учебники на сегменты размером S = 200, 500, 1000 и 2000 слов и вычисляем характеристики графов для каждого сегмента, как отдельного текста с последующим усреднением.

Далее для каждого размера сегментов вычисляются коэффициенты корреляции Пирсона между рассчитанными параметрами графов и номером класса как показателем сложности. Результаты приведены в таблице 1. Жирным шрифтом выделены результаты с корреляцией, большей или равной 0,7, что обычно рассматривается как высокая корреляция.

Выделяются два параметра: ‘общее число связанных вершин’ и ‘доля связанных вершин в общем числе вершин в графе’. Обе характеристики графов имеют высокую степень корреляции со сложностью текста, причем независимо от размера сегментов, так что они могут быть рекомендованы для оценки концептуальной сложности текстов. Также как параметры, характеризующие сложность текстов, можно использовать еще два параметра: ‘число связных компонент’ и ‘средний размер компонент’, однако высокая корреляция отмечается лишь для текстов (сегментов) определенного размера.

[1] Исследование выполнено при поддержке РНФ, грант № 18-18-00436.

Таблица 1. Корреляция параметров графа со сложностью текста

S	Co	NC	CS	MND	TCN	TCN/Co	ASP
200	0.45	**0.86**	0.32	0.53	**0.84**	**0.76**	0.27
500	0.62	**0.86**	0.68	0.45	**0.88**	**0.79**	-0.25
1000	0.56	0.38	**0.73**	0.40	**0.85**	0.69	-0.36
2000	**0.88**	0.40	**0.81**	0.68	**0.86**	**0.70**	-0.54

Схожее по идее исследование для английского языка было проведено в работе Štajner, Hulpus 2018. Однако во многих аспектах эти исследования отличаются. В работе Štajner, Hulpus 2018 для сопоставления используются искусственно упрощенные тексты, мы используем учебники для разных классов — естественные тексты разной сложности. Далее, в работе Štajner, Hulpus 2018 единицей анализа является предложение или абзац, у нас же значительно более длинные сегменты текста. В Štajner, Hulpus 2018 в качестве базы знаний используется DBpedia, у нас же RuThes-Lite. Мы рассматриваем все части речи, в Štajner, Hulpus 2018 же рассматриваются только существительные. Это затрудняет прямое сопоставление наших результатов. Все же можно отметить, что, согласно обеим работам, параметр ‘число связных компонент’ отражает сложность текстов. Но между нашими результатами есть и существенные расхождения. Параметр ‘максимальна степень вершин’ по нашим данным имеет положительную корреляцию со сложностью текстов, в то время как по данным Štajner, Hulpus 2018 — отрицательную. Последнее, вероятно, связано именно с использование в работе Štajner, Hulpus 2018 искусственно упрощенных текстов.

В дальнейшем предполагается реализовать предложенный метод оценки сложности текстов на основе онтологии DBpedia-Russian. Данный подход может быть полезен для оценки реальной, а не поверхностной сложности текстов, что важно в образовательном процессе, в том числе при выборе учебников для самообразования.

Литература

Flesch R. 1948. A new readability yardstick // Journal of Applied Psychology, 32. Pp. 221–233.

Štajner S., Hulpus I. 2018. Automatic assessment of conceptual text complexity using knowledge graphs // Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics. Pp. 318–330.

Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I. 2014. Ruthes-lite, a publicly available version of thesaurus of Russian language Ruthes // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». Т. 13. М.: Изд. центр РГГУ. С. 340–350.

Ivanov V. V., Solnyshkina M. I., Solovyev V. D. 2018. Efficiency of text readability features in Russian academic texts // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». Том 17. М.: Изд. центр РГГУ. С. 277–287.

НЕСВЯЗНОСТЬ ТЕКСТА, ИЛИ О НЕКОТОРЫХ ПОДВОДНЫХ КАМНЯХ НА ПУТИ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА

О. Ю. Инькова

Olga.Inkova@unige.ch

Женевский университет, ИПИ ФИЦ ИУ РАН

Все исследователи, занимающиеся проблемами изучения текста, согласны с тем, что между составляющими его высказываниями существуют отношения связности, позволяющие превратить последовательность фраз в сообщение, имеющее смысл и коммуникативную цель. Несмотря на то, что проблемы связности текста занимают ученых разных областей знания уже более полувека, вопросы о том, что понимать под этими отношениями, какова их природа и т. д., продолжают вызывать дискуссии. Были сделаны многочисленные предложения, которые, однако, являются скорее инструментом семантического анализа текста, чем теоретической моделью его организации, поскольку основное внимание при анализе уделяется, в первую очередь, микроструктуре текста. Это справедливо, прежде всего, для представителей компьютерной лингвистики, которые несколько упрощенно — в силу принципа единственности (uniqueness), сформулированного У. Манном и С. Томпсон (Mann, Thompson 1988: 248), — полагают, что между следующими друг за другом высказываниями существует только одно «риторическое отношение», хотя еще Дж. Хоббс подчеркивал, что при интерпретации текста читатель совершает параллельно несколько когнитивных операций (Hobbs 1979). Это значит, что связность текста устанавливается одновременно между языковыми сущностями разной природы, взаимодействующими между собой.

Наконец, принцип связанности (connectedness) постулирует, что все фрагменты текста должны быть обязательно связаны между собой некоторым «риторическим отношением» (Mann, Thompson 1988: 248). Создание корпусов текстов, аннотированных с точки зрения отношений связности, позволило, однако, заметить, что между соседними высказываниями может не существовать никакого «риторического отношения», что привело, в частности, разработчиков Penn Discourse Treebank (PDTB) к введению метки NoRel (Webber et al. 2019).

В свою очередь, внимание к макроструктуре текста позволило увидеть, во-первых, что структуры более высокого уровня организации текста, например, абзацы, оказывают влияние на интерпретацию структур более низкого уровня, а во-вторых, что в текстах, особенно значительной длины, присутствуют не только сигналы непрерывности сообщения, но и сигналы дискретности, предупреждающие о том, что между предыдущим и последующим контекстом нет непосредственной связи, подробнее см. (Инькова 2019). Эти сигналы играют не менее важную роль для правильной интерпретации текста, чем сигналы связности, устанавливая границы, в рамках которых должна быть интерпретирована информация, содержащаяся в микроединицах. Этот аспект семантического анализа текста остается пока малоизученным.

В докладе предлагается рассмотреть функции двух таких сигналов, часто взаимодействующих между собой: рамочные выражения («adverbiaux cadratifs» в терминологии М. Шароля) и абзацы, при анализе которых мы будем опираться на работы Ж.-М. Адама. Термин «рамочные выражения», близкий термину «детерминант», предложенный Н. Ю. Шведовой, но не получивший развития в российской лингвистике, охватывает такие языковые единицы, которые задают ту рамку (временную, пространственную, тематическую и т. д.), в границах которой должна быть интерпретирована информация, попадающая в их сферу действия, которая часто выходит за пределы предложения. Рамочные выражения позволяют сгруппировать информацию, индексируемую одним критерием, обеспечивая, таким образом, связность текста. Ср.:

В Аргентине по давней традиции служащие учреждений в последний день уходящего рабочего года выбрасывают из окон старые календари, ненужные ведомости и бланки. В деловой части *страны* — Буэнос-Айресе — уже к полудню тротуары и проезжая часть густо покрываются пухлым слоем бумаги. Никто не знает, *как и когда возник этот обычай*. *Не обходится без казусов*. Однажды не в меру разыгравшиеся сотрудники *одной из газет* выкинули за окно весь архив.

В Болгарии в канун Нового года приобретают кизиловые палочки — непременный атрибут новогоднего праздника. <...> [«Новогодние традиции в разных странах мира»: http://welcomeworld.ru/spravochnik-puteshestvennika/traditsii-i-kharakter-narodov-mira/item/411-newyearstraditions/411-newyearstraditions]

Интерпретация выделенной курсивом информации, находящейся за пределами предложения с обстоятельством места «в Аргентине», точнее локализация описанных в ней ситуаций или их составляющих, очевидным образом зависит от этого рамочного выражения, распространяющего свою сферу действия на пять предложений, группируя их в единый семантический блок. Его сфера действия прерывается рамочным выражением «в Болгарии», оформляющим следующий семантический блок. Границы семантических блоков совпадают здесь и с разделением на абзацы.

Этот пример показывает, насколько важно учитывать данный параметр при интерпретации текста, а также ставит ряд вопросов: как мы интерпретируем ситуации при отсутствии рамок, как взаимодействуют их сферы действия при наличии в тексте нескольких рамок, каковы языковые средства, сигнализирующие о завершении сферы действия рамочного выражения и др. В связи с последним вопросом в докладе будут рассмотрены функции абзацев, разбиение на которые, как правило, исчезает при создании корпусов.

Литература

Инькова О. 2019. Логико-семантические отношения: проблемы классификации // Инькова О., Манзотти Э. Связность текста: мереологические логико-семантические отношения. Москва, Издательский дом ЯСК. С. 11–98.

Adam J.-M. 2018. Le paragraphe: entre phrases et texte. Paris, Arman Colin. 224 p.

Charolles M. 1997. L'encadrement du discours // Cahier de Recherche Linguistique. 1997. 6, pp. 1–73. URL: https://hal.archivesouvertes.fr/hal-00665849.

Hobbs J. R. 1979. Coherence and Coreference // Cognitive science. 1979. 3, pp. 67–90.

Mann W. C., Thompson S. A. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization // Text & Talk. 1988. 8 (3), pp. 243–281.

Webber B., Prasad R., Lee A., Joshi A. 2019. The Penn Discourse Treebank 3.0. Annotation Manual. URL: https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf

ВВОДНО-МОДАЛЬНЫЕ СЛОВА КАК НЕДИСКРЕТНОЕ ЯВЛЕНИЕ В ЯЗЫКЕ

В. В. Казаковская
Институт лингвистических исследований РАН
victory805@mail.ru

Н. К. Онипенко[1]
Институт русского языка им. В. В. Виноградова РАН
onipenko_n@mail.ru

В докладе обсуждается проблема порождения и функционирования вводных слов — единиц, которые могут быть отнесены к недискретным языковым явлениям, — в различных типах дискурса. К анализу привлекаются данные раннего диалога «взрослый — ребенок (1;5–6;5)», устного монолога и письменного нарратива школьников (13;11–17;1), художественного нарратива (НКРЯ).

Вводные слова занимают особое место в языковой системе — между полнознаменательными частями речи и служебными, между словами и предложениями. Их семантика принадлежит трем модусам: ментальному, эмотивному и речевому. Нас будут интересовать вводно-модальные слова (*ВМС*), представляющие ментальный модус, — эгоцентрические элементы (Падучева 2017), выражающие вероятностную оценку и относящиеся к сфере эпистемической модальности. Градуальность семантики (различная эпистемическая сила, или интенсивность) ВМС также обусловливает их недискретность.

В отличие от дескриптивных лексем и словоформ, семантика которых может в большей или меньшей степени зависеть от функции и позиции в синтаксической конструкции, ВМС сохраняют свою модусную семантику в разных синтаксических позициях. Самый термин «вводные слова» указывает на то, что определенные лексемы, для которых возможны позиции в структуре простого (внутрисинтаксическая позиция) и сложного (рамочная позиция) предложения, могут быть употреблены и помимо структуры предложения (парентетическая позиция). Так наз. омонимия кратких прилагательных, наречий и вводных слов на синхронном уровне связана с различиями не столько в семантике, сколько в прагматике. Если в исторической перспективе мы наблюдаем семантические различия между внутрисинтаксической и двумя другими позициями, то сейчас эти различия стерлись. И если в нач. XIX — первой пол. XX в. *видимо, по-видимому* и *очевидно* во внутрисинтаксической позиции выражали перцептивность: *Мы разговаривали* ***по-видимому*** *так дружелюбно, что Иван Игнатьич от радости проболтался* (Пушкин), а в парентетической — эпистемическую модальность (Гатинская 2010): *Собака моя, нимало не медля, с сверхъестественными усилиями залезла под диван и,* ***по-видимому****, нашла там много пыли, потому что расчихалась страшно* (Тургенев), то в конце XX — нач. XXI в. их основным значением становится эпистемическое, что подтверждается возможностью изолированного употребления в качестве ответной реплики неуверенного подтверждения: *Сын все уроки сделал? —* ***По-видимому*** (РР).

Теряя синтаксические связи, словоформа приобретает особый коммуникативный статус — становится эгоцентрическим элементом и маркером отношения говорящего к содержанию высказывания (в нашем случае — эпистемическим). Эгоцентрическая сущность проявляется в ограничении личной парадигмы (например, парентетических глаголов) и закреплении за лексической единицей определенных вариантов субъектной перспективы, то есть определенных соотношений между субъектом действующим, субъектом мыслящим, субъектом говорящим и субъектом воспринимающим (адресатом). В художественном произведении ВМС выполняют особые — текстовые — функции. Они обусловлены нарративной стратегией, которую избирает автор: в нарративе от 1-го л. ВМС субъективируют текст и ста-

[1] Исследование выполнено при поддержке гранта РФФИ №18-012-00263 «Семантико-грамматический словарь русских глаголов».

новятся средствами речевой характерологии, в нарративе от 3-го — обнаруживают дистанцию между ним и героем, маркируют переход от одной точки зрения к другой.

Обращение к детскому дискурсу позволяет увидеть зарождение обсуждаемого явления в речевой деятельности (то есть определить начальную семантическую доминанту, синтаксическую позицию ее экспликаторов и их первичную функцию), а также проследить его развитие в новой для ребенка ситуации создания собственного (устного или письменного) текста.

Результаты проведенного анализа свидетельствуют о том, что ВМС являются предпочтительной формой экспликации говорящего / пишущего, причем как в раннем онтогенезе, так и в позднем, как в устной речи, так и в письменной. В частности, по лонгитюдным данным раннего диалога, первыми эпистемическими маркерами в речи русскоязычных детей (*TD, middle SES*) оказались ВМС с семантикой ‘неуверенность в достоверности сообщаемого’, лидирующие впоследствии по разнообразию (*lemmas*) и частоте словоупотреблений (*tokens*). Высокая степень эпистемической силы маркера и частотность в инпуте влияют на скорость его усвоения (Казаковская 2017). Появление новых ВМС (*наверно*(*е*) → *может* (*быть*), *по-моему*, *кажется*, *видимо*, *возможно*, *пожалуй* vs. *конечно* → *действительно*, *на самом деле*, *правда*), увеличение круга пропозиций, попадающих в фокус вероятностной оценки ребенка, начало использования маркеров во вторичных функциях, а также в качестве реактивной реплики — основные направления раннего онтогенеза эпистемической модальности.

В позднем онтогенезе, по данным сочинений восьмиклассников, зафиксировано меньше ВМС (6), чем в устной речи маленьких детей (11), что может показаться на первый взгляд парадоксальным. Между тем их количество заметно увеличивается в работах выпускников (12), равно как изменяется семантическая доминанта (‘неуверенность’ → ‘уверенность’) и усложняются выполняемые маркерами функции. ВМС начинают «работать» не только на уровне высказывания, но и на уровне текста, обнаруживая при этом взаимодействие нескольких точек зрения — собственной, автора (обсуждаемого произведения) и/или персонажа. Одно из объяснений заключается в изменении ментального статуса говорящего / пишущего, сопряженного со степенью развития модели психического (*Theory of mind*) и позволяющего учитывать точку зрения Другого (в частности, чувствовать позицию автора). Разрешительным условием для экспликации отмеченного взаимодействия становится не только когнитивная зрелость, но и уровень развития дискурсивной компетенции. Ценным элементом последней является умение использовать ВМС в сочетании с определенными союзами в качестве «текстовых коннекторов» (Иорданская, Мельчук 2007). Выражая с их помощью структурные и логические связи между компонентами текста, подросток обнаруживает себя в ипостаси автора текста.

Сочетание системно-языкового и онтогенетического подходов к столь сложному лексико-грамматическому объекту, которым являются вводные слова, представляется перспективным направлением и при изучении других недискретных явлений в языке.

Литература

Гатинская Н. В. 2010. Семантико-синтаксическая история слов очивѣсть, очевидно // Русский язык в научном освещении, 1 (19).

Иорданская Л. Н., Мельчук И. А. 2007. Смысл и сочетаемость в словаре. М.

Казаковская В. В. Языковое и когнитивное в усвоении эпистемической модальности // Acta Linguistica Petropolitana. Труды ИЛИ РАН. Т. 13. Ч. 3. 2017. СПб.

Падучева Е. В. 2017. Эгоцентрические языковые единицы // Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). М. (на правах рукописи).

МУЛЬТИКАНАЛЬНЫЙ РЕСУРС И МОДЕЛИРОВАНИЕ ЕСТЕСТВЕННОЙ КОММУНИКАЦИИ

А. А. Кибрик
Институт языкознания РАН, МГУ имени М.В. Ломоносова
aakibrik@gmail.com

Г. Б. Добров
ЗАО «Консультант Плюс»
wslcdg@gmail.com

Н. А. Коротаев
РГГУ
n_korotaev@hotmail.com

При создании компьютерных систем, имитирующих коммуникативное поведение человека, разработчики нередко опираются на собственные, во многом интуитивные, представления о естественной коммуникации и дополняют их технологическими решениями, в том числе методами машинного обучения. Например, в статье Iribarren 2018, посвященной разработке виртуального собеседника Alexa, отражено упрощенное представление о речи как цепочке фонем. Благодаря методам машинного обучения текущие алгоритмы речепорождения уже сейчас работают довольно хорошо. Очевидно, однако, что порождаемая таким образом речь пока не является вполне естественной. Во многом это связано с тем, что используемые методы учитывают лишь часть значимых речевых характеристик. Следует исследовать вопрос о том, какие характеристики уже учтены в текущих системах, а какие только предстоит учесть. В некоторых случаях позитивные знания об этих характеристиках, имеющиеся у лингвистов, могут пригодиться, и мы предлагаем рассмотреть одну возможность такого типа.

На протяжении ряда лет в Институте языкознания РАН осуществлялся проект «Язык как он есть: русский мультимодальный дискурс», см. multidiscourse.ru. Исследовательская группа изучала мультиканальную коммуникацию, то есть передачу значимой информации при помощи вербального компонента, просодии, движений глаз, жестов рук, жестов головы и т. д. Был разработан единый подход ко всем этим коммуникативным каналам, был создан ресурс *Рассказы и разговоры о грушах* / *Russian Pear Chats and Stories* (RUPEX). RUPEX включает: набор мультимедийных файлов, фиксирующих коммуникативное поведение собеседников; аннотации соответствующих записей; поиск по ресурсу. Аннотация является очень подробной, аналогичных ресурсов в доступе, насколько нам известно, нет. Данные RUPEX позволяют составить адекватное представление о том, как ведут себя люди в ходе естественной коммуникации. Рассмотрим некоторые типы данных, которые благодаря RUPEX доступны специалистам по моделированию.

1. Темп речепорождения (ТРП) и темп говорения (ТГ). ТРП — это количество вербальных единиц (фонем, слов, элементарных дискурсивных единиц — ЭДЕ) в единицу времени коммуникации, а ТГ — это количество вербальных единиц в единицу времени вокализации (то есть без учета пауз). Кроме того, известен эффект замедления в конце ЭДЕ, также заслуживающий внимания.

2. Длительность и размещение пауз. Паузы встречаются как на границе ЭДЕ, так и внутри ЭДЕ, и это очень разные типы пауз. Различаются абсолютные и заполненные паузы (ЗП); последние всегда являются признаком речевого сбоя.

3. Интонационное оформление иллокутивно-фазовых значений. В зависимости от роли в локальной структуре дискурса, ЭДЕ приписываются иллокутивно-фазовые значения, просодически выражаемые в первую очередь характером тонального движения в главном акценте ЭДЕ. Завершению сообщения чаще всего сопутствует нисходящий акцент, а фазовой незавершенности — восходящий, однако возможны дополнительные варианты.

4. Частотность и характеристики жестов рук. Как было показано в работе Kibrik and Fedorova 2019, имеются существенные индивидуальные различия между людьми в характеристиках

используемых ими жестов, включая частотность и длительность жестов, амплитуду и скорость жестов, доли двуручных и одноручных жестов, склонность к право- и леворуким жестам.

5. Направленность и длительность фиксаций взгляда. Как было показано в той же работе, собеседник смотрит на других собеседников или на окружение в течение разной доли времени, в зависимости от ряда факторов, таких как его актуальная коммуникативная роль, тип взаимодействия, индивидуальные особенности. Взгляд на собеседника в наших данных далее конкретизируется как взгляд на лицо, на руки, туловище, другие части тела.

Все эти типы данных существенны и уже доступны. По каждому из параметров можно измерить индивидуальные значения, выяснить разброс, использовать средние или типичные значения при разработке моделей. Иллюстрация того, какие данные извлекаются из RUPEX, приведены на примере темповых характеристик (пункт 1 выше) в таблицах 1, 2. Обсчитаны монологические фрагменты корпуса в исполнении 6 говорящих; в табл. 1 учтены все ЭДЕ этих участников, в табл. 2 — выборка из 154 ЭДЕ, в которых сопоставляются начала и концы, включающие по два слога.

Таблица 1. ТРП и ТГ, единиц/с; разброс по говорящим и средние значения по всей выборке

	ТРП					ТГ (слова + ЗП)				
	Фонемы	Слоги	Слова	Слова/ЗП	ЭДЕ	Фонемы	Слоги	Слова	Слова/ЗП	ЭДЕ
min–max	8,61–12,03	3,65–5,22	1,92–2,56	2,14–2,68	0,57–0,69	11,54–15,98	4,90–6,91	2,56–3,36	2,79–3,5	0,73–0,89
ср. знач.	10,20	4,40	2,18	2,38	0,62	13,06	5,63	2,79	3,05	0,80

Таблица 2. Эффект замедления в конце ЭДЕ.
Различие между двумя первыми и двумя последними слогами значимо (Тест Уилкоксона; p=0,01)

	ТН — темп начала ЭДЕ, мс/слог	ТК — темп конца ЭДЕ, мс/слог	эффект замедления: ТК/ТН
min–max (ЭДЕ)	30–627	70–421	0,33–10,18
станд. откл.	84	67	1,28
ср. знач.	152	222	1,87

Мы полагаем, что перечисленные характеристики поведения коммуникантов должны учитываться при создании экологически валидных виртуальных собеседников. В частности, разработчики могли бы опереться на данные по разбросу и средним значениям ТРП, ТГ и эффекта замедления. В качестве пилотного сопоставления наших данных с текущим положением в индустрии мы проанализировали фрагмент одного из рассказов RUPEX, произнесенный голосовым помощником «Алиса» (cloud.yandex.ru/services/speechkit). Согласно нашим результатам, эффект замедления обнаружился, но его среднее значение составило 1,38, что существенно отличается от наших данных (см. табл. 2). Учет данных RUPEX может сделать речь и в целом коммуникативное поведение виртуального собеседника более естественными.

Литература

Iribarren M. 2018. voicebot.ai/2018/11/29/neural-text-to-speech-development-allows-alexa-to-sound-more-realistic.

Kibrik A. A., Fedorova O. V. 2019. A «portrait» approach to multichannel discourse. Proceedings of LREC 2018, Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan. 2019, pp. 1908–1912.

РАЗДЕЛЬНО-СЛИТНЫЕ БИНОМЫ *ЛИХЭЦЫ* КАК ПРОЯВЛЕНИЕ НЕДИСКРЕТНОСТИ В СОВРЕМЕННОМ КИТАЙСКОМ ЯЗЫКЕ

П. О. Кисель

14polink@mail.ru

Санкт-Петербургский государственный университет

Раздельно-слитные биномы представляют весьма дискуссионное явление в современном китайском языке. Специфика, обусловившая актуальность исследования данных единиц, заключается в том, что их компоненты допускают дистанцированное друг от друга положение в предложении за счет вставки различных элементов как в постпозиции к первому компоненту, так и в препозиции ко второму. В отечественном китаеведении такие единицы, в большинстве своем построенные по глагольно-объектной модели, назывались по-разному: глагольно-объектные биномы в (Драгунов 1946), глагольно-именные комплексы в (Маматюк 1986), инкорпорации в (Иванов, Поливанов 2003), «слова-словосочетания» в (Готлиб 1995), «разрывной комплекс» в (Панфилов 2008), «композит-комплексы» в (Рукодельникова 1995) и др. Мы пользуемся транскрипцией предложенного в (Лу 1957) китайского термина 离合词 *líhécí лихэцы* (букв. «раздельно-слитные слова»).

В постпозиции к первому компоненту глагольно-именного образования допускается постановка глагольных счетных слов и синтаксически несамостоятельных единиц из закрытого списка: аспектуальные показатели, маркеры возможности/невозможности, глагольные модификаторы, уточняющие характер протекания действия. Наиболее репрезентативным, однако, представляется оформление именного компонента глагольно-объектного образования, представленное открытым списком вариантов, куда входят счетные комплексы, числительные, местоимения, существительные, прилагательные, а также именные группы в роли косвенного дополнения и обстоятельства длительности в препозиции к прямому дополнению, синтаксически реализуемые как определения.

Разработанные на материале индоевропейских языков критерии выделения слова, в частности, критерий «вставимости» как признак отсутствия у единицы словесного статуса, применительно к материалу современного китайского языка, а именно — к глагольно-именным биномам *лихэцы* остается спорным ввиду своей строгой формальности. Выдвигается гипотеза, что глагольно-именные *лихэцы* различаются характером семантических реляций между компонентами, следствием чего может быть различная степень спаянности компонентов как с семантической, так и с формальной точек зрения.

В рамках семантического подхода мы выделяем такие группы раздельно-слитных слов, как:

1. Лихэцы с максимальной идиоматичностью значения (значение полисиллаба не сводится к сумме значений его переосмысленных компонентов): 吹牛 chuīniú ‘хвастаться’ (букв.: ‘раздувать’ + ‘корова, бык’).
2. Лихэцы, в которых именной компонент является «пустым дополнением» (значение полисиллаба базируется на семантике глагольного компонента):
   2.1. Именной компонент семантически дублирует значение глагольного компонента: 洗澡 xǐzǎo ‘мыться, принимать ванну’ (букв.: ‘мыть, купаться’ + ‘купанье, ванна’).
   2.2 Именной компонент представляет собой генерический, или родовой объект, наиболее характерный для действия, обозначенного компонентом глагольным: 懂事 dǒngshì ‘понимать, разбираться’ (букв.: ‘понимать’ + ‘дело’).
3. Лихэцы, в которых глагольный и именной компоненты употребляются в своих основных значениях и являются семантически значимыми (значение полисиллаба базируется на семантике обоих компонентов): 做梦 zuòmèng ‘видеть сон’ (букв.: ‘делать’ + ‘сон’).

Анализ корпусного материала выявил, что именные компоненты *лихэцы* всех групп допускают в препозиции примерно схожий набор вставок. При этом в группе 1 и подгруппе 2.1 семантическая спаянность полисиллаба, как правило, «побеждает» формальное расчленение, и вставные элементы с семантической точки зрения относятся к глагольному комплексу в целом, что позволяет рассматривать данные единицы как тяготеющие к словам. В случае оформления именного компонента *лихэцы* в подгруппе 2.2 и группе 3 формальное расчленение с различной частотой разрывает семантическую спаянность полисиллаба, что свидетельствует о тяготении данных единиц к словосочетаниям.

Представляется целесообразным применение скалярного подхода к определению статуса раздельно-слитных биномов с выстраиванием шкалы, полюсами которой будут выступать прототипическое «слово» и «словосочетание» соответственно, к которым в разной степени будут тяготеть расчленяемые единицы.

Литература

Готлиб О. М. 1995. О природе и видах синкретизма в современном китайском языке // Актуальные проблемы китайского языкознания: Материалы VII Всероссийской конференции по китайскому языкознанию. М. С. 31–34.

Драгунов А. А. 1946. О зависимых членах предложения в современном китайском языке // Известия Академии наук СССР. Т. 5 Вып. 6. С. 484–489.

Иванов А. И., Поливанов Е. Д. 2003. Грамматика современного китайского языка. М.: Едиториал УРСС.

Маматюк Ю. Д. 1986. Глагольно-именные комплексы в КЯ (к проблеме уровневого разграничения: Автореф. дис. ... канд. филол. наук. М.

Панфилов В. С. 2008. Слово в китайском языке (прототипический подход) // Вопросы языкознания. №2. С. 57–64.

Рукодельникова М. Б. 1995. К проблеме изучения лексических комплексов в современном китайском языке (подтип глагольных комплексов) // Актуальные проблемы китайского языкознания: Материалы VII Всероссийской конференции по китайскому языкознанию. М. С. 107–109.

Лу Чживэй. 1957. Словообразование китайского языка. Пекин: Наука (陆志伟. 汉语的构词法. — 北京: 科学出版社).

СПЕЦИФИКА ОБЪЕКТИВАЦИИ КАТЕГОРИИ НАБЛЮДАТЕЛЯ В НЕЙТРАЛЬНЫХ И ЭМОЦИОНАЛЬНЫХ ТЕКСТАХ ЧЕРЕЗ ПРИЗМУ СЕНТИМЕНТ-АНАЛИЗА[1]

А. В. Колмогорова
nastiakol@mail.ru

А. А. Калинин
verbalab@yandex.ru

А. В. Маликова
malikovaav1304@gmail.com

Сибирский федеральный университет

Сентимент-анализ является одним из направлений развития искусственного интеллекта. Его цель — обеспечить автоматическое, без участия человека, распознавание эмоций в тексте, например, для анализа уровня социальной напряженности в обществе (Ceron et al. 2012), удовлетворенности потребителя качеством услуги или товара (Spatiotis et al. 2020; Abassy 2020).

Проект, в рамках которого выполняется настоящее исследование, направлен на разработку технологии, способной на основе машинного обучения автоматически производить атрибуцию интернет-текстов на русском языке к 9 классам, 8 из которых соответствуют базовым эмоциям, согласно модели эмоций «Куб Лёвхейма», а 9 класс − нейтральный. Такой многоклассовый сентимент-анализ позволит получать данные о динамике изменения эмоционального баланса в отдельных социальных группах интернет-пользователей, что представляется актуальной задачей в контексте «новой чувствительности» дискурса социальных сетей (Кошкарова, Яковлева 2019).

Для реализации проекта был сформирован датасет из 3920 эмоциональных текстов, отобранных первоначально в социальной сети ВКонтакте, в пабликах «Подслушано», «Палата №6», «Карамель», по хештегам (например, #Стыдно — эмоция Стыд/Унижение), а затем оцененных 2000 асессорами с использованием разработанного группой интерфейса для оценки эмоции в тексте.

Подкорпус нейтральных текстов был получен в результате применения следующей процедуры. Нами был собран датасет из 15000 фрагментов из Википедии, которая принималась как эталон нейтральности, и 15000 текстов из паблика «Подслушано», которые рассматривались, как не-нейтральные образцы. После этого был обучен бинарный классификатор на основе LSTM-сети. Он классифицировал тексты на нейтральные и не-нейтральные с точностью 92 % на валидационной выборке.

После этого данный классификатор был применен исключительно для текстов из социальной сети, чтобы автоматически выявить нейтральные или близкие к нейтральным тексты. Таким образом был сформирован нейтральный подкорпус текстов (далее НТ) из ВКонтакте в объеме 49000 токенов.

В данной публикации в качестве референтного для сравнения нейтральных текстов используется эмоциональный подкорпус (далее — ЭП) «Интерес/Воодушевление» (184 074 токена), маркированный, согласно нейрофизиологической модели эмоций Лёвхейма, высоким уровнем всех трех «гормонов эмоций»: дофамина, серотонина и норадреналина.

Цель данной публикации — охарактеризовать одну группу дискриминантных черт нейтральных текстов, а именно − специфику проявления фигуры Наблюдателя. Основной из использованных методов — метод сравнительно-сопоставительного анализа нейтральных текстов на фоне эмоциональных — реализован с опорой на инструментарий корпусного менеджера Sketch Engine, в который были загружены подкорпусы.

В ходе исследования сделан вывод о том, что ключевое отличие эмоциональных текстов от нейтральных базируется в том числе на уровне эпистемологии: в эмоциональных текстах

[1] Исследование выполнено при поддержке гранта РФФИ (проект «Разработка классификатора русскоязычных интернет-текстов по критерию их тональности на основе модели эмоций «Куб Левхейма»« № 19-012-00205)

позиции говорящего (нарратора), наблюдателя совпадают, а в нейтральных говорящий, как правило, не является наблюдателем. Эта особенность наиболее ярко проявляется в текстовой реализации дейктических категорий, поскольку основное свойство дейксиса — «либо совпадение, либо несовпадение пространственно-временных координат описываемого факта, как их мыслит говорящий, с теми координатами, в которых он мыслит себя» (Апресян 1995: 638).

Пространство говорящего, маркированное дейктиками *тут*, *там*, *здесь*, *туда*, *сюда*, в нейтральных текстах проявляется значительно реже (суммарная относительная частотность — 968 ipm), чем в эмоциональных (2741 ipm).

Что касается **категории времени**, то опосредованно специфику восприятия времени говорящим, по-видимому, отражает комбинаторика лексемы *год*: 1) в списке самых частотных частей речи в контекстах с *год* в нейтральных текстах числительные имеют первый ранг, а в эмоциональных — 6; 2) из 20-ти наиболее статистически релевантных коллокатов лексемы *год* в каждом из подкорусов, согласно значениям мер ассоциации T-score, MI, LogDice, 7 — это числительные, но в НП они все представлены в числовом формате и четыре из них − трех-, четырехзначные; а в ЭП доминирует словесный формат числительных, среди которых нет ни одного трех- или четырехзначного, например: (1) *А кроме того школьницы 2010 — 2015 годов очень мало знали о грамотном демакияже и уходе за кожей* (НП); (2) *Пять лет назад, как только женился, ушел в дальнобойщики, были очень нужны деньги* (ЭП). Очевидно, что «ословленный» относительно небольшой отрезок времени имеет более «человеческую мерку» (человек может быть непосредственным наблюдателем происходящего), чем временной период, обозначенный трех- или четырехзначными цифрами. Последний исключает присутствие наблюдателя.

Синтаксическая дистрибуция лексемы *год* в эмоциональных текстах демонстрирует, например, что самой частотной конструкцией является $N_{одуш}$ *лет* Num (gen.) − например, *девушка лет двадцати*. А в нейтральных эта конструкция представлена единичными примерами. Именно присутствие говорящего в качестве наблюдателя позволяет проникать маркерам **субъективной личной оценки**, возраста например, в эмоциональные тексты.

Таким образом, глубинное эпистемологическое различие в позиции наблюдателя, существующее в эмоциональных текстах, с одной стороны, и в нейтральных − с другой, «прорастает» в речи в специфических формах «непосредственного» (слова-дейктики) и «опосредованного» (лексическая и синтаксическая комбинаторика) дейксиса, которые могут быть выявлены при помощи корпусного анализа, формализованы и использованы в качестве дискриминантных черт нейтральных текстов в машинном обучении по прецедентам.

Литература

Апресян Ю. Д. 1995. Дейксис в лексике и грамматике и наивная модель мира. Избранные труды. Т. 2. М.: Школа «Языки русской культуры». С. 629–650.

Кошкарова Н. Н., Яковлева Е. М. 2019. Дискурс новой эмоциональности: коммуникативные практики цифровой реальности // Политическая лингвистика. 5(77). С. 147–152.

Abbassy M. 2020. Opinion Mining for Arabic Customer Feedback Using Machine Learning // Journal of Advanced Research in Dynamical and Control Systems. 12, pp. 209–217.

Ceron A., Curini L., Iacus S. M. 2012. Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France // New Media & Society. 16(2), pp. 340–358.

Spatiotis N., Perikos I., Mporas I., Paraskevas M. 2020. Sentiment Analysis of Teachers Using Social Information in Educational Platform Environments // International Journal on Artificial Intelligence Tools. 29(02). P. 2040004:1–2040004:28.

СРЕДСТВА ОБЕСПЕЧЕНИЯ СВЯЗНОСТИ В СОВРЕМЕННОМ КИТАЙСКОМ ЯЗЫКЕ

Е. Н. Колпачкова
ekolpachkova@gmail.com

Е. И. Мовшович
evinig@gmail.com

Санкт-Петербургский государственный университет

В докладе представлены результаты анализа стратегий поддержания связности текста в китайском языке с особым акцентом на взаимодействии темы/ремы, реализующемся не только в рамках одного предложения, но и в более крупных синтаксических единицах, на средствах повторной номинации как обязательных элементах семантической организации текста. Кроме того, анализируется такой параметр китайского текста, как насыщенность дискурсивными словами и предпочтение выбора между дискурсивными словами той или иной семантики.

Предложение в китайском языке обычно содержит несколько высказываний-сообщений, что затрудняет членение текста на определенные отрезки. Применительно к данной языковой системе о предложении в традиционном понимании можно говорить лишь с некоторыми оговорками, поскольку текстовый отрезок (от точки до точки) будет представлен как цепочка предикаций, разделенных запятыми, но объединенных единой темой, то есть монотематичных и полипредикативных по своей природе. Подобные синтаксические образования организуются через тема-рематическую прогрессию, т. е. фактически через последующую предикацию, «пошагово» реализующую тематизацию ремы, продвигая каждый раз от клаузы к клаузе повествование от данного к новому.

С точки зрения тема-рематической структуры китайский язык следует логически обусловленному типу изложения, когда общим правилом является линейная организация компонентов в сообщении от темы к реме, соответствующая объективному, прямому порядку слов, типичному для данной синтаксической структуры. Элемент в абсолютном начале высказывания в подавляющем большинстве случаев воспринимается как тема, т. е. в системе преобладает *позиционный* способ маркирования темы, когда компонент в инициальной позиции выступает темой, но в целом могут использоваться и другие способы. Так, при *грамматическом* способе маркирования темы применяется клефтинг, реализуемый с помощью рамочной конструкции（是）…的 *(shì)…de*. К *лексическому* способу относятся вводящие в высказывание тему предложные обороты 至于 *zhìyú*, 关于 *guānyú* 'что касается', сфера употребления которых ограничена текстами делового стиля. Наиболее продуктивным в исследуемом языке является *фонетический* способ, в частности его паузальный подвид, когда элемент, являющийся темой в данной структуре, отделяется от остальной части высказывания в устной речи паузой, а в письменном тексте — запятой.

Как одно из важнейших средств обеспечения связности текста в данном исследовании анализировалась повторная номинация, случаи которой в китайском тексте были представлены *именной*, *прономинальной* и *нулевой анафорой*. В рамках исследования стояла задача определить референциальное линейное расстояние, т. е. точное значение контактной или дистантной позиции анафора по отношению к своему антецеденту. Для анализа и количественной оценки данного явления текст разбивался на пары пропозитивных синтагм, потенциально содержащих кореферентные элементы, и высчитывалось «расстояние» между анафорическим местоимением и его антецедентом, при котором местоимение все еще могло быть проинтерпретировано правильно.

Как представляется, система переходов от тема-рематической структуры к пропозиции должна учитывать использование семантических операторов, в число которых предлагается включить дискурсивные слова. Параметр насыщенности текста дискурсивными словами как особым функционально-семантическим и прагматическим классом единиц может послу-

жить дополнительным источником для типологизации текстов и языков. Собранные количественные данные по частотности различных типов дискурсивных слов в китайском языке, описанных в (Касевич и др. 2014; Колпачкова 2014), дают возможность оценить их релевантность для обеспечения целостности и связности текста. Китайский материал выявил высокую частотность показателей сильной эмфазы (就 *jiù* ‘именно, как раз’, 就是 *jiùshì* ‘именно, это и есть’). Особый интерес представляет семантическая область, связанная с употреблением дискурсивных слов, маркирующих контрастивную тему, поскольку тема-рематическая семантика оказывается частным случаем дискурсивной. Сопоставление полученных данных с результатами анализа других восточных языков свидетельствует о высокой частотности дискурсивных слов в китайском тексте.

Литература

Касевич В. Б., Берникова О. А. и др. 2014. Дискурсивные слова восточного текста // Языки стран Дальнего Востока, Юго-Восточной Азии и Западной Африки: Материалы XI Международной конференции. СПб.: СПбГУ. С. 86–90.

Колпачкова Е. Н. 2014. Модальные и дискурсивные слова в современном китайском языке // Языки стран Дальнего Востока, Юго-Восточной Азии и Западной Африки: Материалы XI Международной научной конференции. СПб.: СПГУ. С. 91–94.

ИЕРАРХИЯ ЧЕЛОВЕЧЕСКИХ ПОНЯТИЙ, ЗАДАННАЯ ЗНАЧЕНИЯМИ ПРЕДМЕТНЫХ СУЩЕСТВИТЕЛЬНЫХ

А. Д. Кошелев

koshelev47@gmail.com

Издательский Дом ЯСК

1. Сенсорный концепт и его партитивная система. К человеческим понятиям принято относить основные (непроизводные) значения конкретной лексики. Условимся называть **сенсорным концептом** (понятием) значение предметного существительного (*стул, чашка, озеро, дерево* и под.). Сенсорный концепт (значение) имеет структуру: «Прототип ← Функция», где Прототип это перцептивный (часто визуальный) образ типичного референта слова, Функция — это человеческая интерпретация референта, а стрелка обозначает отношение интерпретации, приписывающее Функцию Прототипу, см., напр., (Кошелев 2019: 35–38). Примеры сенсорных концептов:

СТУЛ = Прототип (образ стула) ← Функция ‘Позволяет сидеть, опираясь спиной’;

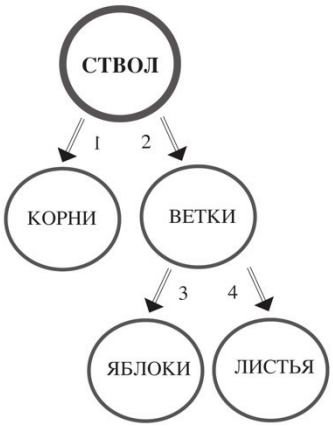
ОЗЕРО = Прототип ← Функция ‘Сохраняет от растекания большую массу воды’;

ДЕРЕВО = Прототип ← Функция ‘Растет само собой и порождает семена’.

Для сравнения: у значений существительных *дурак, сорняк*, *хищник* и под. нет сенсорного прототипа, есть только функциональная составляющая. Поэтому их значения суть **функциональные концепты**. Как показано (Там же: 43–46), сенсорный концепт разделяется на естественные части, среди которых выделяется одна главная часть (ее функция составляет бóльшую часть общей функции концепта) и дополнительные части. Все части концепта интегрируются в партитивную систему посредством бинарных ролевых отношений (нумерованные двойные стрелки), связывающих главную часть с дополнительными частями. Например, концепт ДЕРЕВО разлагается на три части: **Ствол** (главная часть, выделена п/ж), Корни и Ветки. Они интегрируются в систему **партитивную систему**: Корни ⇐1 **Ствол** 2⇒ Ветки.

Аналогичная партитивная система имеет место и для частей. Например, для Ветки (с листьями и плодами) имеем систему частей: Плоды/Семена ⇐3 **Ветка** 4⇒ Листья, см. (1).

(1)



2. Иерархия сенсорных концептов. Партитивная система является главной характеристикой сенсорного концепта. Ее модификации порождают иерархию концептов. В качестве примера рассмотрим процесс образования иерархической классификации растений, см. рис. 1.

Возьмем конкретный концепт ЯБЛОНЯ (1).

Первый вариант модификации: замена элемента на более общий. Заменим видовые характеристики на родовые: «яблоко» — на «плод», а «яблоневый лист» — на «плоский лист». Тогда получится более общий концепт ПЛОДОВОЕ ДЕРЕВО.

7.	**РАСТЕНИЕ**	
6.	**ОДНОСТВОЛЬНОЕ РАСТЕНИЕ**	ДРУГИЕ РАСТЕНИЯ
5.	**ДЕРЕВО**	ПАЛЬМА, и др.
4.	ХВОЙНОЕ ДЕРЕВО	**ЛИСТВЕННОЕ ДЕРЕВО**
3.	**ПЛОДОВОЕ ДЕРЕВО**	НЕПЛОДОВОЕ ДЕРЕВО
2.	**ЯБЛОНЯ**	ГРУША, и др.
1.	**ЯБЛОНЯ-АДМИРАЛ**	ЯБЛОНЯ-САНТАНА, и др.

Рис. 1. Иерархическая классификация растений. Уровни 1-5 задают иерархию (по степени общности) сенсорных концептов, а уровни 6–7 – иерархию несенсорных концептов (6 – гетерогенный, а 7 – функциональный концепты). Темным цветом выделена полная ветвь иерархии от самой общего элемента (**РАСТЕНИЕ**) до самого конкретного (**ЯБЛОНЯ-АДМИРАЛ**).

Второй вариант модификации: замена элемента на более конкретный. Заменим видовую характеристику «яблоко», на подвидовую — на сорт яблока адмирал. Тогда получится более конкретный концепт ЯБЛОНЯ-АДМИРАЛ.

Третий вариант модификации: замена элемента на альтернативный элемент того же уровня общности. Заменим характеристику «яблоко», на характеристику «груша». Тогда получится другой концепт ГРУША того же уровня.

Заменим элемент «плоский лист» на альтернативный «игольчатый лист», тогда получится альтернативный концепт ХВОЙНОЕ-ДЕРЕВО.

Четвертый вариант модификации: придание элементу факультативного статуса (может быть, а может и не быть). Объявим элемент «плод» факультативным. В этом случае получится более общий концепт — ЛИСТВЕННОЕ ДЕРЕВО, включающий плодовые и неплодовые лиственные деревья. Если теперь в этом концепте элемент «плоский лист» заменить более общим элементом «лист», включающим также «игольчатый лист», то получится наиболее общий сенсорный концепт ДЕРЕВО, включающий лиственные и хвойные деревья.

Пятый вариант модификации: удаление элемента. Удалим элемент «плод» и оставим элемент «плоский лист», тогда получится концепт НЕ-ПЛОДОВЫЕ-ДЕРЕВЬЯ (среди хвойных деревьев нет плодовых; кедровые орехи — это семена, а не плоды).

3. Гетерогенные концепты. Шестой вариант модификации партитивной системы. Во многих языках имеются гетерогенные категории (термин А. Вежбицкой (2011: 111)): ОБУВЬ, МЕБЕЛЬ, ПОСУДА и под., строгая дефиниция которых до сих пор не известна (по крайней мере, автору). В рамках предложенной иерархии эти концепты получают естественное определение: **гетерогенный концепт** задается **главной частью** партитивной системы. Все остальные ее части являются факультативными, т. е. могут отсутствовать.

3.1. Рассмотрим гетерогенный концепт ОБУВЬ. Мы полагаем, что главной частью любой обуви является подошва, которая служит прокладкой между подошвой ноги и землей. Она защищает ногу от повреждений при ходьбе. Подошва имеется у пляжных тапочек и кроссовок, у туфель на шпильке и солдатских сапог. В то же время, бахилы не относятся к обуви, поскольку не защищают подошвы ног. По этой же причине и галоши не являются обувью, ср., в частности, выражения: *галоши для обуви / на обувь*. Не относятся к обуви и носки, поскольку они не предназначены для контакта с землей.

3.2. Главной частью гетерогенного концепта ПОСУДА является емкость, в которую помещается еда или питье: тарелка, чашка, рюмка, соусник, кастрюля, сковородка, банка и пр. В то же время, ваза, поднос, пепельница имеют емкость, но не относятся к посуде, поскольку она несет другие функции — в нее не кладут еду или питье.

3.3. Главной частью гетерогенного концепта МЕБЕЛЬ является горизонтальная плоскость, расположенная над поверхностью пола и используемая для размещения на ней человека или его вещей. Такой плоскостью обладают: стол, кресло, диван, книжная полка, шкаф (несколько плоскостей) и пр. Соответственно, не относятся к мебели: пол, спальный мешок, дверь, подоконник (не используется для размещения вещей), ванна и пр. Платяной шкаф может не иметь горизонтальной полки, но ее функцию выполняет горизонтальный шест, на котором висят платья. Он является главной частью шкафа и его можно трактовать как вырожденную горизонтальную плоскость.

Заметим, что наличие гетерогенных концептов косвенно подтверждает иерархичность партитивной системы с главной частью в вершине иерархии. Ср. также применение данного типа категоризации в зоологии. Термины «позвоночные», «кольчатые» обозначают типы животных с некоторой общей (и, по-видимому, также главной) частью строения тела.

4. Функциональные концепты. Наконец, **седьмой вариант модификации** партитивной системы это когда она вся **целиком объявляется факультативной**. В результате возникает функциональный концепт. Так, концепт ДЕРЕВО определяется (примерно) так: 1) постоянно растущий живой организм, вросший одной своей частью в землю, и не способный к самостоятельным перемещениям или действиям; 2) имеющий партитивную систему: Корни (растет в земле), **Ствол** (растет из корней над землей) и Ветки (растут из ствола). Если объявить эту партитивную систему факультативной, то получится функциональный концепт РАСТЕНИЕ, в котором обязательно только функциональное свойство 1), ср. с аналогичными концептами ЖИВОТНОЕ, СОРНЯК, ХИЩНИК.

У многих сенсорных концептов (СТУЛ, ОЗЕРО) кроме функционального компонента, присущего их партитивной системе, нет других, дополнительных функциональных свойств. Поэтому с удалением партитивной системы исчезает и сам концепт.

Введенная иерархия, естественно, относится и к лексическим значениям.

Литература

Вежбицкая А. 2011. Семантические универсалии и базисные концепты. М.: Языки славянских культур.

Кошелев А. Д. 2019. О генезисе мышления и языка: Генезис понятий и пропозиций. Аристотель и Хомский о языке. Влияние культуры на язык. М.: Издательский дом ЯСК.

ТИПЫ ГРАММАТИЧЕСКОЙ ИНФОРМАЦИИ В «СЕМАНТИКО-ГРАММАТИЧЕСКОМ СЛОВАРЕ РУССКИХ ГЛАГОЛОВ» И ИХ ПРИМЕНЕНИЕ В КОМПЬЮТЕРНОМ АНАЛИЗЕ ТЕКСТОВ (НА МАТЕРИАЛЕ ГЛАГОЛОВ С КАУЗАТИВНО-ЭМОТИВНОЙ СЕМАНТИКОЙ)[1]

С. А. Крылов
krylov-58@mail.ru
Институт востоковедения РАН

Е. Н. Никитина
yelenon@mail.ru
ФИЦ «Информатика и управление» РАН

Н. К. Онипенко
onipenko_n@mail.ru
Институт русского языка РАН

М. А. Станкевич
stankevich@isa.ru
ФИЦ «Информатика и управление» РАН

1. В докладе будут освещены результаты междисциплинарного эксперимента, включающего постановку задачи психологом, автоматический анализ текстов авторов с разным психиатрическим статусом, математико-статистическую обработку данных автоматического анализа и последующую когнитивно-лингвистическую интерпретацию полученных данных. Целью эксперимента было выявление значимых грамматических и лексических характеристик, выделяющих группу авторов того или иного психиатрического статуса — в норме и патологии.

2. Материалом для автоматического анализа текстов послужил корпус эссе на тему «Я, другие, мир», написанных в рамках психологического эксперимента авторами с разным психиатрическим статусом, образующими три группы: «здоровые», «депрессия», «шизофрения». Тексты исследовались в связи с употребимостью переходных глаголов каузативно-эмотивной семантики и их возвратных коррелятов с опорой на типы грамматической информации в «Семантико-грамматическом словаре русских глаголов», работа над которым ведется в ИРЯ РАН под руководством Н.К. Онипенко.

3. Глаголы каузации психического состояния (эмоции), или каузативные эмотивы, означают такую ситуацию (каузативную), которая соединяет два события, два положения дел: первое — бытие, действие или состояние чего-либо или кого-либо и второе — эмоциональное состояние субъекта личного. Между двумя положениями дел устанавливаются причинно-следственные отношения. Н.Д. Арутюнова относит глаголы рассматриваемой группы (в ее терминологии — глаголы «эмоционального воздействия») к событийным каузативам, семантика которых содержит «указание на каузированное событие» (Арутюнова (1976) 2002: 176), см. также (Апресян 2013).

4. В ходе исследования были введены новые приемы и параметры анализа глагольных лексем и текста: удалены «омонимичные» глаголы, выделены семантические подклассы в составе эмотивно-каузативных глаголов (позитивы, негативы, амбиваленты, деэмотивы), которые частично могут быть обоснованы с обращением к (Иорданская 1970, Апресян 1974); сформулированы параметры эмоциональности — общая эмоциональность текстов группы и плотность эмотивов в тексте; соотнесенность эмотивной и каузальной составляющей в конструкции и тексте; эмотивный динамизм (отнесенность/неотнесенность эмоции к плану прошлого); произведен анализ экспериенциальной и каузальной составляющих в связи с Я: экспериенцер — собственно-Я, Я, соотнесенное с другими (обобщенно-личное значение), не-Я; Я каузирующее.

[1] Исследование выполнено при поддержке грантов РФФИ № 17-29-02305 и № 18-012-00263

5. Исследование показало, что учет каузативно-эмотивных глаголов и их возвратных коррелятов при анализе письменной речи может внести существенный вклад в понимание и разграничение психиатрических статусов пишущих субъектов. Была выявлена обращенность разных групп испытуемых к определенным типам синтаксических конструкций, определенным значениям категорий времени и лица, а также семантическим подклассам эмотивных глаголов.

6. Интерпретация языковых форм и их значений позволила сформулировать предположения о поведенческих закономерностях, характерных для лиц с тем или иным психиатрическим статусом.

7. Исследование стимулировало к формулированию идей для дальнейшего компьютерного анализа текста: (1) учет семантической структуры глагола — именных составляющих конструкции в единстве их формы (падеж), категориального значения (семантический, или категориальный, класс имени) и семантической роли (типовое обобщенное значение именной группы, зависимое от семантики глагола) (ср. с понятием «связанная синтаксема» в (Золотова 1988)), что позволит выйти за пределы морфологии, подняться до уровня синтаксической конструкции и ее компонентов; (2) кластерный подход, т. е. учет комбинаторной силы нескольких признаков (например, низкая эмоциональность в сочетании с высоким уровнем логизма (шизофрения); низкая степень позитивизма в сочетании с большим количеством форм прош. вр. (депрессия) и т. п.).

8. Работа с языковым материалом каузативно-эмотивных глаголов в рамках психологического эксперимента оказалась полезной и для собственно лингвистической стороны дела. Для лексики и синтаксиса это касается нового разграничения семантических подклассов эмотивов, наблюдений за изменением эмотивной семантики в контексте отрицания. Для грамматической семантики и лексикографической практики получены новые аргументы в пользу так называемых «третьеличных» глаголов.

Литература

Апресян В. Ю. 2013. Семантика эмоциональных каузативов: статус каузативного компонента // Компьютерная лингвистика и интеллектуальные технологии: по материалам Международной конференции «Диалог 2013». Вып. 12(19). М. С. 43–57.

Апресян Ю. Д. 1974. Лексическая семантика: Синонимические средства языка. М.: Наука.

Апресян Ю. Д. 2006. Фундаментальная классификация предикатов // Языковая картина мира и системная лексикография. М. С. 75–110.

Арутюнова Н. Д. 2002. Предложение и его смысл. М.

Золотова Г. А. 1988. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. М.

Иорданская Л. Н. 1970. Словарные статьи ВОСХИЩАТЬСЯ и ВОСТОРГАТЬСЯ // И. А. Мельчук, Л. Н. Иорданская. Попытка лексикографического толкования группы русских слов со значением чувства // Машинный перевод и прикладная лингвистика. Вып. 13. М. С. 3–26.

ОСОБЕННОСТИ ЖАНРА ПОПУЛЯРНЫХ МЕДИЦИНСКИХ РЕКОМЕНДАЦИЙ В КОНТЕКСТЕ ЗАДАЧ КОГНИТИВНОГО АССИСТИРОВАНИЯ[1]

Ю. М. Кузнецова

kuzjum@yandex.ru

Федеральный исследовательский центр «Информатика и управление» РАН

Когнитивное ассистирование представляет собой перспективное прикладное направление искусственного интеллекта, связанное с предоставлением пользователям разносторонней поддержки. В отличие от технологий, называемых интеллектуальными агентами, призванных содействовать поиску и анализу информации, функционал когнитивных ассистентов (КА) включает в себя также помощь в планировании и осуществлении деятельности, в поддержании и развитии ее мотивации (Смирнов и др. 2019). Особенно полезными данные функции оказываются по отношению к деятельности с отсроченным результатом, такой, как длительное обучение или формирование здорового образа жизни. Здесь требующиеся от человека усилия не приводят к мгновенным очевидным достижениям, поэтому оказание стимулирующей поддержки со стороны заинтересованного партнера по общению может иметь критическое значение, что демонстрирует медицинская практика по преодолению поведенческого саботажа при сопровождении хронических заболеваний (Кузнецова 2019). Та же практика доказывает значимость наряду с содержательными и формальных особенностей коммуникации врача и пациента.

Предметом коммуникации КА с его «хозяином» является план действий, который ассистент разрабатывает, применяя свои возможности по интеллектуальному анализу информации, в ответ на инициирующий запрос пользователя. Этот план, помимо адекватности общему намерению человека, должен быть достаточно подробным и содержать определенные промежуточные цели, последовательное достижение которых обеспечит, во-первых, переживание успешности и повышение мотивации к продолжению усилий, а во-вторых, доверие к КА как к источнику полезных и исполнимых советов. В терминах деятельностного подхода А. Н. Леонтьева можно сказать, что в коммуникации КА должны быть представлены все три уровня психологической структуры деятельности, а именно: 1) мотив (М), определяющий, *зачем* должна быть осуществлена рекомендуемая активность, 2) цели (Ц), указывающие на то, *что* именно должно быть сделано для реализации мотива, и 3) задачи (З), конкретизирующие, *как* выдвинутые цели могут быть достигнуты с учетом текущих условий, обстоятельств и возможностей. Достижение цели, как следует из теории А.Н. Леонтьева, сопровождается положительными эмоциями, придает полученному промежуточному результату положительный личностный смысл и поэтому создает энергетические предпосылки к продолжению усилий. Конкретно сформулированные задачи обеспечивают, помимо экономии сил и времени, возможность на этапе контроля точно оценить достигнутый результат по сравнению с планировавшимся, что также имеет мотивирующее значение.

В целях оптимизации работы подсистемы порождения речи КА в сфере формирования здорового образа жизни был проведен анализ рекомендаций по поддержанию здоровья — жанра популярно-медицинского дискурса, к которому прибегают специалисты при общении с широкой аудиторией в просветительских, ориентационных и фасцинативных целях (Календр 2015). Предметом нашего анализа стали тексты 26 популярных медицинских рекомендаций, основанных на материалах ВОЗ (Кобринский и др. 2019). В плане отражения в рекомендациях описанной выше психологической структуры деятельности были выделены следующие параметры:

[1] Работа частично поддержана грантами РФФИ 18-29-22027мк и 18-29-22047мк.

1. Полнота схемы деятельности М-Ц-З, представленной в рекомендации: *Для повышения уровня физической активности* (М) *рекомендуется вести дневник, в котором вы будете фиксировать свою физическую активность в течение дня по трем параметрам* (Ц): *формальная нагрузка (тренировки, если есть); неформальная двигательная активность; неподвижность (сидение, лежание)* (З) — полная структура; *Обратиться к врачу* (Ц) *для установления причин бессонницы* (М) — отсутствует компонент З.
2. Последовательность компонентов деятельностной схемы, выражающаяся в актуальном членении высказывания: прямая (в первом из приведенных примеров) или инвертированная (второй пример и пример ниже), что определяет психологические нюансы, связанные с восприятием относительной субъективной значимости темы и ремы в высказывании.
3. Мотивирующий аспект: эксплицитное мотивирование, характерное для рекомендации как провокативного речевого жанра (приведенные выше примеры), или имплицитное мотивирование, например, путем предоставления аффектогенной информации: *Увеличение потребления оливкового масла* (Ц) *на 25 г в сутки* (З) *связано с достоверным снижением риска инсульта на 24%* (М).

Оценка реального перлокутивного эффекта рекомендаций в зависимости от сочетания данных параметров — задача специального психолингвистического исследования. Предварительно можно считать, что целям КА в наибольшей степени соответствует представленность в тексте полной схемы М-Ц-З. Используемые технологии позволяют КА учитывать в коммуникации индивидуальные особенности конкретного пользователя, в частности, выбирать последовательность представления компонентов схемы М-Ц-З и способ выражения мотивирующего аспекта.

Литература

Календр А. А. 2015. Рекомендация как жанр популярно-медицинского дискурса // Известия Волгоградского государственного педагогического университета. №8 (103). С. 142–147.

Кобринский Б. А., Кадыков А. С., Полтавская М. Г., Благосклонов Н. А., Ковелькова М. Н. 2019. Принципы функционирования интеллектуальной системы динамического контроля факторов риска и формирования рекомендаций по здоровьесбережению // Профилактическая медицина. Т. 22. №5. С. 78–84. DOI: 10.17116/profmed20192205178

Кузнецова Ю. М. 2019. Психологические барьеры формирования саморегуляции здоровьесберегающего поведения // Вестник экспериментального образования (Электронный журнал). №4. URL: http://www.ppacademy.ru/wp-content/uploads/2019/11/142.pdf

Смирнов И. В., Панов А. И., Скрынник А. А., Чистова Е. В. 2019. Персональный когнитивный ассистент: концепция и принципы работы // Информатика и ее применения. Т. 13. Вып. 3. С. 105–113.

КОНТЕКСТНОЕ ПОЛЕ ДИАЛОГА И ПЛАТФОРМА ЭМОЦИОНАЛЬНЫХ ПЕРЕЖИВАНИЙ В АСПЕКТЕ АЛГОРИТМОВ КОММУНИКАЦИИ

Н. В. Лавров

vasilylavrov@yandex.ru

Санкт-Петербургский государственный педиатрический медицинский университет

Искусственный интеллект интегрируется в социальные отношения, создавая технологии, осуществляющие обработку естественного языка. Определение алгоритмов, задающих построение вербальных посылок с учетом эмоциональной насыщенности сообщений, является актуальной задачей конструкторов искусственного интеллекта. Задача решается с привлечением данных из области нейроинформатики и нейрофизиологии. В процессе диалога происходит обмен информационными посылками, отражающими фрагменты образов тех объектов, которым посвящен диалог. Результаты исследования формирования образа из дискретных информационных фрагментов под влиянием изменяющегося эмоционального состояния (Лавров 2016; Лавров, Рудинский 2016) расширяют знания о принципах коммуникации в условиях дефицита достоверной информации и при действии помехи. Отмечено, что участники переговоров неодинаково воспринимают полученные посылки в силу их коннотации с индивидуальным контекстом. Каждая посылка обладает эмоциональной составляющей, поэтому и ее формирование, и восприятие сопряжено с провокацией ассоциаций. Семантическая ассоциативность языковых единиц, используемых в информационных посылках, или, иначе говоря, коннотация, препятствует коммуникации. Препятствие преодолевается, во-первых, посредством согласования правил ведения устных и письменных переговоров. Имеется в виду конвенциональность используемых терминов, соблюдение этикета и установленных канонов, а также договор о процедуре переговоров, как это предусмотрено в медиации. Во-вторых, эффективность переговоров зависит от степени детализации и совпадения образов, которые выстраиваются у коммуникантов при рассмотрении предмета переговоров. Второй фактор является главным. Дело в том, что конвенциональность не обеспечивает конструктивного диалога, если рассматриваемый объект неодинаково воспринимается участниками переговоров, а построенные образы отличаются по структуре фрагментов и по системной организации. Совпадение образов, присущих обсуждаемому объекту, достигается благодаря общности контекстного поля. Эта общность вместе с принятой конвенциональностью создают условия для формирования коммуникативной системы, которая лежит в основе конструктивных переговоров (Лавров 2016).

Кроме двух отмеченных обстоятельств, выделяется третье. Эффективность коммуникации и мобилизация интеллекта обусловлена эмоциональными переживаниями участников переговоров. Именно эмоциональная составляющая переговоров мобилизует каналы коммуникации. Здесь уместно обратить внимание на принципиальное различие природы эмоций и эмоциональных проявлений. Эмоция — это не субъективный ответ на позитивные или негативные обстоятельства, а эволюционно заданный и генетически закрепленный паттерн экстренной перенастройки организма в ряду стандартных поведенческих ситуаций. Стабильность эмоционального механизма исключает сознательное управление. Эмоциональные проявления переживаются многообразием чувств, возникающих под воздействием мотиваций, актуализированных ситуационными обстоятельствами. Субъективная оценка обстоятельств происходит под влиянием внешних и внутренних факторов и поддается сознательному контролю (см. рис. 1).

Проведенное в 2001–2019 годах исследование позволило проследить динамику переговоров участников конфликтов по мере согласования представлений о предмете спора. Подробно методика изложена в публикации (Лаврова и др. 2013). При выяснении предмета спо-

ра регистрировали интересы и ожидания участников конфликта, а также их мнения относительно мотиваций поведения оппонента. Степень совпадения служила показателем, свидетельствовавшим о состоянии контекстного поля.

В докладе обсуждаются результаты наблюдений за тем, как переход от разрушительных мотиваций к восстановительным (Лаврова и др. 2013) влияет на вербальную и невербальную коммуникацию участников конфликтов и на принятие согласованного решения. Демонстрируется и обсуждается конструкция и влияние платформы эмоциональных переживаний (см. рис. 1) на формирование общего контекстного поля медиативных переговоров.

Рис. 1. Обстоятельства коммуникации в проблемной ситуации

Литература

Лавров В. 2016. Между мозгом и психикой. Образная психология. Saarbrücken: OSM.

Лавров В. В., Рудинский А. В. 2016. Формирование матрицы цельного образа при раздельном восприятии элементов комплексного объекта // Международный журнал прикладных и фундаментальных исследований. 7 (1), с. 91–95.

Лаврова Н. М., Лавров В. В., Лавров Н. В. 2013. Медиация: принятие ответственных решений. М.: ОППЛ.

КОМБИНИРОВАННЫЙ ТРИЛИНГВИЗМ КАК ИСТОЧНИК МОДЕЛИРОВАНИЯ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ[1]

Ю. Е. Лещенко

naps536@mail.ru

Пермский государственный национально-исследовательский университет

Т. И. Доценко

tamaradotsenko@bk.ru

Т. С. Остапенко

osttania@yandex.ru

Пермский государственный гуманитарно-педагогический университет

Многочисленные геополитические и социокультурные изменения конца XX — начала XXI в. (образование Евросоюза, открытость границ, активные миграционные процессы, создание двуязычных семей) привели к тому, что большая часть современного мира характеризуется как триязычная. Такая социальная ситуация обусловила закономерный поворот современной лингвистики от эпохи одно- и двуязычия к три- и мультиязычию. Наиболее общее определение мультилингвизма предложено в работе Р. Франческини, в которой в качестве основной черты мультилингвизма называется «фундаментальная человеческая способность обществ, групп и индивидов осуществлять коммуникацию на нескольких языках» (Franceschini 2009: 33–34). В рамках мультилингвального подхода языковая компетенция триязычного индивида осмысляется не как простая сумма двух или более языков, а как единая лингво-когнитивная конфигурация, характеризующаяся сложными процессами взаимодействий между языками (Herdina, Jessner 2002). В связи с этим традиционная бинарная оппозиция «естественный — учебный», прилагаемая к билингвизму, стала неприемлемой по отношению к трилингвизму. При классификации вариантов трилингвизма по способу усвоения языков различаются следующие варианты: 1) естественный способ усвоения всех трех языков; 2) естественный способ усвоения первого языка (Я1) и учебный способ усвоения второго языка (Я2) и третьего языка (Я3); 3) естественный способ усвоения Я1 и Я2 и учебный способ усвоения Я3. Таким образом, по способу усвоения языков триязычие может быть либо *естественным* (1), либо *комбинированным* (2,3). В последнем случае триязычие варьирует в зависимости от билингвальной базы, представляя собой либо *комбинированное триязычие, сформированное на базе учебного билингвизма* (2), либо *комбинированное триязычие на базе естественного билингвизма* (3) (Доценко и др. 2013).

Очевидно, что изучение разных вариаций комбинированного трилингвизма должно осуществляться в рамках единой модели, позволяющей применять универсальные паттерны для их описания. Такая интегрированная модель может быть положена в основу исследований искусственного интеллекта, а именно — послужить источником для моделирования искусственных нейронных сетей, репрезентирующих мультилингвальную речевую способность. Экспериментальным аналогом искусственной нейронной сети является ассоциативно-вербальная сеть, воспроизводящая основные закономерности взаимосвязей языковых единиц в сознании говорящего и реконструируемая по данным ассоциативных экспериментов с носителями одного / нескольких языков.

В данной работе представлена модель мультилингвальной ассоциативно-вербальной сети, построенной на материале экспериментов с комбинированными трилингвами (носителями коми-пермяцко — русского двуязычия естественного типа, изучающими английский язык в учебных условиях) разных возрастных групп. Результаты моделирования демонстри-

[1] Исследование выполнено при поддержке гранта РФФИ 17-29-09074 «Комбинированное триязычие и его влияние на языковую и когнитивную деятельность индивида: интегративная модель».

руют, что траектория развития комбинированного трилингвизма может быть представлена в виде трех последовательно сменяющих друг друга стадий:

А) первичное, относительно устойчивое состояние, которое характеризуется преимущественным сходством системы ассоциативных связей Я3 с системой связей доминантного языка (в нашем случае — русского);
Б) промежуточное, неустойчивое состояние, которое характеризуется изменением системы ассоциативных связей Я3 под влиянием активной конкуренцией систем связей двух усвоенных естественным образом языков (русского и коми-пермяцкого);
В) вторичное, относительно устойчивое состояние, характеризующееся сходством системы ассоциативных связей Я3 одновременно с системой ассоциативных связей Я1 и Я2.

Все три стадии сопровождаются, во-первых, интенсивной конкуренцией трех языковых систем в сознании говорящего и, во-вторых, активацией и формированием межъязыковых связей. Последовательное продвижение трилингва от одной стадии к другой открывает перспективу для дальнейшего освоения мультилингвального континуума.

Основные черты ассоциативно-вербальной сети в ее промежуточном неустойчивом состоянии (стадия Б), реконструированные на материалах ассоциативных экспериментов со студентами 2–5 курсов коми-пермяцко-русского отделения, изучающими английский язык, проявляются в следующем:

1. Триязычной сети свойственны только смешанные узлы с межъязыковыми и внутриязыковыми связями: их соотношение в сети составляет 0,41 и 0,59 соответственно. Степень покрытия пространства узлов межъязыковыми и внутриязыковыми связями колеблется в зависимости от функциональной активности каждого отдельного языка.
2. Межъязыковые связи формируют единую межъязыковую триаду, в которой наиболее сильными (двунаправленными и высокочастотными) оказались русско–коми-пермяцкие связи, менее сильными (однонаправленными и среднечастотными) стали англо–русские связи, слабыми (однонаправленными и низкочастотными) — англо–коми-пермяцкие связи.
3. Узлы трех языков оказались заполненными переводными и непереводными связями: без наличия хотя бы одного из этих типов связей коммуникация на трех языках становится принципиально невозможна. И те, и другие связи в англоязычных узлах представлены в равных долях — по 0,5. В русских и коми-пермяцких узлах большая доля приходится на непереводные связи: 0,86 и 0,64 соответственно.
4. Среди непереводных связей в узлах всех трех языков преобладают синтагматические (контекстуальные) связи, которые являются структурообразующими для речевого механизма и направлены на организацию уровня связности любого речевого высказывания.

Выявленные закономерности могут учитываться при построении систем искусственного интеллекта по принципу ассоциативно-вербальных сетей, репрезентирующих мультилингвальную ситуацию.

Литература

1. Franceschini R. 2009. Genesis and development of research in multilingualism: Perspectives for future research // Exploration of Multilingualism: Development of Research on L3, Multilingualism and Multiple Language Acquisition / L. Aronin, B. Hufeisen (eds). Amsterdam: John Benjamins. P. 27–61.

2. Herdina P., Jessner U. 2002. A Dynamic Model of Multilingualism: Perspectives of Change in Psycholinguistics. Clevedon; Buffalo; Toronto; Sydney.

3. Доценко Т. И., Лещенко Ю. Е., Остапенко Т. С. 2013. Кодовые переключения как межъязыковые взаимодействия в ситуации комбинированного билингвизма (на фоне становления профессиональной лингвистической компетенции) // Вопросы психолингвистики. № 2(18). С. 78–89.

ИСПОЛЬЗОВАНИЕ ПОКАЗАТЕЛЕЙ ДВИЖЕНИЯ ГЛАЗ ПРИ ИЗУЧЕНИИ ВОСПРИЯТИЯ ТЕКСТОВ НА РУССКОМ И АНГЛИЙСКОМ ЯЗЫКАХ

О. В. Ломтатидзе
Olya.l@mail.ru

А. С. Алексеева
nurochka_k@mail.ru

Уральский федеральный университет, Институт естественных наук и математики

Е. С. Исакова
isakova.evg@mail.ru

Уральский государственный медицинский университет

Чтение представляет собой сложный многоуровневый когнитивный процесс, включающий в себя прежде всего визуальное восприятие текста и извлечение информации с использованием механизмов рабочей памяти и селективного внимания (Иванов 2015: 16). Помимо чтения на родном языке нередко информация поступает в виде печатного материала на иностранном языке, чаще всего на английском. Чтение текстов на английском языке может вызывать дополнительные трудности, связанные как со сложностью формирования навыка, так и с более глобальной по сравнению с родным языком проблемой понимания (Соколова 2012: 64). При этом, в зависимости от развития навыков, при восприятии текстов могут применяться различные стратегии чтения (Сыса 2014: 228). Показатели движений глаз, определяемые методом электроокулографии, могут определяться когнитивными процессами, обеспечивающими реализацию процессов чтения, и связаны, с одной стороны, со свойствами предлагаемого текста, а с другой стороны, с особенностями визуального восприятия (Иванов 2015: 8).

В основу метода электроокулографии (ЭОГ) положено использование собственных электрических свойств глазного яблока. Электрическая ось глазного яблока примерно совпадает с его оптической осью, поэтому она может служить индикатором направления взора. Движения глаз регистрируются с помощью электродов, которые устанавливаются вокруг глазной впадины. При ЭОГ исключается прямой контакт прибора с глазом и не нарушаются естественные условия зрительной активности (Полевая 2017: 9).

В ходе использования метода могут быть зафиксированы саккады (макро- и микросаккады) — быстрые движения глазных яблок при перемещении взгляда с одной точки фиксации взора на другую при произвольном взгляде на предмет. Основными характеристиками, характеризующими саккады в методе ЭОР, являются амплитуда саккады, мВ, — прямо пропорциональна углу поворота глаз, длительность и скорость саккады. Между саккадами происходят фиксации на объекте, слове или предложении, во время которых происходит распознавание объектов. В силу анизотропности сетчатки движения глаз могут быть использованы как индикаторы перцептивного процесса и связанных с ним форм деятельности (Барабанщиков 2010: 244).

В экспериментальной части исследования приняли участие 40 студентов в возрасте от 18 до 25 лет, имеющих различные направления подготовки и уровни владения английским языком от А0 (стартового) до В2 (средне-продвинутого). Испытуемым предъявлялись два отрывка: на русском и английском языке. Отрывки были уравнены по типу и размеру шрифта, количеству строк и слов в строке, частотности слов, уровню сложности текста. Регистрация проводилась на инструментально-программном комплексе в системе Biopac Student Lab. В ходе работы снимались такие показатели, как амплитуда саккады, мВ, доля саккад в общем времени прочтения, % и скорость саккады.

Анализ результатов работы показал, что амплитуда ($F=3.68$, $p<0.05$) и скорость саккад ($F=1.3*107$, $p<0.05$) достоверно выше при чтении русского текста, чем английского. Результа-

ты, полученные по амплитуде саккад, отвечают имеющимся в литературе данным об увеличении длины саккады при развитии навыка чтения и свидетельствует о повышении когнитивной нагрузки при переходе от чтения отрывка на родном языке к чтению отрывка на иностранном языке. Скорость саккад, не связанная напрямую со скоростью чтения за счет возможного разброса продолжительности фиксации и количества регрессивных саккад, по-видимому отражает более легкую ориентацию в пространстве текста на родном языке по сравнению с иностранным.

Скорость саккад при чтении на русском языке достоверно ($F=5.48$, $p<0.05$) выше у испытуемых в группе с небольшим знанием английского языка (уровень А1). По-видимому, начало формирования навыка овладения иностранным языком активирует мозговые механизмы зрительного восприятия речи при чтении текста. Дальнейшее изучение английского языка (в нашем исследовании группы испытуемых с уровнем В1- В2) требует дополнительных психофизиологических ресурсов, вследствие изменения соотношения моторного и семантического чтения, что приводит к уменьшению скорости саккад, поскольку при увеличении скорости саккад выраженность выносливости (эргичность) уменьшается (Белых 2015: 143).

Проведенный корреляционный анализ показал, что структура взаимосвязей исследуемых параметров различается в зависимости от уровня знания английского языка. Количество и сила значимых корреляционных связей увеличивается с увеличением уровня овладения иностранным языком. В группах с уровнем знания языка А0 и А1 связей между показателями окуломоторной активности нет. В группе А2 и В1 наблюдаются связи между одноименными показателями восприятия (длительность и амплитуда саккады) текстов при чтении на русском и английском языках, при этом сила связей увеличивается с увеличением уровня овладения иностранным языком ($rs=0.55$ и $rs=0.82$ соответственно, $p<0,05$). В группе В2 структура корреляционных связей включает в себя взаимосвязи всех исследуемых параметров ($r=0.62\div0,82$, $p<0,05$) как внутри, так и между пробами на русском и английском языках. По-видимому, усложнение структуры связей, в совокупности с их усилением являются следствием формирования новых комплексных языковых связей, которые приводят к образованию единой структуры чтения. Исследованные параметры движения глаз имеют четко выраженную изменчивость в зависимости от уровня изучения иностранного языка и могут выступать маркерами формирования определенных когнитивных навыков восприятия текстов на родном и иностранном языках, отражая ведущие психофизиологические механизмы зрительного восприятия иноязычной письменной речи.

Литература

Барабанщиков В. А., Жегалло А. В. 2010. Методы регистрации движений глаз: теория и практика // Психологическая наука и образование, № 5. С. 240–254.

Белых Т. В., Иголкина Н. И., Зинченко Е. М. 2015. Взаимосвязь формально-динамических различий и особенностей глазодвигательной активности в процессе запоминания значения иностранных слов// Айтрекинг в психологической науке и практике М.: Когито-Центр.

Иванов В. В. 2015. Особенности окуломоторной активности у детей младшего школьного возраста в процессе чтения текстов различной сложности: автореф. дис. … канд. биол. наук. Архангельск.

Полевая А. В., Демарева В. А., Парин С. Б., Полевая С. А. 2017. Практикум по методу eye-tracking. Н. Новгород: Нижегородский госуниверситет.

Соколова Л. В., Черкасова А. С. 2012. Функциональное взаимодействие корковых зон головного мозга // Arctic Environmental Research. Биология. С. 59–65.

Сыса Е. А. 2014. Стратегии чтения на иностранном языке в техническом вузе // Вестник Сургутского государственного педагогического университета. №5 (32). С. 227–231.

СПЕЦИФИКА ДВИЖЕНИЙ ГЛАЗ ПРИ ЧТЕНИИ У ДЕТЕЙ С ДИСЛЕКСИЕЙ

А. А. Лопухина
alopukhina@hse.ru

С. М. Гольдина
smgoldina@edu.hse.ru

А. К. Лауринавичюте
alaurinavichute@hse.ru

О. В. Драгой
odragoy@hse.ru

НИУ «Высшая школа экономики»

Дислексия — это избирательная неспособность овладения навыками чтения, она относится к наиболее распространенным трудностям обучения, от которой в той или иной мере страдают около 10% детей (Sprenger-Charolles et al. 2011). Современные теории предполагают неврологические основы дислексии (Stein 2017) и среди причин ее возникновения называют нарушения фонологической обработки (Ziegler, Goswami 2005). Существующие видеоокулографические исследования показали, что дети и взрослые с дислексией испытывают больше трудностей при чтении длинных и низкочастотных слов (Rello et al. 2013; однако см. Hyönä, Olson 1995), делают больше саккад, а также чаще перечитывают слова (Jainta, Kapoula 2011), чем их типично развивающиеся сверстники. Между тем до сих пор практически не исследовалось, как трудности в фонологической обработке связаны с особенностями движений глаз при чтении у детей с дислексией. Наше исследование использует метод видеоокулографии и направлено на изучение того, как навыки фонологической обработки и свойства читаемых слов влияют на особенности движений глаз при чтении у русскоязычных школьников с фонологической дислексией в сравнении с типично развивающимися школьниками.

Метод. В исследовании приняли участие 76 учеников 1–4 классов: 29 детей с диагностированной фонологической дислексией (средний возраст 9.5 лет; 10 девочек) и 47 типично развивающихся детей (средний возраст 8.5 лет; 26 девочек). Дислексия была диагностирована специалистами-логопедами в Центре патологии речи и нейрореабилитации и варьировалась от легкой до тяжелой степени выраженности. Все дети прошли тест на невербальный интеллект «Прогрессивные матрицы Равена» (Raven 1938) и скрининг на первичные нарушения слуха и имели результаты в пределах нормы. Кроме того, мы дополнительно определили, что все типично развивающиеся дети находились в группе нормы согласно «Стандартизованной методике исследования навыка чтения» (СМИНЧ; Корнев 1997). Все дети прошли поведенческое тестирование, оценивающее их навыки фонематического анализа при помощи семи упорядоченных от простого к сложному субтестов из батареи тестов «Звуковой анализ русского языка» (ЗАРЯ; Дорофеева и др. 2019).

Материалом исследования движений глаз при чтении стал корпус из 30 предложений для исследования чтения у младших школьников, разработанный на основе Русского корпуса предложений (Laurinavichyute et al. 2019). Все слова (N = 228) в 30 предложениях были размечены по длине, частотности, частеречной принадлежности и количеству морфем. В ходе исследования дети читали предложения про себя, в то время как движения их глаз записывались при помощи видеоокулографа EyeLink 1000 Plus или EyeLink Portable Duo (SR Research) с частотой дискретизации 1000 Гц. После трети предложений детям задавались вопросы на понимание прочитанного. Мы оценивали, как на длительность фиксаций при чтении будут влиять свойства слов (длина, частотность, часть речи, морфемная сложность), амплитуда входящей саккады, особенности участников (класс обучения, принадлежность к группе нормы/дислексии и баллы по семи субтестам ЗАРЯ), а также взаимодействие между группой и остальными предикторами. При анализе мы контролировали длину предшествующего и последующего слов.

Результаты и выводы. По результатам анализа мы выявили следующие статистически значимые различия в длительности фиксаций при чтении между группами участников:

1) Во-первых, дети с дислексией дольше фиксируются на более длинных словах и словах с большим количеством морфем по сравнению с типично развивающимися детьми, что свидетельствует о трудностях лексической обработки у детей с дислексией. При этом фактор частотности слова оказывает сопоставимое влияние на чтение детей с дислексией и без нее: у всех детей меньшая частотность слова провоцирует более длительные фиксации на этом слове. Эти результаты лишь частично соотносятся с выводами предыдущих исследований (Hyönä & Olson 1995; Rello et al. 2013). Кроме того, дети с дислексией делают более короткие фиксации на служебных словах и местоимениях, что может говорить о менее глубокой обработке этих слов.

2) Во-вторых, исследование амплитуды входящей саккады показало, что дети с дислексией хуже извлекают информацию о слове, чем типично развивающиеся дети, если сделали саккаду не в оптимальное место внутри слова (середина слова), а ближе к концу слова. В этом случае для обработки слова детям с дислексией требуется сделать корректирующую саккаду, что замедляет чтение.

3) Наконец, трудности со сложным, но не с простым фонематическим анализом у детей с дислексией коррелируют с общим временем чтения и длительностью перечитывания слова. Такой результат означает, что дефицит при фонологической дислексии носит не элементарный характер нарушения фонематического слуха, а проявляется при повышенной нагрузке на фонологическую систему в особенно сложных заданиях.

Исследование выполнено при поддержке РФФИ, грант № 17–29–09122

Литература

Дорофеева С. В., Решетникова В. А., Лауринавичюте А. К., Ахутина Т. В., Драгой О. В. 2019. Исследование взаимосвязи навыков фонематической обработки и навыков чтения // Когнитивная наука в Москве: новые исследования. Материалы конференции 19 июня 2019 г. Под ред. Е. В. Печенковой М. В. Фаликман. М.: ООО «Буки Веди», ИППиП, 151–156.

Корнев А. Н. 1997. Нарушения чтения и письма у детей: Учебно-методическое пособие. СПб.: МиМ.

Hyönä J., Olson R. K. 1995. Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency // Journal of Experimental Psychology: Learning, Memory and Cognition 21(6), 1430–1440.

Jainta S., Kapoula Z. 2011. Dyslexic Children Are Confronted with Unstable Binocular Fixation while Reading // PLoS ONE 6(4), e18694.

Laurinavichyute A. K., Sekerina I. A., Alexeeva S., Bagdasaryan K., & Kliegl R. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian // Behavior research methods, 51(3), 1161–1178.

Raven J. C. 1938. Raven's progressive matrices: sets A, B, C, D E. Melbourne: Australian Council for Educational Research.

Rello L., Baeza-Yates R., Dempere-Marco L., Saggion H. 2013. Frequent Words Improve Readability and Short Words Improve Understandability for People with Dyslexia // IFIP Conference on Human-Computer Interaction. Berlin, Heidelberg: Springer, 203–219.

Sprenger-Charolles L., Siegel L., Jiménez J., Ziegler J. 2011. Prevalence and reliability of phonological, surface, and mixed profiles in dyslexia: A review of studies conducted in languages varying in orthographic depth // Scientific Studies of Reading 15, 498–521.

Stein J. F. 2017. Does dyslexia exist? // Language, Cognition and Neuroscience 33(3), 313–320.

Ziegler J. C., Goswami U. 2005. Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory // Psychological Bulletin 131(1), 3–29.

СПОСОБНОСТЬ ДАВАТЬ ПОКАЗАНИЯ У ЛИЦ С НАРУШЕНИЯМИ СЛУХА И РЕЧИ

Н. Е. Лысенко

nlisenko@yandex.ru

ФГБУ НМИЦПН им. В. А. Сербского Минздрава РФ

Целью публикации является обсуждение вопросов, возникающих при производстве судебно-психологической экспертизы (СПЭ) лиц с дефицитарностью слуха и речи: 1) связана ли дефицитарность речи и способность давать показания; 2) можно ли доверять показаниям, которые даются на спонтанном жестовом языке; 3) на какие критерии нужно опираться при производстве СПЭ обследуемого с дефицитарностью речи.

Одним из наиболее часто назначаемых видов СПЭ является оценка способности давать показания в ходе производства по уголовному делу. Способность считается сохранной, если у человека не нарушены, во-первых, способность воспринимать происходящее, во-вторых, способность воспроизводить случившееся. Однако, если человек может воспроизводить информацию, значит ли это, что он может ее сообщать? Дача показаний предполагает, что воспринятая информация не просто воспроизведена, но и сообщена в понятной форме сотрудникам судебно-следственных органов.

В пп. 2 п. 4 ст. 56 УПК РФ указано, что свидетель может «давать показания на родном языке или языке, которым он владеет» и допрос может производиться с участием переводчика. Но в редких случаях подэкспертные с дефицитарностью слуха вследствие социально-педагогической запущенности не владеют языком глухонемых, что исключает возможность перевода. При исследовании способности давать показания трудности вызывает экспериментально-психологическая проверка повторяемости показаний и сохранности памяти на детали обстоятельств. Проведение экспериментально-психологического исследования в таких случаях требует использования тестовых методик, которые пригодны для сохранного сенсорного качества.

В феврале 2020 года была назначена комплексная судебная психолого-психиатрическая экспертиза потерпевшего О. Согласно материалам уголовного дела, двигаясь в ночное время на велосипеде вдоль тротуара, находясь на участке местности, расположенном примерно в 20 метрах от строящегося дома, О. упал в котлован, образованный на месте проводившихся строительно-ремонтных работ. По данным судебно-медицинской экспертизы действиями строительной компании по признаку опасности для жизни ему был причинен тяжкий вред здоровью. В протоколе допроса О. в качестве потерпевшего указано, что следователем предпринималась попытка диалога с О., установлено, что он «не контактен», не может говорить, читать, писать. Сурдопереводчик И. показала, что О. не понимал язык жестов и ничего не мог пояснить; попытка общения с ним с помощью буквенного обозначения жестов (дактология) также не удалась. Со слов свидетеля Ж. (опекун потерпевшего), в протоколе допроса указано, что «в ходе «общения-демонстрации» жестов, она поняла, что он катался на велосипеде, разогнался быстро, время было ночное, в какой-то момент он упал в котлован», «место падения он показал ей самостоятельно», «в разрыве была вода». Среди прочих, следователем был поставлен вопрос к экспертизе: мог ли потерпевший О. по своему психическому состоянию правильно воспринимать обстоятельства, имеющие значение для уголовного дела и давать показания?

При экспериментально-психологическом исследовании поведение О. было адекватным ситуации — он легко шел на контакт, войдя в кабинет, выказывал приветствие кивком головы, пропускал вперед сотрудников женского пола. О. чутко реагировал на обращенные к нему вопросы, внимательно отслеживал жесты и мимику собеседника, пытался жестами рассказать о себе и о случившемся. Произносил отдельные звуки, некоторые целые слова («мама», «папа», «Жанна»), называл свое имя — Олег.

При исследовании когнитивной сферы с использованием преимущественно визуального стимульного материала, нарушений памяти, внимания и мышления не выявлено: О. улавливал смысл заданий и алгоритм выполнения, удерживая в работе инструкции. Ему был доступен поиск простых обобщающих понятий, он правильно понимал и передавал суть ситуаций межличностного взаимодействия на визуальном материале, не соглашался с заведомо неверными решениями, предложенными экспертом. В личностной сфере О., на фоне имеющегося дефекта развития, связанного с глухонемотой, микросоциальной и педагогической запущенностью на ранних этапах развития, выявлялись добродушие, живость и адекватность эмоциональных проявлений при развитой жестовой коммуникативной способности, позволяющей адаптироваться в социуме, осведомленность в социальных и правовых нормах поведения, сформированность бытовых навыков, настроенность на преодоление сложных ситуаций.

В процессе производства экспертизы О. самостоятельно произносил лишь отдельные звуки, слоги, однако активно жестикулировал. Так, складывая ладони домиком, показывал, что в день произошедшего был дома, затем поехал кататься на велосипеде (показывал руками, как крутил педали и быстро ехал); попеременно раскрывая и собирая пальцы правой руки чуть выше уровня головы показывал, что везде светили фонари, но в одном месте свет не горел, было темно и плохо видно дорогу (закрывал глаза руками). Затем, сложив руки «рыбкой» и подавая тело вперед, показывал, как «полетел», ударился, испытывал боль (тер место удара на спине). Сжимаясь всем телом, показал, что испытывал физические мучения; не мог дышать, было горячо, когда упал на трубы, не с первого раза вылез из ямы. После этого он пошел домой, стучался в дверь, но Ж. не открывала, и ему пришлось ждать утра, очень хотел спать и засыпал сидя, так как лечь не мог из-за сильной боли в спине. Сообщал, что утром Ж. открыла дверь, он ей показал дорогу к яме, где остался лежать его велосипед, и потом его отвезли в больницу (показывает крест пальцами на лбу).

Описанный случай показывает, что при проведении допросов лиц с дефицитарностью слуха и речи, у сотрудников правоохранительных органов возникают сомнения в сохранности их способности давать показания. При экспертно-психологическом исследовании таких лиц необходимо уделять внимание не только уровню развития интеллектуально-мнестической сферы, но и надтекстовым способам выражения мыслей, опирающимся на специфику русского жестового и сопровождающего его звукового дискурса (Ганина 2005; Прозорова 2009; Борисова 2016). Значимыми становятся сохранность способности понятно излагать на индивидуальном языке жестов и звуков ответы на вопросы, наполненность высказываний корректной коннотацией. При сохранности интеллектуальной сферы изложение обстоятельств юридически значимого события на привычном русском жестовом языке с сопровождением мимики и звуков, не является ограничением способности к даче показаний.

Литература

Ганина В. В. 2005. Невербальные компоненты коммуникации, отражающие эмоциональные реакции человека: гендерный аспект (на материале англ. яз. худож. текстов). Автореф. дисс. … канд. филолог. наук. М.: РГБ.

Прозорова Е. В. 2009. Маркеры локальной структуры дискурса в русском жестовом языке. Автореф. дисс. … канд. филолог. наук. М.: МГУ.

Борисова Л. В. 2016. Лингвофилософские аспекты русского жестового дискурса // Экономические и социально-гуманитарные исследования № 4 (12), с. 63–68.

ЛЕКСИЧЕСКОЕ РАЗНООБРАЗИЕ У МОЛОДЫХ И ПОЖИЛЫХ ЛЮДЕЙ: ЭКСПЕРИМЕНТ НА ЗАКАНЧИВАНИЕ ПРЕДЛОЖЕНИЙ

С. А. Малютина Е. Д. Лапин М. Д. Терехина
s.malyutina@gmail.com
Национальный исследовательский университет «Высшая школа экономики»

Одним из основных факторов, обуславливающих индивидуальные различия в понимании и порождении языка, является возраст. С возрастом происходят изменения на всех уровнях языка, от фонологического до дискурсивного (Burke, Shafto 2008). Парадоксальными можно назвать изменения на лексическом уровне. С одной стороны, с возрастом непрерывно увеличиваются знания о мире, а с ними и словарный запас. С другой стороны, с возрастом постепенно затрудняется сам процесс лексического доступа к словам: лексический доступ замедляется, пожилые люди чаще испытывают ощущение «на кончике языка», они более уязвимы для психолингвистических эффектов конкуренции между словами-«соседями» и т. д.

Учитывая подобную двунаправленность возрастных изменений на лексическом уровне, сложно однозначно предположить, как они могут влиять на лексическое разнообразие при порождении языка: будет ли оно повышаться у пожилых людей в связи с бóльшим словарным запасом или же понижаться в связи с более затруднённым лексическим доступом. Действительно, эмпирические данные по этой теме пока неоднозначны. И данные о выполнении экспериментальных заданий (например, задания на порождение слов из какой-либо семантической категории за определённый период времени), и анализ спонтанной речи показывают противоречивые результаты по возрастным различиям в лексическом разнообразии (Kavé & Goral 2017).

Целью нашего исследования было привнести новые эмпирические данные по теме возрастных различий в лексическом разнообразии. Мы использовали экспериментальное задание на заканчивание предложений на материале русского языка. Для этого мы составили стимульный материал из 120 предложений, в которых было пропущено последнее слово — существительное. Предложения варьировались по предсказуемости последнего слова: от высокой («*Учитель написал пример мелом на …*») до низкой («*На рынке фермер недорого продавал …*»). Предложения были сбалансированы по части речи и синтаксической роли предпоследнего слова, а также по ожидаемой абстрактности/конкретности и одушевлённости последнего слова.

Экспериментальная парадигма была запрограммирована в программе DMDX. Участники должны были пословно читать предложения на экране компьютера, нажимая на кнопку для перехода к следующему слову. Вместо последнего слова предъявлялся прочерк, и участники должны были завершить предложение вслух одним словом. Предложения предъявлялись в рандомизированном порядке.

В основном эксперименте приняли участие 86 молодых (средний возраст 22 года, диапазон 18–32 года) и 86 пожилых (средний возраст 68 лет, диапазон 60–82 года) людей. Для каждого стимульного предложения в каждой группе участников в качестве меры лексического разнообразия была вычислена информационная статистика *H*. Аналогично её применению в (Snodgrass & Vanderwart 1948), статистика *H* вычислялась по следующей формуле, где *k* — количество уникальных завершений в возрастной группе, p_i — вероятность завершения *i* в возрастной группе:

$$H = \sum_{i=1}^{k} p_i \log_2\left(1/p_i\right)$$

Статистика *H* оказалась значимо выше у молодых людей, чем у пожилых ($p < 0{,}0001$), что указывает на меньшее лексическое разнообразие при заканчивании предложений пожилыми людьми.

Однако данный эксперимент не предоставляет каких-либо данных о том, почему именно лексическое разнообразие было понижено у пожилых людей. Нельзя было исключить, что этот эффект мог быть обусловлен не столько психолингвистическими причинами, сколько общей нейрокогнитивной стратегией экономии усилий. Многие исследования в области неязыковых когнитивных функций показывают, что пожилые люди могут осознанно или неосознанно «экономить» когнитивные ресурсы (Hess 2014). Поэтому мы провели дополнительный эксперимент, призванный проверить, не является ли меньшее лексическое разнообразие у пожилых людей искусственным эффектом вследствие экономии усилий в контексте прохождения эксперимента.

По процедуре и стимульному материалу дополнительный эксперимент был идентичен основному, однако в нём были изменены инструкции. В отличие от основного эксперимента, участникам давалась инструкция закончить предложение не «любым подходящим по смыслу словом», а «подходящим словом, которым, по вашему мнению, не закончат предложение другие участники». Эта инструкция была призвана мотивировать участников осознанно отказаться от стратегии экономии когнитивных ресурсов и использовать максимально разнообразные лексические средства.

В дополнительном эксперименте приняли участие 20 молодых (средний возраст 23 года, диапазон 19–32 года) и 20 пожилых (средний возраст 72 года, диапазон 61–81 год) людей. Несмотря на изменение инструкций, был воспроизведён результат основного эксперимента: статистика *H* вновь оказалась значимо выше у молодых людей, чем у пожилых ($p < 0{,}001$), то есть лексическое разнообразие при заканчивании предложений вновь оказалось ниже у пожилых людей.

Таким образом, данное исследование продемонстрировало меньшее лексическое разнообразие у пожилых людей, чем у молодых, при выполнении экспериментального задания на заканчивание предложений. Дополнительный эксперимент показал, что эти различия не являются результатом экономии когнитивных усилий, а отражают психолингвистические изменения в лексическом доступе.

Литература

Burke D. M., & Shafto M. A. 2008. Language and aging. In F. I. M. Craik & T. A. Salthouse (eds.), The handbook of aging and cognition (p. 373–443). Psychology Press.

Hess T. M. 2014. Selective Engagement of Cognitive Resources: Motivational Influences on Older Adults' Cognitive Functioning // Perspectives on Psychological Science, 9(4), pp. 388–407.

Kavé G., & Goral M. 2017. Do age-related word retrieval difficulties appear (or disappear) in connected speech? // Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition, 24(5), pp. 508–527.

Snodgrass J. G., & Vanderwart M. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity // Journal of Experimental Psychology: Human Learning and Memory, 6(2), pp. 174–215.

ОБ АВТОМАТИЧЕСКОЙ ИДЕНТИФИКАЦИИ ТЕКСТОВ С ЗАДАННОЙ МОДАЛЬНОСТЬЮ (НА МАТЕРИАЛЕ ТЕКСТОВ СЕТЕВОЙ КОММУНИКАЦИИ С ТЕМАТИКОЙ ПРОФЕССИОНАЛЬНОГО КРИЗИСА)[1]

В. А. Мишланов
vmishlanov@yandex.ru

Л. А. Каджая
kadzhaya@psu.ru

Пермский государственный национальный исследовательский университет

Е. Н. Махмутова
Московский государственный институт международных отношений (МГИМО-университет) МИД РФ

И. В. Смирнов
ivs@isa.ru
ФИЦ «Информатика и управление» РАН

1. В условиях обострения экономического кризиса чрезвычайно возрастает актуальность исследований, связанных с проблемами развития человеческого капитала. Представляется, в частности, что в настоящее время весьма востребованными становятся междисциплинарные изыскания, ориентированные на оптимизацию цифровых технологий информационного и психологического сопровождения человека, оказавшегося в ситуации нового профессионального выбора. Особое внимание в ходе психо- и социометрических исследований профессиональных сообществ должно быть уделено задаче обнаружения кризисных явлений, могущих вызвать негативные изменения в психологическом состоянии как отдельного человека (разочарование, неуверенность в себе, утрата перспектив и растущая апатия), так и тех или иных социальных групп и общества в целом.

Аффективно насыщенные психологические состояния человека, надо думать, стимулируют коммуникативную активность, в том числе в сетевом контенте, в связи с чем появляется благоприятная возможность привлечения к решению указанных задач методов искусственного интеллекта (ИИ).

2. Апробированные в компьютерной лингвистике практики смыслового анализа текстов на естественном языке (в частности, с использованием методов *тематических групп слов* или *ключевых слов* (Ананьева и др. 2016)) ориентированы на считывание пропозитивного (диктумного) содержания текста, тогда как для измерения психического состояния индивидуумов и настроений в социальных группах на передний план выдвигается проблема обучения машины пониманию модальности текста — идентификации субъективных компонентов текста, дающей ключ к реконструкции «ego-смыслов» (Бенвенист 1974), в том числе коммуникативных интенций говорящего, часто не выражаемых сегментными средствами языка.

Автоматическое извлечение информации из текста обеспечивается ныне машинным анализом на всех языковых уровнях (Толдова, Бонч-Осмоловская 2019: 515, 517). Проблема, однако, заключается в том, что алгоритмы анализа отдельной коммуникативной единицы (высказывания) в большинстве случаев недостаточны не только для считывания имплицитных смыслов (обычно легко устанавливаемых в реальном речевом акте благодаря сверхсегментным маркерам), но даже для разрешения полисемии (без чего не может быть достигнуто понимание эксплицитной фактуальной информации). А между тем развитие современной «инфосферы» (в частности, возникновение такого коммуникативного феномена, как соци-

[1] Исследование проводится при финансовой поддержке РФФИ в рамках научного проекта № 19-29-07163 мк.

альные сети) ставят перед исследователями ИИ новые задачи, требующие извлечения из текста имплицитной информации (Толдова 2019: 533).

Алгоритмы семантического анализа в оптимальном варианте не ограничиваются выделением ключевых слов и маркирующих дискурс коллокаций, а охватывают широкий комплекс явлений, относящихся к речевой системности (Кожина 2014) и отражающих, в соответствии со сложившимися дискурсивными правилами, как специфику речемыслительных действий автора текста, так и его эмоционально-волевые состояния (интенции), в том числе не выражаемые сегментными показателями (представленные нулевыми маркерами модуса, значение которых закрепляется за высказываниями с заданными семантическими и формально-грамматическими параметрами).

3. Для решение очерченных здесь задач от лингвистов и программистов требуется выявление как можно более широкого круга формальных (сегментных) речевых признаков модусных смыслов в текстах, относящихся к определенным речевым жанрам и обусловленных определенными дискурсивными правилами. Автоматическую идентификацию модусного содержания вербального текста предлагается основывать на методе реляционно-ситуационного анализа (РСА) (Осипов и др. 2008), который дал хорошие результаты при разработке моделей распознавания воплощаемых в тексте ментальных действий. При таком подходе используются лингвистические шаблоны, в которых заданы системно-речевые (дискурсивные) признаки-маркеры, объективирующие ментальные (речемыслительные) действия и состояния (в том числе интенции автора текста).

Полагаем, что настройка лингвистических анализаторов в системах ИИ на восприятие модусных компонентов содержания текста и поиск методов решения возникающих при этом задач значимы не только в прикладном аспекте, но и с точки зрения теоретического развития целого комплекса наук, вовлеченных в исследование проблем ИИ.

Эмпирической базой исследования послужили размеченные вручную (интроспективно) тексты, тематически связанные с проблемами профессионального образования и кадровой мобильности, при этом внимание фокусировалось на средствах выражения оценочного содержания и волюнтативной модальности. Для разработки программ автоматического мониторинга сетевого контента используются как лингвистические методы, на основе которых формируются *тематические группы слов*, так и методы ИИ, в частности, метод выделения *ключевых слов* (Мбайкоджи и др. 2012). Сопоставление результатов этих двух подходов к решению проблемы тематической идентификации текстов позволит определить перспективы оптимизации систем автоматического анализа.

Литература

Ананьева М. И., Девяткин Д. А., Зубарев Д. В., Осипов Г. С., Смирнов И. В., Соченков И. В., Тихомиров И. А., Швец А. В., Шелманов А. О. 2016. TextAppliance: поиск и анализ больших массивов текстов // КИИ-2016. Труды конференции. Т. 3. С. 220–228.

Бенвенист Э. 1974. О субъективности в языке // Бенвенист Э. Общая лингвистика. М.: Прогресс. С. 292–300.

Кожина М. Н. 2014. Речеведение. Теория функциональной стилистики. М.: Флинта: Наука.

Мбайкоджи Э., Драль А. А., Соченков И. В. 2012. Метод автоматической классификации коротких текстовых сообщений // Информационные технологии и вычислительные системы. № 3. С. 93–102.

Осипов Г. С., Смирнов И. В., Тихомиров И. А. 2008. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений, 2. С. 3–10.

Толдова С. Ю. 2019. Извлечение информации из текста // Кибрик А. Е. и др. Введение в науку о языке. М.: Буки Веди. С. 528–534.

Толдова С. Ю., Бонч-Осмоловская А. А. 2019. Автоматическая обработка текста // Кибрик А. Е. и др. Введение в науку о языке. М.: Буки Веди. С. 513–527.

ФАКТОР ПРОСОДИКИ В ВОСПРИЯТИИ КОММУНИКАТИВНОЙ СЦЕНЫ С ПРОСТРАНСТВЕННЫМ РАЗДЕЛЕНИЕМ ИСТОЧНИКОВ РЕЧИ И РЕЧЕПОДОБНОЙ ПОМЕХИ[1]

Е. А. Огородникова
elena-ogo@mail.ru

О. В. Лабутина
oliolo@mail.ru

Институт физиологии им. И.П. Павлова РАН

И. Г. Андреева
ig-andreeva@mail.ru

А. П. Гвоздева
kukumalu@mail.ru

Ю. А. Баулин
ghost7104@mail.ru

Институт эволюционной физиологии им. И.М. Сеченова РАН

В контексте современных исследований биологических и искусственных нейронных сетей, создания интерфейсов «мозг-компьютер» актуальной задачей остается изучение психофизических коррелят и механизмов слухового восприятия речи в условиях сложной коммуникативной сцены, близкой к ситуации «вечеринки» (Cherry 1953). Известно, что в процессы восприятия в таких условиях с необходимостью вовлечены сенсорные, информационные и лингвистические компоненты речевого взаимодействия. Наиболее значимыми из них являются: пространственная избирательность и свойства бинаурального слуха, голосовые характеристики дикторов; контекстная связность и интонация речевого высказывания (Андреева 2018; Bronkhorst 2015; Darvin et al. 2003).

В отношении пространственной избирательности одной из наиболее изученных является ситуация разделения источников речи и маскера по азимутальной координате. Для нее разработан ряд моделей, а также показана разница в уровнях разборчивости речи порядка 12 дБ (Litovsky 2012). В значительно меньшей степени изучен вопрос в отношении пространственных конфигураций, при которых целевой речевой и маскирующий источники расположены под одинаковым азимутальным углом, но имеют разную удаленность от слушателя. В этом направлении был выполнен ряд исследований, в которых расстояние до источника речи изменялось только в пределах не более 1 метра (Westermann, Buchholz 2015). Их результаты свидетельствовали о влиянии расстояния на разборчивость речи, особенно, если маскер был перцептивно близок к целевому сигналу (речеподобная помеха).

В нашей работе диапазон исследуемых расстояний был увеличен в соответствии с общими параметрами удаленности коммуникантов при естественном речевом взаимодействии в условиях так называемого ближнего поля — от 1 до 4 м. Моделировались 2 типа пространственных сцен: 1) источники речи и речеподобной помехи расположены перед испытуемым и разделены по расстоянию на 1, 2 и 3 м (Andreeva et al. 2019); 2) речеподобный шум распределен вокруг головы испытуемого, источник речи локализован в соответствии с вариантами в предыдущей схеме (Gvozdeva et al. 2019). Оценивалась вероятность обнаружения целевого речевого сигнала интенсивностью 50 дБ УЗД при разном уровне маскера (55, 58, 61 или 64 дБ УЗД). Для этого использовали метод постоянных рядов с двухальтернативным, двухинтервальным вынужденным выбором. В экспериментах приняли участие 16 (сцена 1) и 17 (сцена 2) взрослых испытуемых с нормальным слухом. Речевые сигналы представляли собой 8 двусложных слов русской речи с близкими показателями частотности по данным Национального корпуса русского языка (Ляшевская, Шаров 2009), произнесенных 4 дикторами разного пола с частотой основного тона голоса (ЧОТ) от 110 до

[1] Работа выполнена в рамках государственного задания (темы № АААА-А18-118013090245-6 и АААА-А18-118050790159-4) при частичной поддержке РФФИ (проект № 18-015-00296).

240 Гц. В качестве речеподобной помехи использовали результат микширования всех тестовых речевых стимулов (babble-шум). Измерения производились в условиях звукоизолированной безэховой камеры объемом 62.5 куб. м.

Полученные данные свидетельствовали о зависимости вероятности обнаружения речевого сигнала на фоне речеподобной помехи от разделения их источников в пространстве. Степень проявления этого эффекта была максимальной в условиях сцены 1 и при расстояниях между источниками речи и маскера, превышающих 2 м. При этом нахождение целевого речевого источника в близкой или далекой от слушателя позиции не оказывало значимого влияния на результат восприятия. В то же время условно просодические характеристики речевых сигналов, в качестве которых рассматривались тембр голоса (пол диктора, индивидуальные особенности) и слогоритмическая структура тестовых слов, выступали как значимый фактор, влияющий на вероятность выделения целевого речевого сигнала.

Таким образом, было экспериментально подтверждено, что пространственное разделение близких по перцептивным характеристикам источников речи и маскирующего сигнала в условиях, моделирующих ситуацию «вечеринки», способствует обнаружению речевой «цели». Полученные данные позволяют утверждать, что в условиях свободного поля сенсорные (бинауральные) и лингвистические (просодика) признаки оказываются наиболее эффективными с точки зрения выделения целевого речевого сигнала на фоне речевого маскера в пределах диапазона расстояний коммуникативного взаимодействия (1–4 м).

Результаты работы представляют особую важность для развития технических средств слухопротезирования, создания слуховых интерфейсов при недоступности зрительного канала связи, повышения эффективности бимодальных интерфейсов «мозг-компьютер», разработки систем виртуальной реальности и роботизированных устройств с возможностями пространственного выделения целевого источника речи в условиях многокомпонентной акустической среды.

Литература

Андреева И. Г. 2018. Пространственная избирательность слуха при распознавании речи на фоне речевого маскера // Физиология человека. № 2. С. 124.

Ляшевская О. Н., Шаров С. А. 2009. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник.

Andreeva I. G., Dymnikowa M., Gvozdeva A. P., Ogorodnikova E. A., Pak S. P. 2019. Influence of separation by the distance of the sources of speech and noise on speech signal detection // Acta Acust. United Acoustica. V. 105, pp. 484–491.

Bronkhorst A. W. 2015. The cocktail-party problem revisited: early processing and selection of multi-talker speech // Atten. Percept. Psychophys. V. 77. N5. P. 1465.

Cherry E. C. 1953. Some experiments on the recognition of speech, with one and with two ears // J. Acoust. Soc. Am. V. 25. N5, p. 975.

Darvin C. J., Brungart D. S., Simpson B. D. 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one or two simultaneous talkers // J. Acoust. Soc. Am. V. 114, pp. 2913–2922.

Gvozdeva A., Smirnova V., Labutina O. 2019. Speech detection in spatially distributed speech-like noise // Neural Networks and Neurotechnologies. SPb: VVM, pp. 52–59.

Litovsky R. Y. 2012. Spatial release from masking // Acoustics today. N8, pp. 18–24.

Westermann A., Buchholz J. M. 2015. The influence of informational masking in reverberant, multi-talker environment // J Acoust Soc Am. V. 138 (2), pp. 584–593.

«ЯЗЫК И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»: НЕФОРМАТНЫЕ СООБРАЖЕНИЯ О ФОРМУЛИРОВКЕ ТЕМЫ ФОРУМА

П. Б. Паршин

pparshin@mail.ru

Московский государственный институт международных отношений (Университет) МИД РФ

Московский государственный лингвистический университет

Нижеследующие тезисы носят неформатный характер и фактически претендуют на пленарный доклад, к представлению которого автора никто не приглашал. Однако воспроизведенная выше формулировка темы такова, что до всякой попытки высказаться по какому-либо из вопросов, перечисленных в «Объявлении» в качестве интересных для обсуждения, она вызывает реакцию типа «Не могу молчать». Посему, сознавая свою уязвимость, осмелюсь все-таки поделиться некоторыми мыслями, опуская при этом формальные ссылки — на них просто нет места.

Что за «и»? Союз *и* в обсуждаемой конструкции может быть нейтральным маркером конъюнкции, но может указывать и на наличие каких-то понятийных регионов, «горячих точек», в которых изучение языка и накопленные знания о нем, с одной стороны, и упражнения в моделировании интеллектуальной деятельности и полученные при этом результаты — с другой оказываются в з а и м о р е л е в а н т н ы м и. По мнению автора, таких регионов (было), как минимум, четыре: это (1) попытки проникнуть в устройство, содержание и механизмы мышления через язык (программа далеко не новая, в отрефлексированном виде представленная уже в Средние Века у модистов, а реально работавшая еще в античности); (2) дискуссия о том, является ли адекватное речевое поведение однозначным индикатором наличия интеллекта (тест Тьюринга — «Китайская комната» Дж. Сёрла — «Корейская комната У. Рапопорта); (3) усилия, направленные на выявление роли знаний в процессах ЕЯ общения и, соответственно, на разработку средств их извлечения, представления и использования с прикладной целью компьютерного моделирования такового общения (по факту в основном п о н и м а н и я ЕЯ, признаваемого заведомо асимметричным относительно производства речевых произведений, — канонический искусственный интеллект и Natural Language Processing как его ядро в 1970–1980-х гг.); и (4) формирование когнитивной науки как теории искусственного интеллекта (В.М. Сергеев), а также когнитивной лингвистики, предполагающей использование знаний о мыслительных структурах и процессах как инструмента объяснения лингвистических фактов; вопреки распространенному мнению, ИИ стоял куда ближе к колыбелям Cognitive Science и Cognitive Grammar, нежели психология. При этом надо понимать, что для когнитивной лингвистики ИИ был источником скорее познавательных индульгенций, метафор и гипотетических объяснительных конструктов, проверяемых языковым материалом, нежели фактуальных знаний о когнитивных процессах и системах. Утверждая обратное, легко попасть в порочный круг (см. п. (1) выше).

Интересно, что в формулировке темы форума конструкции с прилагательным *когнитивный* отсутствуют.

Существуют ли в настоящее время такие точки интенсивного взаимодействия? Автор в этом серьезно сомневается (ср. также п. 3 ниже), но рад бы был услышать возражения.

Два канона и два облака. Конъюнкты в формулировке темы, если не сужать их понимание в оговоренном выше духе, оказываются весьма разнородными и открытыми для различных пониманий; фактически это некоторые понятийные облака. О языке не приходится и говорить, но что понимается хотя бы под искусственным интеллектом? Исследовательское направление, или артефакт, способный демонстрировать поведение, которое мы готовы

признавать разумным (между тем, с критериями разумности все обстоит не так просто), или конкретная разновидность такого артефакта, основанная на нейросетевых технологиях и способная обрабатывать Big Data, как можно порой заключить из публикаций последних лет — правда, в основном все-таки не собственно научных? История ИИ насчитывает без малого 80 лет (без предыстории), и идея нейронных сетей возникла у У. Мак-Калока и У. Питтса раньше, чем сложилась ньюэлловско-саймоновская парадигма символьной обработки информации, в которой нашлось место лингвистам, одно время даже вполне комфортабельное, позволявшее не только решать прикладные задачи в интересах NLP-систем (сейчас эта аббревиатура узурпирована совсем другим видом деятельности), но и выдвигать идеи, значимые для развития ИИ. Канонический ИИ представлял собой своего рода инженерную семиотику, работавшую с дискретными объектами, вполне привычными для лингвистов (ср. языки представления знаний); нейронные же сети предполагают использование математических моделей, разработанных для описания недискретных объектов. А стало быть — и других профессиональных компетенций. Язык, конечно, тоже не вполне дискретен, и поле для взаимодействия здесь имеется, но каноническая лингвистика не приспособлена для такого рода исследований, как не приспособлена она и для описания реального дискурса, ибо всегда была — и остается — наукой «сворачивающей», направленной на выявление общих закономерностей и их компактное моделирование (признанный шедевр — «Восьмикнижие» Панини), а детальное описание реального дискурса рискует оказаться куда более пространным, чем он сам. Излишне говорить о том, что канонический ИИ описывает и, возможно, отчасти объясняет структуры и функции мышления (и языка), а нейросетевой… см. «Я, робот» А. Азимова: три закона роботехники как поведенческие ограничители и неизвестно что в позитронном мозгу: здравствуй, черный ящик. И то, что нейросетевой ИИ работает, дает основания не только для оптимизма, но и для пессимизма.

Язык и «другие вопросы, укладывающиеся в общую тематику форума». Сказанное в этом пункте менее всего является критикой в адрес программного комитета: поиск потерянного под фонарем, потому что там светлее, или в бюро находок, потому что там много чего есть, — это, в принципе, вполне рациональное и продуктивное поведение. Нельзя, однако, не заметить, что бóльшая часть перечисленных в «Объявлении» вопросов (за исключением первого и последнего, третий — под большим вопросом), безусловно могущих быть уложенными в «двухоблачную» формулировку темы и интересно там повернутыми, все-таки не представляется первостепенно значимыми для разговора о взаимодействии языка и ИИ. Скорее они отражают структуру интересов и научные активы составителей этого списка. Что, опять же, является вполне правомерным подходом и весьма информативно в плане характеристики некоторых участков современного отечественного исследовательского поля.

В порядке извинения. Чтобы не выглядеть критиком, не желающим иметь дело ни с какой эмпирией, автор готов предъявить в каком-либо виде дополнение, посвященное когнитивным (Sic!) процессам в понимании рекламы.

РЕЧЕВОЕ ПОВЕДЕНИЕ АКТОРОВ СОЦИАЛЬНЫХ МЕДИА В ГОРОДСКИХ КОНФЛИКТАХ: НЕЙРОСЕТЕВОЙ ПОДХОД

М. А. Пильгун

mpilgun@iling-ran.ru

Институт языкознания РАН

В докладе будет представлена модель предиктивной аналитики городских конфликтов на материале речевого поведения акторов социальных медиа с использованием нейросетевых технологий. Цифровая среда позволяет автоматически получать большие объемы данных в реальном времени, реализовать междисциплинарные проекты (Kullkarni and al. 2018; Gonçalves 2017; Meiselwitz 2019; Shibuya 2020 и др.). Цель исследования — анализ восприятия, речевых и коммуникативных особенностей акторов в конфликтном онлайн-взаимодействии, а также построение прогнозных моделей. Материалом послужили социальные сети, видео, микроблоги, блоги, мессенджеры, форумы, отзывы, посвящённые строительству хордовых магистралей (Северо-западной, Северо-восточной и Юго-восточной) г. Москвы. Сбор данных: 1.08.2019 00:00 — 30.09.2019 23:59. Объем контента — 9 670 141, активных акторов — 681, лояльность — 43,5, вовлеченность — 9 639, аудитория — 4 455 912. Одна из проблем использования нейросетевых технологий — воспроизводимость результатов в данном исследовании решается с помощью параллельного статистического анализа данных, описания динамических процессов с использованием стохастических дифференциальных уравнений. В ходе исследования был разработан алгоритм построения моделей на основе комбинации описания динамических процессов с помощью нейросетевых технологий и стохастических дифференциальных уравнений (Lototsky, Rozovsky 2017). Применение знаний в области теории случайных процессов при решении прикладных задач нелинейной динамики сосредоточенных и распределенных систем при учете шумов и флуктуаций успешно используются, например, в финансовой сфере (Zhang, Jianfeng 2017). Между тем, подобный интегрированный подход для исследования речевых реакций, текстового анализа, предиктивной аналитики развития конфликтов не применялся. Исследование позволило проанализировать восприятие пользователей социальных медиа, речевое и коммуникативное поведение, выявить конфликтогенные цифровые зоны, основные причины конкретных городских конфликтов, прогнозировать направление роста агрессии и разработать рекомендации для нейтрализации ситуации.

References

Gonçalves A. 2017. Social Media Analytics Strategy. Apress.

Kullkarni, Vivek, Kern M. L., Stillwell D., Kosinski M., Matz S., Ungar L., Skiena S., Schwartz H. A. 2018. Latent Human Traits in the Language of Social Media: An Open-Vocabulary Approach // PLoS ONE. November 28, 13 (11).

Lototsky S. V., Rozovsky B. L. 2017. Stochastic Partial Differential Equations. Springer International Publishing.

Meiselwitz G. (ed.) 2019. Social Computing and Social Media. Communication and Social Communities. Springer International Publishing.

Shibuya, Yuya 2020. Social Media Communication Data for Recovery. Springer Singapore.

Zhang J. 2017. Backward Stochastic Differential Equations. From Linear to Fully Nonlinear Theory. Springer-Verlag New York.

К ВОПРОСУ ОБ ИЗМЕРЕНИИ КОЛИЧЕСТВА И ЦЕННОСТИ ИНФОРМАЦИИ В ЯЗЫКОВОЙ ЕДИНИЦЕ ВТОРИЧНОЙ НОМИНАЦИИ (НА ПРИМЕРЕ ИДИОМ АНГЛИЙСКОГО ЯЗЫКА)

И. Г. Подгорбунская

mir_63@mail.ru

СибГУ им. академика М. Ф. Решетнева

Язык как самоорганизующаяся система неразрывно связан с теорией информации. Причиной более пристального изучения информации послужил научно-технический прогресс и создание искусственного интеллекта. Теперь информация рассматривается как научное понятие, у которого требуется выявить наиболее характерные свойства. В связи с этим была создана теория информации, которая опиралась на исчисление ее количественной меры (Шеннон, Колмогоров и др.). Ученые пытались решить проблему исчисления ценности и полезности информации с точки зрения ее использования.

В 1960 году А. А. Харкевич предложил определение ценности как «приращение вероятности достижения данной цели в результате использования данной информации» (Харкевич). Были также предприняты попытки определить количество семантической информации с помощью математических средств (Р. Карнап и др.). Использование информационных методов получило широкое распространение как в лингвистике, так и в других науках.

Целью нашего исследования является решение вопроса определения количества и ценности информации в единице вторичной номинации — фразеологической единице с помощью семантической теории информации.

Язык, как открытая система, на пути своего развития постоянно подвергается изменениям. Принципы и характер этих изменений можно объяснить с точки зрения теории информации. В каком-то смысле процесс изменений в неявной форме уже присутствует в наших представлениях. Предполагается, что процесс обмена информацией между коммуникантами должен быть оптимальным, передача и обработка должны происходить в кратчайшее время и с наименьшими затратами для достижения прагматического эффекта.

Информация тесно связана со значением и поэтому без четкого определения значения единицы языка не будет возможности определения количества информации. С одной стороны, в знаково-познавательных и знаково-коммуникативных ситуациях значение выражает свойство информационной инвариантности, а с другой — значение представляет двуединое семантическое отношение, определяющее соответствие знака и объекта, знака и предмета.

В семантической теории информации также важно определение самого понятия «информация». В работах Ю. А. Шрейдера, Е. К. Войшвилло, Д. Харраха используется концепция разнообразия, а также отмечается важность оценки и анализа сообщения от свойств самого приемника сообщений, от тезауруса коммуниканта.

Концепции измерения количества информации и ее ценности неразрывно связаны между собой, что подтверждает положение о взаимосвязи семантических и прагматических свойств информации. Несмотря на разные подходы к измерению ценности информации, они имеют общие черты.

Фразеологическая единица, как языковая единица, сопряженная с родо-исторической памятью, уникальна. В соответствии с известными на сегодняшний день подходами решается вопрос измерения количества и ценности информации, заключенной во фразеологической единице в процессе коммуникации.

Литература

Аничков И. Е. 1997. Идиоматика и семантика // Труды по языкознанию. СПб. С. 188–197.

Апресян В. Ю., Апресян Ю. Д. 1993. Метафора в семантическом представлении эмоций // Вопр. языкознания. № 3. С. 27–35.

Арутюнова Н. Д. 1988. От образа к знаку // Мышление. Когнитивные науки. Искусственный интеллект. М.: ЦС филос. (методол.) семинаров при президиуме АН СССР.

Голицын Г. А., Петров В. М. 2010. Информация и биологические принципы оптимальности: Гармония и алгебра живого. М.: Книжный дом «ЛИБРОКОМ».

Иванова С. А., Суетин А. Г. 2019. Хождения по дебрям информации, или Алгоритмы понимания: Познание в эпоху незнания: самоучитель работы с информацией. М.: ЛЕНАНД (Синергетика от прошлого к будущему, № 93).

Колмогоров А. Н. 1965. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. Т. 1, вып. 1. С. 3–11.

Ньютон И. 1989. Математические начала натуральной философии = Philosophiæ Naturalis Principia Mathematica: [пер. с лат.]; ред. и предисл. Л. С. Полака; пер. и комм. А. Н. Крылова. М.: Наука.

Степанов А. М., Бухтояров А. А., Брагин В. В., Мирской А. И. 2018. Человек как мерная система в развитии современного Знания и технологий // Информационное право, №4 (58). С. 33–42.

Урсул А. Д. Информация: Методологические аспекты. Изд. 2-е. М.: ЛЕНАНД, 2020.

Харкевич А. А. 1960. О ценности информации // Проблемы кибернетики, вып. 4. М.

Шамис А. Л. 2005. Пути моделирования мышления. М.: Комкнига.

Шеннон К. Э. 1963. Работы по теории информации и кибернетике. М.

Шрейдер Ю. А. 1969. О семантических аспектах теории информации // Информация и кибернетика. М.: Наука.

Fodor J. A. 1975. The Language of Thought. N. Y.: Thomas Y. Crowell company.

Harrah D. 1967. Communication: a Logical Model. Cambridge, MA.

Pinker S. 2007. The Stuff of Thought. Language as a Window to the Human Nature. N. Y.: Penguin Group.

ИНТЕРНЕТ-РЕСУРС «МАЛЫЕ ЯЗЫКИ РОССИИ»: СПОСОБЫ ВИЗУАЛИЗАЦИИ ЯЗЫКОВЫХ ДАННЫХ

К. К. Поливанов
polivanov.studio@gmail.com

О. А. Казакевич
kazakevich.olga@gmail.com

Н. В. Сердобольская
serdobolskaya@gmail.com

З. М. Халилова
zaira.khalilova@gmail.com

Е. М. Будянская
budyanskaya.lena@gmail.com

А. П. Евстигнеева
evstigap@gmail.com

К. О. Мищенкова
karinam6@mail.ru

Д. Д. Мордашова
mordashova.d@yandex.ru

С. В. Покровская
sofie.v.pokrovskaya@gmail.com

Е. А. Ренковская
jennyrenk@gmail.com

Институт языкознания РАН

Сайт «Малые языки России» создается на базе Лаборатории исследования и сохранения малых языков Института языкознания Российской академии наук как общедоступный информационный интернет-ресурс, содержащий материалы о структуре и функционировании миноритарных языков России и их локальных вариантов.

Главной задачей данного ресурса является систематизация имеющихся сведений по конкретным языкам. В число этих сведений входят не только уже опубликованные материалы, но и неопубликованные полевые данные, полученные от специалистов по конкретным языкам. При этом на сайте представлены как современные данные, которые планируется регулярно пополнять, так и более ранние материалы, по большей части относящиеся к прошлому столетию.

Статья по каждому языку представляет собой шаблон, содержащий следующие типы информации: генетическая принадлежность, территория распространения, языковые контакты, сферы функционирования языка, динамика развития языковой ситуации, структура языка, сведения об исследователях и исследовательских центрах, список основных публикаций и электронных ресурсов, информация о наличии административной и общественной поддержки. Пользовательский интерфейс строится с использованием шаблонов разметки html5, каскадных таблиц стилей CSS3, а также js и jquery библиотек. В частности, используются библиотеки leaflet для создания интерактивных карт, а также JavaScript библиотека D3 для создания динамических интерактивных визуализаций и схем.

При разработке данного шаблона возник целый ряд сложностей, связанных с визуализацией данных о языке. В докладе будут рассмотрены пять из них.

1. Представление генеалогии

Основные задачи: визуализация полной картины генеалогии (от уровня диалекта до уровня языковой семьи) с возможностью перехода на конкретные фрагменты; удобство использования; адекватное отображение данных.

Решение: были разработано две интерактивные схемы, различающиеся способом визуализации данных. Первая схема содержит более традиционное представление генетической принадлежности в древовидной форме. Имеется возможность сворачивать и разворачивать узлы дерева для фокусировки внимания на его отдельных фрагментах. Вторая схема содержит представление в форме вложенных друг в друга окружностей. Для наглядности различается визуальное оформление разных уровней генетической классификации. Уровни языка и диалекта отличаются от вышестоящих уровней контрастным цветом, расшифровка которого дается в легенде под схемой.

2. Представление территории распространения языка

Основные задачи: визуализация распространения в целом, а также отдельно по населенным пунктам; удобство использования; адекватное отображение данных.

Решение: была разработана интерактивная карта, на которой в соответствующих локациях маркерами отмечены язык и диалекты. Для наглядности каждому диалекту присваивается свой цвет; значения используемых цветов даются в легенде под картой. Для населенных пунктов, где локализовано более одного диалекта, был введен особый маркер, состоящий из нескольких цветов. При клике на маркер во всплывающем окне возникает перечень диалектов.

3. Представление сфер функционирования языка

Основные задачи: визуализация полной картины сфер функционирования в рамках одного пользовательского экрана; удобство и скорость получения информации.

Решение: была разработана интерактивная таблица, в которой представлены основные сферы функционирования языка (был выделен список из 16 таких сфер). Для удобства просмотра и восприятия данных таблица была сделана следующим образом: в ячейках таблицы представлены названия сфер функционирования и отмечено, используется ли язык в соответствующей сфере. При клике на ячейку открывается всплывающее окно с развернутой информацией об особенностях функционирования языка в данной сфере.

4. Представление фонетических данных

Основные задачи: визуализация фонетических и фонологических данных в единой и доступной форме; удобство и скорость получения информации; объективность и репрезентативность.

Решение: ведется разработка графических представлений, в которых одновременно отражаются данные фонетики и фонологии. Звуковой строй каждого языка записывается в системе МФА. При этом структуры представления строятся с учетом особенностей русскоязычной терминологии. Предпринимаются попытки стандартизировать представление с целью выработки оптимального варианта структуры, релевантного для любого языка России. Всё это призвано упростить сопоставление звуковых систем разных языков.

5. Представление динамики развития языковой ситуации

Основные задачи: визуализация общей картины динамики, позволяющая отслеживать изменение во времени различных количественных показателей.

Решение: в настоящее время разрабатываются графики, отображающие изменение величины этнической группы, численности членов группы, считающих этнический язык родным, и численности тех, кто сообщил о владении этим языком. Основой для построения графиков служат данные переписей населения разных лет, данные местных администраций, а также оценки исследователей.

ИНСТРУМЕНТЫ ДЛЯ СОЗДАНИЯ КОРПУСА ЛИТЕРАТУРНЫХ ТЕКСТОВ НА МАЛОРЕСУРСНЫХ ЯЗЫКАХ

В. С. Сазонова
ghosttown.vs1@gmail.com

Е. И. Гавришина
9katenok9@gmail.com

Школа лингвистики, НИУ ВШЭ

Задачи изучения, документации и сохранения малых языков часто сталкиваются с проблемой нехватки ресурсов, так как для реализации большинства из них необходимы оцифрованные и удобные в использовании материалы. В то время как для крупных языков существует множество объемных, функциональных корпусов, адаптированных под различные задачи, для подавляющего большинства малоресурсных языков нет общей базы аннотированных текстов. Сложности связаны как с малым количеством текстов в электронном формате, так и с отсутствием механизмов распознавания и разметки для каждого конкретного языка. Таким образом, для решения подобного рода задач необходим ряд предварительных действий и инструментов: оцифровка текстовых данных, последовательное исправление ошибок, оценивание результатов, морфологическая разметка, выравнивание оригинального текста и перевода (alignment).

Наш проект начался в рамках мастерской «Корпус литературных текстов на языках России» в НИУ ВШЭ под руководством Е. Л. Клячко. Основной задачей было разработать пайплайн на примере эвенкийского языка, для которого у нас имелся корпус размеченных полевых записей в формате .eaf, эвенко-русский словарь в формате .dsl и оцифрованные литературные тексты на эвенкийском и русском языках (параллельно или последовательно). Кроме того, мы использовали морфологический анализатор, разработанный Анной Зуевой для эвенкийского языка (https://github.com/zu-ann/apertium-evn).

В работе мы также опирались на некоторые существующие проекты на смежные темы, например, создание корпуса аннотированных текстов для малых языков Сибири (http://siberian-lang.srcc.msu.ru/ru/textspage), использование методов оптического распознавания для уральских языков (https://github.com/nikopartanen/ocr-tutorial, Partanen, Rießler 2019), создание электронных ресурсов для использования в лингвистических исследованиях обско-угорских языков (http://www.babel.gwi.uni-muenchen.de, Schön, Wisiorek 2017) и создание корпуса текстов на мансийских диалектах (Zhornik, Sizov 2018).

Актуальность данного проекта мы видим в том, что для эвенкийского, как и для многих других малоресурсных языков, нет общедоступного параллельного корпуса литературных текстов. Обеспечить существование корпусов для всех языков довольно сложно, потому что процесс оцифровки и аннотации вручную не эффективен, а инструментов автоматической обработки для малых языков не так много. Поэтому мы считаем, что создание универсальных инструментов значительно помогло бы исследователям при составлении корпусов, а в дальнейшем, возможно, и для сохранения малых языков в целом.

Цель проекта заключается в том, чтобы реализовать разработанный пайплайн создания корпуса для эвенкийского языка и обеспечить его доступность в Интернете. Мы старались делать инструменты наиболее универсальными, чтобы ими можно было пользоваться в других проектах после внесения минимальных правок в изначальный код.

Первым этапом нашей работы стало получение словаря корней и глосс из изначального корпуса эвенкийского языка в формате .eaf с помощью скрипта на Python. Затем полученную информацию было необходимо транслитерировать из латиницы в кириллицу, чтобы в дальнейшем использовать один алфавит для всех инструментов и итогового корпуса литературных текстов. Чтобы добиться высокого качества оптического распознавания, для обучения инструментов необходимо обеспечить как можно большее количество словоформ, которые

могут встретиться в тексте, поэтому был также написан скрипт для генерации словоизменительных парадигм для лемм из словаря, полученного на первом этапе, при помощи анализатора для эвенкийского языка, упомянутого выше. С той же целью был реализован парсинг эвенко-русского словаря в формате .dsl, то есть получен полный список имеющихся лемм и соответствующих переводов на русский язык. Для пополнения итогового корпуса были использованы отсканированные и распознанные с помощью алгоритмов OCR (оптическое распознавание символов) тексты на эвенкийском с переводом на русский, для которых применялось выравнивание и морфологический анализ. На финальном этапе проекта планируется все обработанные тексты с дополнительной морфологической информацией и параллельным переводом на русский язык сделать доступными онлайн.

Таким образом, в **результате** работы были получены специальные инструменты, которые в изначальном виде применимы только для эвенкийского языка, как, например, транслитерация, а также ряд универсальных инструментов, таких как получение необходимой информации из корпуса в формате .eaf и парсинг словаря в формате .dsl, которые могут быть использованы при создании корпусов для других малоресурсных языков.

Литература

Partanen N., Rießler M. 2019. An Ocr System for the Unified Northern Alphabet // Fifth International Workshop on Computational Linguistics for Uralic Languages. Pp. 77–89.

Wisiorek A., Schön Z. 2017. Ob-Ugric database: Corpus and lexicon databases of Khanty and Mansi dialects // Acta Linguistica Academica. V. 64, №. 3. Pp. 383–396.

Zhornik D. O., Sizov F. O. 2018. Linguistic variation and minor languages corpora: a case study of mansi dialects // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции Диалог (Москва, 30 мая — 2 июня 2018 г.). V. 27.

ФОРМИРОВАНИЕ ЭКВИВАЛЕНТНЫХ ОТНОШЕНИЙ МЕЖДУ ЗНАКОМ И РЕФЕРЕНТОМ ЗАВИСИТ ОТ ТИПА ПОСЛЕДНЕГО[1]

М. В. Самулеева
samuleeva@gmail.com
Московский институт психоанализа

А. А. Смирнова
annsmirn1@gmail.com
МГУ имени М.В. Ломоносова

З. А. Зорина
zoyazorina17@gmail.com
МГУ имени М.В. Ломоносова

Использование языка человеком связано со способностью к символизации, т. е. установлению эквивалентных отношений между знаком и референтом (Zentall et al. 2014). Эквивалентные отношения обладают свойствами симметричности, рефлексивности и транзитивности (Sidman et al. 1982). Для изучения механизмов формирования эквивалентных отношений используют задачу выбора по условному соответствию образцу: субъекта обучают выбирать стимул B, если образец A (A-B) и стимул C, если образец B (B-C). Затем в тесте оценивают возможность спонтанного появления новых отношений: например, между образцом B и стимулом A (отношение симметричности; B-A); между образцом A и стимулом A (отношение рефлексивности; A-A); между образцом A и стимулом C (отношение транзитивности; A-C; Sidman et al. 1982; Самулеева и др. 2019). Испытуемые-люди с такими тестами обычно справляются (Quezada Velázquez et al. 2018). Моделирование и изучение этого процесса возможно в экспериментах с животными. С тестом на понимание симметричности отношений животные справлялись лишь в нескольких исследованиях, в которых им ранее демонстрировали возможность таких отношений (Frank, Wasserman 2005). Целью нашего исследования было выяснить, могут ли отношения между знаком и обозначаемым стать симметричными спонтанно, т. е. без такой демонстрации. Исследование проводили на 8 серых воронах (*Corvus cornix L.*) разного возраста: их подкрепляли за выбор изображения двух одинаковых по размеру фигур (кругов, эллипсов, квадратов и т. п.), если образцом было изображение буквы «S», и разных по размеру фигур, если образец — «V». При обучении по очереди использовали 12 стимулов для выбора. Обучение проводили до достижения не менее 80% правильных выборов в 96 пробах подряд ($p<0.0001$). Провели два эксперимента с двумя группами птиц, которые отличались моментом предъявления теста на симметричность. В первом эксперименте выбору по условному соответствию образцу с двумя первыми стимулами для выбора (кругами одинакового или разного размера) удалось обучить двух серых ворон из четырех (потребовалось 2328 и 2248 проб). После этого провели первый тест на понимание симметричности отношений, в котором образцы и стимулы для выбора поменяли местами. Доля правильных решений в 48 тестовых пробах у обеих ворон не отличалась от случайного уровня (43.8% и 54.2%). После этого птиц обучили выбору по условному соответствию образцу с оставшимися 10 стимулами и провели второй тест на понимание симметричности отношений, в котором у одной из птиц доля правильных выборов превышала случайный уровень (64.6%, $p=0.02$; $n=48$). Для того, чтобы выяснить, с чем птицы связали знаки «S» и «V» — с конкретными стимулами или с обобщенными представлениями о сходстве и различии, провели тесты на перенос, в которых использовали новые стимулы для выбора, различающиеся либо по знакомому признаку (размеру), либо по новому признаку (форме). Птицы справились с обоими типами тестов: доля правильных выборов в первом у обеих ворон составила 79.2% ($p=0.0008$; $n=24$); а во втором — 66.7% ($p=0.03195$; $n=24$) и 75.0% ($p=0.0033$; $n=24$). Следовательно, вороны связали знаки «S» и «V» с обобщенными представлениями о сходстве и различии. После этого обе птицы справились с третьим тестом на по-

[1] Работа поддержана грантом РФФИ № 20-015-00287

нимание симметричности отношений (66.7%, p=0.03195, n=24; и 79.2%, p=0.00077, n=24). Положительный результат третьего теста мог быть обусловлен либо тем, что птицы к моменту его проведения связали знаки «S» и «V» с представлениями о сходстве и различии, либо с тем, что в ходе двух предыдущих тестов им продемонстрировали возможность симметричных отношений (хотя во всех тестовых пробах птица получала корм в случае любого выбора). Поэтому во втором эксперименте с новой группой ворон провели только один тест на понимание симметричности отношений: после завершения обучения и тестов на перенос на новые стимулы для выбора. Из четырех ворон выбору по условному соответствию образцу с кругами одинакового или разного размера удалось обучить только одну (потребовалось 3168 проб). После завершения обучения с оставшимися 10 стимулами эта птица справилась как с тестом на перенос правила выбора на новые стимулы, различающиеся по знакомому признаку — размеру (83.3%, p=0.0001, n=24), так и с тестом с новыми стимулами, различающимися по новому признаку — форме (70.8%, n=24; p=0.011). Эта ворона, связавшая знаки «S» и «V» с понятиями сходства и различия, далее справилась с тестом на понимание симметричности отношений (95,8%, p<0.0001, n=24) без какой-либо предыдущей демонстрации возможности таких отношений. Таким образом, мы обнаружили, что если референтом является не отдельный объект, а обобщенное представление о классе объектов (понятие), то это положительно влияет на формирование эквивалентных отношений.

References

Самулеева М. В., Смирнова А. А., Зорина З. А. 2019. Методы исследования механизмов формирования эквивалентных отношений между знаком и обозначаемым у человека и животных // Экспериментальная психология. Т. 12, №. 4. С. 91–105.

Frank A. J., Wasserman E. A. 2005. Associative symmetry in the pigeon after successive matching-to-sample training // Journal of the Experimental Analysis of Behavior, V. 84, №2. Pp. 147–165.

Quezada Velázquez A. G., Padilla Vargas M. A., Flores Aguirre C. J. 2018. Equivalence class formation in 11-month-old pre-linguistic infants // Acta Colombiana de Psicología. V. 21, №1, Pp. 271–289.

Sidman M., Rauzin R., Lazar R., Cunningham S., Tailby W., Carrigan P. 1982. A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children // Journal of the Experimental Analysis of Behavior. V. 37, №1, Pp. 23–44.

Zentall T. R., Wasserman E. A., Urcuioli P. J. 2014. Associative concept learning in animals // Journal of the Experimental Analysis of Behavior. V. 101, №1. Pp. 130–151.

ТРАНСЛИНГВАЛЬНЫЕ ОСОБЕННОСТИ ЗНАКОВ (НА ПРИМЕРЕ РЕКЛАМНЫХ ВЫВЕСОК ЦЕНТРАЛЬНЫХ УЛИЦ ГОРОДА МОСКВЫ)

У. В. Смирнова

yanasmirnova@hotmail.com

Московский городской педагогический университет

В работе предлагается провести анализ транслингвальных особенностей знаков лингвистического ландшафта мегаполиса на примере рекламных вывесок центральных улиц города Москвы. Представляется, что транслингвальные знаки могут стать серьезным препятствием для искусственного интеллекта.

Актуальность исследования транслингвальных знаков обусловлена необходимостью исследования новой модели коммуникации — транскультурации и связанной с ней новой структуры знака, которая в исследованиях характеризуется с помощью понятий «транслингвальность» / «транслингвизм» и «транс-семиотизация». Такой знак представляет собой гибрид разных типов знаков, например (символ и икона), или гибрид символических знаков разных естественных языков.

Транскультурация представляет собой модель, складывающуюся в условиях противодействия: во-первых, гегемонии монокультуры, во-вторых, мультикультурализму. Более того, хотя сама модель несет черты глобализованного культурного пространства, по своим идеологическим принципам она призвана противостоять гиперкультуре (как ризоме разорванных частных я). Ее основные черты описываются с помощью понятий: 1) транскультурная идентичность, не исключающая и не подавляющая возможные множественные идентичности внутри единого целого; 2) культурный полилог, условия к которому создаются наличием «нескольких культурных точек отсчета» и, как следствие, возможностью курсирования между культурами, «сохраняющими право на непрозрачность» (Тлостанова 2011: 133, 138).

Транслингвальные знаки демонстрируют эти абстрактные положения в наиболее яркой форме. Необходимо подчеркнуть, что транслингвизм не является переключением или смешением кодов, но представляет собой концепцию, исследующую язык как когнитивную способность, подразумевающую наличие единого коммуникативного репертуара языковых средств и рассматриваемую в контексте холистических воззрений на язык (Canagarajah 2018; Garcia, Kleifgen 2018; Lee 2017; Li 2017; Wagner 2018).

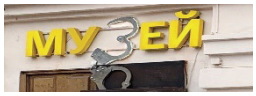
Транслингвизм реализует свойство транскультуральной модели коммуникации адаптироваться к сложной глобальной и локальной социолингвистической ситуации, осуществляя социолингвистическое конструирование языка и собственной идентичности. Выход языка за границы, такое его свойство, как текучесть, приводят к возникновению сложных семиотических форм, проявляющих деятельностный и речетворческий характер языка.

Транслингвизм появляется там, где есть транслингвальная ситуация. Рекламная вывеска становится проекцией траснлингвальных смыслом и соединением нескольких систем координат. Для анализа в работе были выбраны знаки рекламных вывесок центральных улиц города Москвы — Старый Арбат, Новый Арбат, Тверская. Методы исследования включали метод описания лингвистического ландшафта, семиотический анализ. Аналитическими категориями классификации знаков были: географическая дистрибуция, содержащиеся языки, порядок следования знаков, закономерности комбинации.

Общее количество проанализированных вывесок — 105, из них 14 были признаны транслингвальными. С точки зрения географической дистрибуции более всего транслингвальных знаков было обнаружено на улице Новый Арбат. С точки зрения представленности

естественных языков следует отметить, что были обнаружены знаки — гибриды русского и английского. Порядок следования знаков включал расположение знака-иконы внутри символьного знака и гибрид символов английского и русского языков в разной последовательности.

Особый интерес представляют закономерности создания знаков на примере трех групп знаков: гибриды «символ-икона», гибриды «символ культуры-символ», «символ-символ».



Конструирование знака «икона-символ» (М❤СКВА, К☕ФЕ, МУ🖇ЗЕЙ) следует логике введения знака-иконы внутрь границ целостного знака символа. Знак-икона встроен в знак-символ на основании игрового сходства графического изображения и символьного знака, что подчеркивает творческий характер транслингвального знака. Особенностью данного знака становится тот факт, что знак-икона разворачивает в границах знака или предикацию (пить кофе, любить Москву), или номинацию («история телесных наказаний»).

Знаки гибриды «знак культуры-символ» на первый взгляд не входят в число транслингвальных знаков, т.к. представляют собой транслитерированный перевод некой российской культурной реалии — Matreshka, Skazka, Kuzina. Однако их можно признать гибридами культурного артефакта одной лингвокультуры и графической формы, принадлежащей другой лингвокультуре. Весьма интересна в этих знаках взаимонаправленность тенденций: с одной стороны, осуществляется проекция русской лингвокультуры в глобализированное пространство, в котором английский язык приобрел статус лингва-франка, с другой стороны, происходит обратная проекция глобализированной англо-американской культуры в русскую идентичность. Наблюдается остранение знака культуры, превращение его в чужой или кажущийся чужим.

Знаки «символ-символ» #FARШ, Eat Market фуд-холл, the Дача выявляют разный порядок следования символов естественных языков, но в целом демонстрируют динамичный характер границ нового конструируемого знака.

Литература

Тлостанова М. В. 2011. Транскультурация как модель социокультурной динамики и проблема множественной идентификации // Вопросы социальной теории. Т. 5, с. 126–149.

Canagarajah S. 2018. Translingual practice as spatial repertoires: expanding the paradigm beyond structuralist orientations // Applied linguistics. Vol 39 (1), pp. 31–54.

Garcia O., Kleifgen J. A. 2018. Educating emergent bilinguals. Columbia University: Teachers College Press.

Lee J. W. 2017. The politics of translingualism: After Englishes. London and New York: Routledge.

Li W. 2016. Multi-competence and the Translanguaging Instinct // V. Cook (ed.). The Cambridge Handbook of Linguistic Multi-Competence. Cambridge: Cambridge University Press, pp. 533–543.

РАСПРЕДЕЛЕНИЕ КВАЗИСИНОНИМОВ В ТЕЗАУРУСЕ ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ[1]

В. Д. Соловьев
maki.solovyev@mail.ru
Казанский федеральный университет

Н. В. Лукашевич
louk_nat@mail.ru
Московский государственный университет

Современные электронные тезаурусы WordNet (Fellbaum 1998) и его варианты для других языков, в том числе тезаурус русского языка RuWordNet (Loukachevitch et al. 2018), являются одними из самых востребованных компьютерных ресурсов в задачах автоматической обработки текстов. Такие ресурсы строятся на основе синсетов — синонимических рядов, связанных семантическими отношениями гипо-гиперонимии, антонимии и некоторых других. Для практического применения тезаурусы должны включать десятки тысяч слов и выражений, что требует значительных ресурсов по их развитию и поддержанию, включая тестирование качества их формализованных описаний.

В данной работе мы предлагаем подход к тестированию существующих тезаурусов на основе слов-квазисинонимов. Квазисинонимы — это относительно малоизученное понятие, обозначаемое разными словами, например, в словаре Апресян (ред.) 2004 для обозначения этого уровня семантической близости используется слово ‘аналог’. Тем не менее это весьма важное понятие, используемое в ряде задач обработки текстов.

Представляется интересным проследить, как квазисинонимы располагаются в тезаурусах. В данной работе мы рассматриваем тезаурус RuWordNet, а в качестве квазисинонимов берем пары слов-аналогов по словарю Апресян (ред.) 2004. Нулевая гипотеза состоит в том, что аналоги должны располагаться в тезаурусе достаточно близко друг к другу. Степень близости трактуется как расстояние между словами в графе семантических связей тезауруса. Согласно работе Loukachevitch 2019, близкими уместно считать слова, располагающиеся на расстоянии 4 и меньше шагов по ребрам графа.

При обнаружении пар слов-аналогов, находящихся на бóльших расстояниях, можно предположить некий пробел в тезаурусе. Таким образом, развиваемый в данной работе подход, направлен на решение практической задачи повышения качества тезаурусов — выявление возможных пропусков слов, их значений или семантических отношений между ними. В более широком смысле таким способом определяется уровень когерентности двух языковых ресурсов — тезауруса и словаря.

В данной работе мы рассматриваем аналоги в двух частях речи — глаголах и прилагательных. Нами составлен список пар <слово, его аналог> по словарю Апресян (ред.) 2004, доступный по адресу https://kpfu.ru/kompleksnyj-analiz-struktury-i-soderzhaniya-366287.html. Список содержит 1601 пару глаголов и 685 пар прилагательных. Для расчета расстояния и генерации всех возможных путей между словами в RuWordNet применен стандартный алгоритмом ‘поиск в ширину’ — один из методов обхода графов. В общей сложности получено 138505 путей между парами глаголов-аналогов и 10837 путей — для прилагательных. Перечень путей доступен по тому же адресу.

Распределение пар аналогов по расстояниям приведено в таблице 1.

Как видно из таблицы, для прилагательных более 90% пар аналогов находятся на близком расстоянии, а для глаголов даже почти 99%. Это указывает на хорошую согласованность описаний слов в тезаурусе RuWordNet и пар аналогов в словаре Апресян (ред.) 2004. Сформулированная нулевая гипотеза в целом подтверждается. Тем не менее есть и определен-

[1] Исследование выполнено при поддержке РФФИ, гранты № 18-00-01238, 18-00-01226 в составе комплексного проекта № 18-00-01240 (К).

ное число расхождений, которые были подвергнуты тщательному анализу с привлечением данных различных словарных и корпусных источников.

Таблица 1. Число пар аналогов в зависимости от расстояния

Расстояние	Число пар прилагательных	Процент	Число пар глаголов	Процент
1	171	30,65	392	27,80
2	150	26,88	447	31,70
3	122	21,86	386	27,38
4	63	11,29	170	12,06
5	45	8,07	13	0,92
6	7	1,25	2	0,14

Для примера рассмотрим пару слов *коварный — каверзный*. В RuWordNet кратчайшие пути между ними имеют длину 6, причем промежуточные вершины на этих путях семантически далеко отстоят от рассматриваемых слов. Одна из причин этого состоит в том, что не существует удовлетворительных универсальных классификаций прилагательных. Вероятно, переход от прилагательных к существительным является хорошим подходом к их классификации и, соответственно, установления коротких путей.

Рассмотрим слова *коварство* и *каверза*. Первое из этих слов указано в RuWordNet в качестве однокоренного слова и также в качестве частеречного синонима к *коварный*. Второе из этих слов — *каверза* — не связано с *каверзный* отношениями этого типа. Предлагается указать *каверзу* как частеречный синоним к *каверзный.*

Для слова *каверза* в RuWordNet указано значение ‘шалить’ с гиперонимом ‘баловство’. Рассмотрим значение слова *каверза* в словарях. Викисловарь дает такие значения: *злая интрига, происки, затеваемые с целью запутать что-либо*, *повредить кому-либо*; *коварная, злая шалость*, *проделка*. Аналогичные значения и в словаре Ушакова. В корпусах легко найти такие контексты: *Он чувствовал, что готовится какая-то каверза со стороны союзников* (А. Толстой. Похождения Невзорова и Ибикус).

Как видим, слово *каверза* имеет определенно негативный смысл (а не только ‘баловство’) и явно сближается со словом *коварство*. Предлагается добавить в RuWordNet второе значение слова *каверза*: *каверза 2* (злая интрига) с гиперонимом *дурной поступок*. В этом случае длина пути между исходными словами сокращается до 4: *коварный — коварство — дурной поступок — каверза 2 — каверзный*. Первое и четверное отношения — частеречная синонимия, второе и третье — гипо-гиперонимическая связь, причем первое присутствует в RuWordNet. Для выстраивания короткого пути потребовалось добавить одно новое значение слова и два отношения. Следует отметить, что это не единственная возможность установить короткий путь между рассматриваемыми словами. Вообще наличие нескольких путей одинаковой длины является нормой.

Анализ пар слов-аналогов с большим расстоянием между ними позволил выявить следующие типичные пробелы в тезаурусе:

1) пропуск слов, находящихся в отношении частеречной синонимии (*каверзный — каверза*);
2) пропуск значений слов, присутствующих в стандартных словарях русского языка (*громкий* — ‘получивший широкую известность’);
3) пропуск отношений между однокоренными словами (*злодейство — злой*).

Учет перечисленных пробелов позволит повысить качество тезауруса RuWordNet. Ранее подобный подход к верификации и усовершенствованию тезаурусов, насколько нам известно, не применялся.

Литература

Апресян Ю. Д. (ред.) 2004. Новый объяснительный словарь синонимов русского языка. Москва; Вена: Языки славянской культуры: Венский славистический альманах.

Fellbaum Ch. 1998. WordNet: An electronic lexical database. MIT Press.

Loukachevitch N. 2019. Corpus-based Check-up for Thesaurus // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics. Pp. 5773–5779.

Loukachevitch N., Lashevich G., Dobrov B. 2018. Comparing Two Thesaurus Representations for Russian // Proceedings of Global WordNet Conference GWC. Pp. 35–44.

О СПОСОБЕ ИСЧИСЛЕНИЯ ПАРАМЕТРОВ В ОБРАЗНОМ ПРЕДСТАВЛЕНИИ

А. И. Тимофеев

timofiev@yandex.ru

ОАО Национальный институт авиационных технологий

Научный Совет по робототехнике и мехатронике РАН

Научный Совет Business System Laboratory

Отсутствие кардинального решения актуальной проблемы в манипуляционной робототехнике и в протезостроении — обеспечение надежности захвата неориентированных объектов сложных форм — связано с отсутствием активных сил — проекций силы веса — в новых точках контакта до отрыва объекта от исходного положения и неизбежным появлением их в процессе манипулирования в соответствии с физическими законами нашей планеты.

Это генерирует высшую степень неопределенности физической ситуации в системе «Кисть-Объект» (К-О), связанную с информационным разрывом на семантическом уровне Настоящего времени, когда необходимо принимать решение в исходных условиях, и Будущим временем, где реализация этого решения будет происходить в других условиях.

Проект «Искусственная “разумная” рука», базирующийся на моделировании как способов организации, так и функциональных принципов двигательного акта руки человека с применением пространственного тактильного осязания, в т.ч. архитектоники функциональной системы человека — системы захвата, призван обеспечить кардинальное решение упомянутой проблемы с увеличением границ радиуса сферы активного манипулирования в пространстве на основе возможностей оценки физической ситуации в Кисти, прогнозирования надежности захвата с принятием решения в образном представлении — результат моделирования как афферентных, так и эфферентных биологических связей и их синтеза. Это принципиально отличается от разработанных зарубежных образцов современных биопротезов и манипуляторов, лишенных, например, афферентных (обратных) связей с указанными возможностями.

Одно из особенностей формализованных описаний междисциплинарных знаний и принятого решения проблемы — применение внутреннего невербального образного языка — языка инженерии знаний в качестве способа исчисления параметров образов при восприятии, хранении, обработке и передаче информации, где:

— кодирование языка является внутренней связью семиотической структуры отношений точек контакта в качестве признака закрепощений виртуальных перемещений Объекта,
— синтаксис языка содержит правила построения любого образа и подразделяется на порождающие (с целью обеспечения процедуры принятия решения) и интерпретирующие, выводящие на семантику,
— морфология языка определяет как структуру, так и причинно-следственные связи содержания образов с другими образами единого информационного пространства системы К-О.

Таким образом упомянутый внутренний образный язык является средством познания физической ситуации в системе К-О на семантическом уровне.

В целом упомянутое решение актуальной проблемы представляет собой дедуктивную логическую систему с применением псевдофизической логики, использующей пространственные каузальные отношения точек контакта в информационно управляемых процессах прогнозирования надежности захвата в условиях неопределенности. В результате упомянутый способ исчисления параметров обеспечивает семантическую выразительность как процедуры принятия решения, так и параметров самого решения с такими качествами информации, как когнитивная относительность, ясность и сжатие формы.

Литература

Колмогоров А. Н. 1964. Жизнь и мышление как особые формы существования материи. О сущности жизни. М., Наука.

Тимофеев А. И. 2007. Об информационных процессах самоорганизации (на примере технической модели системы захвата) // Вестник Международной Академии наук (Русская секция), под. ред К. В. Судакова, №1. С. 57–62.

Timofeev A. I. 2015. Technical Thinking systems are on the threshold of our Society // Horizon Research Publishing Corporation, Universal Journal of Management, vol. 3(9), USA. Pp. 372–375.

Timofeev A. 2015. Civilization and Technical Thinking systems. LAMBERT Academic Publishing.

НЕКАНОНИЧЕСКИЕ ЭЛЕМЕНТАРНЫЕ ДИСКУРСИВНЫЕ ЕДИНИЦЫ (ЭДЕ) В ТЕКСТАХ БЛОГОВ (НА МАТЕРИАЛЕ РАЗМЕТКИ Ru-RSTreebank)[1]

С. Ю. Толдова
toldova@yandex.ru
НИУ ВШЭ

Т. В. Давыдова
tdadidik@gmail.com
ФИЦ ИУ РАН

М. В. Кобозева
kobozeva@isa.ru
ФИЦ ИУ РАН

Д. Б. Писаревская
dinabpr@gmail.com
ФИЦ ИУ РАН

Введение

Согласно многим исследованиям, в текстах блогов проявляются признаки устной речи (см., например, Scheffler et al. 2019: 52). Эти признаки проявляются не только на лексическом уровне. Устная речь оказывает влияние и на дискурсивную структуру блогов. Так, в блогах существенно больше различных типов непредикативных (в том числе неполных) элементарных дискурсивных единиц (ЭДЕ) по сравнению с письменной речью. Рассмотрению роли неканонических ЭДЕ и их конфигураций в блогах посвящена настоящая работа. В качестве материала используется подкорпус блогов (104 текста) из русскоязычного корпуса Ru-RSTreebank (https://rstreebank.ru/dataset) с разметкой в терминах риторических отношений (Mann, Thompson 1988).

Неканонические ЭДЕ (ЭДЕ, не являющиеся финитными клаузами), описанные в Кибрик, Подлесская 2000, соотносятся со стилевыми приемами, имитирующими разговорный жанр, в текстах блогов. Наш анализ показывает, что авторы блогов намеренно используют такую имитацию, «манипулируя» неканоническими ЭДЕ (именными предикатами, парцелляцией, вынесением топика и т. п.), для управлением вниманием читателя и для повышения экспрессивности текста. В работе предлагается типизация таких приемов с точки зрения механизмов фокусирования внимания пишущего-читающего.

Для выделения ЭДЕ мы используем критерии, основанные на пунктуации: точка является маркером границы ЭДЕ. Иногда одну и ту же функцию имеют как неполные ЭДЕ, рассматриваемые в Кибрик, Подлесская 2000, так и синтаксически полные ЭДЕ. Мы используем для таких случаев префикс ‘квази-’.

Квази-ЭДЕ как средство фокусирования и эмфазы (Тип 1)

1. Квази-эхо

Достаточно распространенным приемом является повтор ЭДЕ с ментальными глаголами, глаголами речи, восприятия и эмоций. Такое повторение по структуре напоминает разговорное эхо (Кибрик, Подлесская 2000). Говорящий в новой ЭДЕ переформулирует, уточняет предыдущую. Новая формулировка не является конкретизацией факта или события. Такие конструкции не содержат фактической информации, они предваряют значимую часть фабулы, которая следует за повторами, служат средством привлечения внимания к ней и эмфазы (1).

(1) [*Ну, такая вот картина рисуется, если почитать новости*.] [*Я буду рад ошибаться, но пока выглядит так*.]

В контекстах наречных квази-эхо конструкций парцеллят служит средством интенсификации.

(2) [*Тестировщик должен находить ровно столько багов, сколько починят*.] [*Не меньше*.] [*Но и не больше*.]

[1] [1] Исследование выполнено при поддержке РФФИ, грант № 17-29-07033.

2. Проспективные субклаузальные ЭДЕ

Именные ЭДЕ в блогах используются для выноса темы:

(3) [***Большая чаша с благовониями.***] [*Купила для себя маленькую в Бурятии, теперь жгу дома*.]

Полную ЭДЕ могут предварять ЭДЕ, состоящие из маркеров оценки и привлечения внимания:

(4) [Кстати.] [На пражских домах всегда два номера — на красной табличке и на синей.]

3. Ретроспективные субклаузальные ЭДЕ

Авторы блогов используют приращения и парцелляции:

(5) [*Вокруг шастают гаишники туда-сюда, туда-сюда*.] [*С бумажками и без*.]

Многоступенчатая парцелляция имитирует разговорный способ подачи информации «последовательное нанизывание смыслов».

(6) *Предупреждал вас козлов Уильям наш Шекспир, вот прямо вопиил*, …: [*«Не пей вина, Гертруда!»*] [*Козленочком станешь*.] [*А то и целым козлом*.]

Тема может интродуктироваться с помощью ретроспективной субклаузальной ЭДЕ (катафорическая ЭДЕ).

(7) [*А вот и моя цель*.] [***Книжный супермаркет***.]

Квази-ЭДЕ как средство дефокусирования повествования (Тип2)

Серия так называемых назывных предложений, связанных отношениями Конъюнкция или Последовательность, может использоваться для дефокусирования [Ирисханова 2014]: центр внимания смещается с фактов на эмоциональное состояние автора (иногда для связи ЭДЕ используется дефис):

(8) [*Монголия!*] [*Юрты,*] [*юрты,*] [*небоскрёб,*] [*юрты!*]

(9) [*Отель*]-[*чемодан*]-[*метро.*] [*Тайм-чек.*]

Во вступительной части текста встречаются серии синтаксически разноструктурных ЭДЕ, связанных многоядерными отношениями, которые в сжатом виде задают исходную ситуацию: назывные предикации «выстраиваются» в один ряд (а иногда сочиняются) с финитными.

(10) [*Погоду вы сами знаете (– 9–13 С),*] [*а я по дурости не очень тепло оделся,*] [...]. [*Майка,*] [*кофта,*] [*полуосенняя куртка,*] [*толстые джинсы без ничего больше*], [*ну хоть ботинки хорошие*]. [*1. Киевский вокзал, 8.40 утра.*]

Выводы

Неканонические ЭДЕ, активно используемые в блогах, наряду с другими средствами выразительности текста, маркируют значимые для пишущего фрагменты истории, создавая либо эффект нагнетания саспенса, либо выводя на передний план повествования топик или эмоциональную оценку автором значимой части фабулы (Тип 1). Последовательности ЭДЕ, включающие цепочки назывных предложений (Тип 2), маркируют дефокусирование, смещение внимания с центрального события на оценку автора либо на задание исходной сцены.

Литература

Ирисханова О. К. 2014. Игры фокуса в языке. Семантика, синтаксис и прагматика дефокусирования. М.: Языки славянской культуры.

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

Mann W., Thompson S. 1988. Rethorical Structure Theory: Toward a functional theory of text organization // Text — Interdisciplinary Journal for the Study of Discourse 8(3), January, pp. 243–281.

Scheffler T., Aktaş B., Das D., Stede M. 2019. Annotating Shallow Discourse Relations in Twitter Conversations // Proceedings of the Workshop on Discourse Relation Parsing and Treebanking, pp. 50–55.

ЖАНРОВЫЕ И МОДУСНЫЕ ОСОБЕННОСТИ ЭЛИЦИТИРОВАННЫХ ДИСКУРСОВ[1]

М. Н. Усачёва

mashastroeva@gmail.com

Институт языкознания РАН; ГИРЯ им. А. С. Пушкина

В последнее время полевые лингвисты все чаще проводят эксперименты с целью собрать корпус текстов, в котором частота исследуемого языкового явления намного выше, чем в корпусе спонтанных текстов. Этот метод, который О. В. Фёдорова называет элицитацией дискурсов, отличается от собственно экспериментальных методов «отсутствием требования жесткой контролируемости экспериментальных условий, а также отсутствием необходимости установления причинно-следственных связей между явлениями» (Фёдорова 2016: 8–9). Отличие текстов, записанных в ходе таких экспериментов, от спонтанных состоит не только в том, что говорящих подталкивают к использованию определенных языковых единиц, но и в структуре дискурса, в том числе жанровой. В докладе мы остановимся на некоторых жанровых особенностях элицитированных дискурсов, записанных от носителей коми-язьвинского и бесермянского наречий, а также сделаем выводы о том, как эти особенности можно использовать при планировании экспериментов в поле.

Изучение жанровых особенностей элицитированных дискурсов важно для полевых лингвистов, которые проводят эксперименты. В текстах разных жанров и модусов преобладают разные лексические, фразеологические и грамматические единицы (Бахтин 1996: 159). Например, в нарративе лексика, выполняющая дейктические функции, либо отсутствует, либо интерпретируется иначе, чем в устной речи (Падучева 1996). Известно, что в нарративах и спонтанных диалогах могут использоваться разные видовременные формы: французское Passé Simple и претериты в языках банту употребляются в нарративах, а перфект во многих западноевропейских языках и русский общефактический несовершенный вид — в диалогах. Что касается синтаксиса, из работы (Коваленко 2006) можно сделать вывод, что эллипсис гораздо более характерен для устной речи, чем для письменной — в частности, для письменных нарративов. Жанр и модус текстов задаются конкретной экспериментальной методикой. Знание о том, какие конструкции частотны в текстах данного жанра и модуса, поможет выбрать методику, стимулирующую испытуемых использовать нужные исследователю языковые единицы.

В таблице ниже представлены данные о наличии определенных лексических элементов, грамматических форм и синтаксических конструкций в двух группах текстов — нарративах и квазиспонтанных диалогах. Нарративы представляют собой записи онлайн-комментирования фрагментов мультфильмов «Петр и Петруша», «Злыдни», «Теремок» и «Три медведя» и оффлайн-пересказов бесермянских переводов текстов сказок «Звери в повозке», «Маковое зернышко», «Девочка и медведь», а также изложения фрагмента сюжета «Эпоса о Гильгамеше». В ходе экспериментов по онлайн-методике сначала демонстрировался ролик целиком, затем экспериментатор последовательно показывал ролик короткими отрывками, а испытуемые по ходу просмотра подробно описывали их содержание. Оффлайн-методика состояла в том, что испытуемые прочитывали напечатанный бесермянский текст и пересказывали всё, что запомнили. Различие «онлайн vs оффлайн» является базовым для методов, ориентированных на получение нарративов (Фёдорова 2016); мы считаем, что полученные в ходе соответствующих экспериментов тексты имеют несколько разные жанровые особенности. Квазиспонтанные диалоги были записаны в ходе экспериментов по методике референ-

[1] Исследование проводится при поддержке гранта РНФ №18-18-00462 «Коммуникативно-синтаксический интерфейс: типология и грамматика».

циальной коммуникации (РК). В ходе экспериментов один из испытуемых объяснял другому, какие манипуляции тот должен проделать с находящимися перед ним предметами. При этом испытуемые слышали, но не видели друг друга.

Доля дейктических элементов в таблице представляет собой отношение количества слов с дейктической семантикой к общему числу словоупотреблений в подкорпусе. Процент форм второго прошедшего времени, которое в бесермянском наречии имеет эвиденциальное значение, посчитан как отношение количества этих форм к общему количеству глагольных форм в соответствующем подкорпусе. Для эллиптических конструкций с нанизыванием падежей приведены абсолютные цифры, поскольку данное синтаксическое явление является низкочастотным.

Из таблицы видно, что доля дейктических элементов в квазиспонтанных РК-диалогах почти в 2 раза выше, чем в онлайн-нарративах. Можно сделать вывод, что при исследовании дейксиса лучше использовать методики, ориентированные на получение диалогов. Доля эвиденциального прошедшего времени не так высока в онлайн-комментировании, как в оффлайн-пересказах. Следовательно, для исследования эвиденциальных форм лучше выбрать оффлайн-методику. (Тот факт, что доля искомых форм в РК-диалогах практически такая же, как в спонтанных текстах, объясняется тем, что формы эвиденциального прошедшего не были целевыми для соответствующих экспериментов). При исследовании эллипсиса мы рекомендуем выбирать методики, ориентированные на получение диалогов, а не нарративов-монологов: как видно из таблицы, редкая стратегия оформления конструкций с эллипсисом вершины ИГ (нанизывание падежей) в нарративах не встретилась ни разу.

	оффлайн-нарративы	онлайн-нарративы	диалоги (РК)	спонтанные тексты
лексика: дейктические элементы	—	4% (40 из 1000)	7,5 % (333 из 4434)	—
грамматика: эвиденциальное прошедшее время	10% (258 из 2550)	4% (101 из 2383)	1,5% (40 из 2650)	1,7% (256 из 15372)
синтаксис: нанизывание падежей при эллипсисе вершины ИГ	0 примеров	0 примеров	71 пример	3 примера

Литература

Бахтин М. М. 1996. Собрание сочинений. Т. 5: Работы 1940–1960 гг. М.: Русские словари.

Коваленко Е. В. 2006. Эллиптические структуры как элемент художественного текста (на материале англоязычной оригинальной и переводной литературы). Дисс. … к.ф.н. СПб. Рукопись.

Падучева Е. В. 1996. Семантические исследования. Семантика времени и вида в русском языке. Семантика нарратива. М.: Языки славянской культуры.

Федорова О. В. 2016. Психолингвистические исследования дискурса в полевой лингвистике. Социо- и психологические исследования, вып. 4. С. 7–18.

ЛИНГВОКОГНИТИВНЫЙ МЕТОД АЛГОРИТМИЗАЦИИ ПРОЦЕССОВ ЛОГИЧЕСКОГО МЫШЛЕНИЯ

О. Л. Филипеня
filipock1@yandex.by

В. В. Ткаченко
tkachenko_v@tut.by

Объединенный институт проблем информатики Национальной академии наук Беларуси

В машинном переводе текстов с одних языков на другие, организации речевого интерфейса, ведении электронных баз данных остро проявляется необходимость лингвистической поддержки решения задач, относящихся к технологиям «искусственного интеллекта» и объединяемых проблематикой когнитивной лингвистики (Белоногов 2004, Попова 2007).

В настоящее время отсутствуют действующие решения для моделирования алгоритмическими методами психических процессов. В самом психическом процессе возникает нечто доступное органам чувств — непосредственно воспринимаемое здесь и сейчас с модуляцией текущего восприятия чем-то ранее закрепленным в памяти. Этот процесс характеризуется переходом от восприятия, представления к понятию о предмете или явлении как форме мышления (Большая советская энциклопедия).

Идея, предлагаемая к дискуссии, основывается на предположении возможности создания особого словаря, воплощающего толкование слова вне контекстной зависимости. Например, для слова «ключ», таким токованием может быть выражение «средство доступа». Средство доступа к механизму замка, к тональности партитуры, к подземному источнику воды, к символьному наполнению знаков шифрованного сообщения. В отличие от подхода, который основан на разделении семантического пространства языка (значений единиц конкретного языка) и концептосферы (Попова 2007), то есть упорядоченной совокупности единиц мышления народа, и который используется при «когнитивной интерпретации» результатов лингвистического описания значений лексических единиц с применением статистического метода построения ассоциативных полей концептов, предлагаемый детерминистский подход отвечает правильности приоритетов по Белоногову в соотношении декларативной и процедурной компонент систем автоматической обработки информации (Белоногов 2004). В дополнение необходимо указать, что этот подход может рассматриваться как альтернатива способу «исчисления языковых смыслов», предложенному В. В. Мартыновым, и развиваемому на его основе универсальному семантическому коду (УСК) как средству снятия неопределеннозначности естественного языка на множестве его глагольных форм (Boyko 2017).

В предлагаемом подходе понятие о чем-либо рассматривается как элемент сознания — отображенное в мышлении что-либо. Отображение понятий вне психического процесса осуществляется записью с помощью знаковой системы языка после перевода описываемого образа в символьную форму — внутреннюю речь. Не являясь тождеством, символьная форма для отождествления себя с каким-либо понятием требует своего употребления в одном и том же смысле (Философский словарь / Под ред. И. Т. Фролова. — с. 371).

В языке понятия преобразуются не только в термины, но и слова с множественностью смыслов. Проблема многозначности толкования слов, не являющихся терминами, является одной из проблем точного перевода. По нашему убеждению, множественность толкования связана с отсутствием полноты описания, а отсутствие полноты — с абстрактностью понятийного моделирования, что следует из следующих рассуждений. Воспринимаемое явление анализируется на множестве рецепторов органов чувств с последующим понятийным синтезом на нейронной сети. Неполный синтез абстрактен, то есть не соответствует критерию системной целостности, но слово, которое соответствует абстрактному понятию, осмысливается как символ и лингвистически используется как целое. Конфликт построения целого разрешается при абстрактном восприятии за счет возможности ассоциативной памяти дополнить об-

раз до полного, а неполнота смыслового содержания слова устраняется установлением контекстных зависимостей. При этих операциях становятся очевидными проблема адекватного смыслового наполнения — полноты и целостности представления, и проблема корректного логического наполнения фразы словами. Особенностью завершающей стадии семантической интерпретации является ее прохождение вне рамок логики на индивидуальной структуре мозга, определяющей субъективизм восприятия. Так толкование посредством языка сопряжено с понятийным осмыслением посредством приобретенного опыта и носит субъективный характер.

Осмысление характеризует переход к иному базису — от символа к понятийному моделированию, от знака как материального воплощения символа к его представлению как понятия в сознании. Речь идет об интерпретации — сопряжении разных качеств в одном процессе. Именно интерпретация, в этом смысле, приобретает особую значимость вместе с приобретением статуса актуальности проблемного направления, связанного с логикой осмысления текста средствами программного обеспечения электронно-вычислительных машин. В отличие от математического базиса, реализующего в философском аспекте количественное наполнение взаимодействующих качеств, интерпретатор психических функций осуществляет преобразование разных качеств — сигнатуры возбужденных синапсов в поток ощущений, доступных сознанию как процессу, фиксирующему этот поток, и в поток представлений как процесс удержания в сознании ранее зафиксированное памятью. Понятно, что простая сеть, осуществляющая количественные преобразования вне контекста взаимодействия качеств, нежизнеспособна в плане реализации функции понимания.

Проблема толкования семантической многозначности разрешима более простыми средствами, доступными в философии логики. Для ее решения анализируется отношение между термином и его отражением в сознании — понятием. Как показано ранее (Филипеня 2017), повышение абстрактности при мысленной замене некорректно примененного термина термином с более обобщающим значением, исправляет ситуацию с установлением адекватности выражения.

Вторым приемом преобразования неадекватных с позиции приобретенного опыта выражений в допустимые является подмена по аналогии отдельных элементов в понятийной модели и формирование нового выражения с корректной конкретизацией. В докладе рассматривается сопряжение отдельных явлений по аналогичным признакам, которое обосновывает логичность операций толкования при исчислении смысла.

Выводы. Для эффективной обработки текста средствами электроники необходимо устранить многозначность смысловой обработки на этапе составления промежуточного словаря. При таком подходе открывается возможность записи словаря как граф-схемы с абстрагированием в направлении вершины и конкретизацией в направлении к основанию. В этом случае результат операций анализа и синтеза формально сопоставляется с движением по сети графа и допускает алгоритмизацию логики мышления.

Литература

Белоногов Г. Г. 2004. Компьютерная лингвистика и перспективные информационные технологии. М.: Русский мир.

Попова З. Д., Стернин И. А. 2010. Когнитивная лингвистика. М.: АСТ, Восток-Запад.

Филипеня О. Л., Ткаченко В. В. 2017. Систематизация по выделенному признаку в контексте влияния на эффективность вовлечения терминов в процесс мышления // Материалы VII международной междисциплинарной конференции Когнитивные Штудии. Вып. 7, Минск: БГПУ. С. 182–187.

Boyko I. M. 2017. Semantic Coding for Semantic Knowledge Inference // Open Semantic Technologies for Intelligent Systems Conference (OSTIS-2017), Minsk, Belarus, pp. 158–162.

О ВИДАХ СМЫСЛОВ И ПРИМЕНЕНИИ ИХ В ИСКУССТВЕННОМ ИНТЕЛЛЕКТЕ

А. Б. Хомяков

alexander.xom@gmail.com

Компания ЭТМ, департамент BigData&AI

В докладе предполагается привести анализ когнитивных процессов, которые мы интерпретируем как смыслы высказываний. И приводятся доводы, которые раскрывают причины и способы возникновения смыслов, например, в результате рефлексии или метафорического переноса. В статье также дается анализ того, какой вид смысла доступен для реализации в искусственном интеллекте в ближайшем будущем, чтобы «машины» стали понимать смысл.

Тезисы

1. В ИИ есть проекты, успешно генерирующие связные статьи на заданную тему в NYT (GPT-2 OpenAI), могут подражать стилю великих писателей, даже «сочинять» стихи. Но осмысленными их до сих пор назвать невозможно. Все нейросети страдают синдромом Уильямса. Причины этого стоят отдельного анализа, но часть причин будет приведена в докладе как ответ на вопрос «Почему нейросети не могут создавать тексты со смыслом?».
2. Смысл высказывания не так однороден, как может показаться. Смысл иногда вполне известен нам так, например, как референт книги или анонс фильма. Иногда же смысл — это цель, которая прослеживается в рассуждениях, которые мы и анализируем. Другие варианты, которые мы называем смыслами, — это предназначение, наше отношение к изложенному, влияние на нас цели автора, подразумеваемое «между строк», но не передаваемое прямо в тексте и т. п. В докладе смысл будет рассмотрен с когнитивной точки зрения, что позволяет увидеть различные виды смыслов и определить их признаки.
3. **Виды смыслов.** В докладе будут рассмотрены следующие виды смыслов:
   a. Краткий пересказ, изложение высказанного в нескольких словах как ответ на вопрос «Что же он хотел сказать?». Или ответ на вопрос «Почему? Зачем?». Здесь вряд ли требуются пояснения, это тривиальный вид смысла, но в список он должен войти. Это логицируемый смысл.
   b. Смысл как выражение цели, намерения или причины высказывания. Цель может быть в выражении отношения, что прямо не отражается в тексте. «Он не хотел этого» может быть смыслом описания происшествия с непреднамеренным нанесением ущерба. Такая причина или цель может пониматься как смысл сказанного. Это логицируемый смысл, потому что всегда может быть точно выражен, если он понимается.
   c. Аналогия или метафора, он же «художественный» смысл. Высказывание приводится только как пример, аналогия того, о чем хочет сказать автор. И то, что он хочет сказать, становится смыслом высказанного. То же самое, что передаваемое «между строк». Все художественные произведения направлены на выражение какого-то смысла, который прямо, буквально в них не выражен. Выражаясь метафорически, автор не всегда может даже сам сказать, что общего у выраженного словами и тем смыслом, который он хотел передать с их помощью. Это не всегда логицируемый смысл.
   d. Мистический смысл. Это смысл — загадка, тайна, всегда передается при помощи специального многозначительного языка. Цель высказывания как раз и состоит в создании такой загадки. Ощущение таинственности возбуждает, как и страх, и по-

этому такой смысл многими любим. Автор намеренно не высказывает смысл прямо, так как логицирование убивает это возбуждение от тайны. Это нелогицируемый смысл.

4. **Почему смысл для нас так важен**, и мы во всем его стремимся увидеть? Для ответа на этот вопрос можно заметить, что во время диалога мы постоянно себя спрашиваем: «О чем это я?» или «Зачем это я?». Такая рефлексия цели участия в диалоге происходит постоянно, чтобы «не потерять мысль», и порой незаметно для нас. Стремление выразить свою мысль, приведя как можно больше аргументов, приводит нас к необходимости помнить и возвращаться к цели. Иначе мы забудем, о чем хотели сказать, после первого же предложения, как бывает при синдроме Корсакова. Именно по этой причине в процессе чтения мы в том же подозреваем и автора, пытаемся понять, что он хотел сказать, какой цели добивается. И получается, что мы ищем смысл по примеру того, как сами его придаем своему тексту. Эмпатия заставляет нас думать о тех же чувствах у другого.
5. **Понимание.** Знание смысла интерпретируется как понимание сказанного. Но это не совсем верно. Понимание — это активация и построение всех знаний, которые связаны с высказанным, но эксплицитно не выражены в тексте. Это может быть похоже на такой пример, приведенный Владимиром Селегеем: «Джей приглашена к Джеку на день рождения, и она подумала, понравится ли ему воздушный змей. Она идет в свою комнату и потрясет копилку. Из копилки не раздалось ни звука». Для нас это следование предложений одного из другого кажется естественным и понятным. Но ни одна программа компьютерной лингвистики (NLP) не сможет распознать связь этих предложений без подсказок. И все эти латентные знания «здравого смысла» мы только подразумеваем, когда произносим приведенный текст. *Это и есть понимание*.
6. **Проблема смысла в искусственном интеллекте.** Какие из указанных выше видов смысла можно было бы реализовать в искусственном интеллекте (ИИ)? Какие наиболее важные проблемы смысла в ИИ можно решить в ближайшее время? В докладе будет представлено, что на сегодняшнем этапе в ИИ могут быть реализованы только смыслы по цели или намерению, для чего еще предстоит разработать алгоритмы интерпретации текста относительно цели, к чему ближе всего автоматическое аннотирование. В то же время понимание текстов невозможно без решения проблемы common sense, но пока даже не намечено путей решения такой проблемы, хотя некоторые успехи продемонстрированы (проект COMET).

Литература

Common Sense Comes Closer to Computers, Quantamagazin, April 30 2020, https://www.quantamagazine.org/common-sense-comes-to-computers-20200430/

Гаспарян Д. Трансцендентальное измерение смысла и проблема искусственного интеллекта, Academia.edu, март 2020.

МЕНЕДЖМЕНТ МУЛЬТИМОДАЛЬНОЙ КОММУНИКАЦИИ СИНХРОННЫХ ПЕРЕВОДЧИКОВ

Е. В. Чистова

kovelena82@mail.ru

ФГАОУ ВО «Сибирский федеральный университет»

Цель доклада заключается в представлении результатов исследования соотношения вербальной и невербальной коммуникации синхронных переводчиков в условиях мультимодальной коммуникации как когнитивного события. Данные предназначены для разработки дизайна тренажера-симулятора в виртуальной среде, способствующего более эффективной подготовке синхронных переводчиков.

Мультимодальная коммуникация (Кибрик и др. 2019) представляет собой совокупность эксплицитных и имплицитных взаимосвязей различных семиотических модусов — ресурсов, интегрированных из вербальных и невербальных знаковых систем и предназначенных для эффективного понимания происходящих вокруг событий. Необходимо отметить, что, согласно исследованиям автора программы «бесшумные сообщения» Альберта Мехрабиана, 7% любого сообщения передаются через слова, 38% — через определенные вокальные элементы и 55% — через невербальные компоненты (Miletich 2015). Таким образом, **актуальность** данного исследования обусловлена, во-первых, малоизученностью невербального компонента в мультимодальной коммуникации переводчиков, а во-вторых, возросшим интересом к изучению виртуальных коммуникативных агентов — собеседников (Allwood, Ahlsén 2009; Robison 2009).

Научная новизна заключается в том, что впервые в качестве *объекта* исследования выбрана когнитивная переводческая деятельность в аспекте соотношения вербальной и невербальной коммуникации как важнейшая составляющая менеджмента мультимодальной коммуникации.

Для фиксации речевой деятельности и сопроводительных жестов переводчиков используется *метод анализа когнитивного события* (Колмогорова 2019). Для аннотирования мануального канала в ситуации когнитивного диссонанса применяется *методика, разработанная в среде ELAN* (tla.mpi.nl/tools/tla-tools/elan), позволяющая обрабатывать жесты, адапторы и смены позы. Исследование проводится *на материале* видеозаписей синхронных переводчиков, сопровождающих реальные мероприятия.

Применение данных методов и принципов распределенной когниции к переводческой деятельности позволяет решить ряд важных *задач*: 1) описать средства невербальной коммуникации, сигнализирующие о когнитивном диссонансе в стрессовой ситуации; 2) выявить процент соотношения вербальных и невербальных компонентов в переводческом акте на материале видеозаписей работы красноярских переводчиков; 3) разработать кейсы для тренинга устных переводчиков в виртуальной среде, способствующей развитию навыков невербальной коммуникации в переводе.

В ходе исследования осуществлена разметка данных в объеме 34 эпизодов, составляющих по времени 59 минут. В разметке выделено 18 слоев, по 9 измерителей на каждого партнера: речь, вокализация, паузация, неречевые звуки, мимика, взгляд, положение головы, жесты, действия (см. рис. 1).

В результате исследования были выявлены следующие *средства мультимодальной коммуникации, указывающие на когнитивный диссонанс переводчика*: паузация в диапазоне 56,06–61,01секунд, вокализация (эмм, ымм, ээээ, растягивание гласного в слове), нерегулярный жест руками (например, вертикальное положение руки вместо доминирующего горизонтального), смена взгляда и положения головы (например, с оратора на блокнот), смена

положения тела (принятие более вертикального положения), листание блокнота, экрана ноутбука и др.

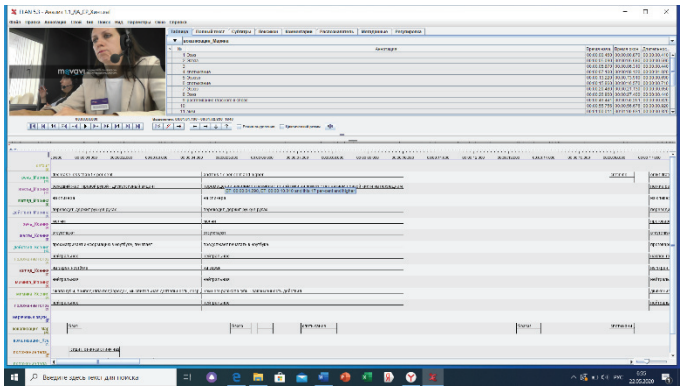


Рис. 1. Фрагмент анализа когнитивного события в среде ELAN 5.3

Ключевым выводом является фиксация доминирующей позиции *жестов собственного менеджмента мультимодальной коммуникации* синхронных переводчиков над жестами, сопровождающими фактическую информацию в речи, что не свойственно для человека в обыденном общении (согласно исследованию Allwood, Ahlsén 2009). К жестам собственной коммуникации относятся, например: режущий жест рукой, расставляющий дискурсивные акценты в речи переводчика, который таким образом осуществляет внутренний контроль над смысловым началом и завершением фразы, регулирует темп и равномерность подачи информации реципиентам. Этот факт показывает, что синхронный переводчик способен одновременно не только воспринимать и воспроизводить информацию, но и параллельно осуществлять несколько видов невербальной коммуникации.

References

Кибрик А. А., Коротаев Н. А., Федорова О. В., Евдокимова А. А. 2019. Единая мультиканальная аннотация как инструмент анализа естественной коммуникации // Диалог. С. 288–303.

Колмогорова А. В. 2019. Лингвистика когнитивная и экологичная: к вопросу о перспективах применения концепции когнитивной экологии в лингвистических исследованиях // Экология языка и коммуникативная практика. № 3 (18). С. 19–28.

Allwood J., Ahlsén E. 2009. Gestures that precede and accompany speech — An analysis of their functions and use in the design of virtual agents in different activities // Multimodal Communication — from Human Behaviour to Computational Models. Proceedings of the NODALIDA 2009 workshop. Vol. 6, pp. 3–9.

Miletich M. 2015. Accounting for nonverbal communication in interpreter-mediated events in healthcare settings // Translation and Translanguaging in Multilingual Contexts, 1:2. Pp. 162–181. DOI 10.1075/ttmc.1.2.02mil.

Robison J. 2009. Affective Support in Narrative-Centered Learning Environments // Affective Computing and Intelligent Interaction. Proceedings of The Doctoral Consortium. Amsterdam, pp. 17–25.

КАЧЕСТВЕННЫЙ АНАЛИЗ ОШИБОК В ТЕКСТАХ НЕЙРОСЕТИ ПОРФИРЬЕВИЧ

С. А. Шаповал
Высшая школа экономики
sv.shapoval@gmail.com

Нейросеть Порфирьевич представляет собой сеть GPT-2 (Radford et al. 2019) компании OpenAI Илона Маска, обученную на русском корпусе М. Гранкиным. Оригинал настолько хорошо умеет генерировать осмысленные тексты, что полная версия некоторое время не выкладывалась в общий доступ из-за потенциальной опасности ИИ, способного легко производить правдоподобные фейки. Нейросеть Порфирьевич (https://porfirevich.ru) генерирует тексты на русском языке, продолжая введенные пользователем несколько слов или предложений. Получившийся в результате текст выглядит настолько осмысленным, что во многих случаях неотличим от текста, написанного человеком. Это обеспечивается, в частности, когезией, которая создается вполне обычными «человеческими» способами: с помощью местоимений, анафоры («О чем думает нейросеть*? Зачем ей нужно всё это? И самое главное, зачем ей вообще надо? Зачем она хранит эти воспоминания?*», синонимов («Йо! Я сыт! *Едрит твою налево! Я теперь наелся! И я больше не голоден!*») и т. п.

Со времен «Корчевателя» (Гельфанд 2008) машины научились многому. Можно ли (будет) каким-либо образом отличить тексты разного происхождения — например, по тем ошибкам, которые делают авторы?

Нами проанализированы тексты нейросети, доступные в ее Галерее (https://porfirevich.ru/gallery) и полученные в ходе эксперимента. В качестве стимула использовались фрагменты «Гости съезжались на дачу», «Это какие-то неправильные пчелы», «Мороз и», «На свадьбе у Ольги Ивановны» и др. начала.

Тесты на осмысленность, например Winograd Schema Challenge (Levesque H. 2011), нейросети сейчас в целом проходят, но типичные ошибки остаются, к ним относится переключение тем бессмысленным образом. Формально это выражается в отсутствии связи между высказываниями, однако имплицитная смысловая цельность текста (когерентность) присутствует, она достраивается читателем. Так, между началом «Гости съезжались на дачу» и продолжением «*Я, кажется, уже говорил тебе, что в той комнате с колоннами была огромная бочка*» формально связи нет, но презумпция связанности двух рядом стоящих предложений заставляет воспринимать «ту комнату с колоннами» как дачную, заполнять скважину (Жинкин 1972); ср. пример N. Enkvist «Джон болеет дома, а формула уксусной кислоты — СН3СООН», который обретает смысл, если рассмотреть фразу как ответ на вопрос «Где Джон? Он обещал сказать мне формулу уксусной кислоты» (Цит. по: Литневская 2015: 120).

При явном (и ярком) отсутствии связности можно говорить об эффекте обманутого ожидания: «Купил мужик шляпу, а она ему *для пущего правдоподобия выстрелила в спину из пулемета*». Непредусмотренный (?) комический эффект переводит подобного рода тексты в разряд художественных; такое же впечатление производят тексты Зощенко, Хармса и др. мастеров абсурда. Наблюдения показывают, что Порфирьевич способен играть с многозначностью, антонимией, внутренней формой и другими языковыми явлениями, ср. «Что происходит на свете? *И есть ли на свете что-нибудь более беспросветное?» — думала она, с тяжелым вздохом шагая к палисаднику, за которым стоял московский автобус*». Вопрос о степени сознательности производства каламбуров остается открытым.

Даже беглый обзор показывает, что ошибки в текстах нейросети встречаются нечасто, т.к. когезии Порфирьевич научился, а когерентность обеспечивается читателем, но нам удалось выявить несколько типов ошибок, которые нарушают общее впечатление связности и цельности текста. Важно, однако, что такие же ошибки мог бы сделать и человек.

Рассмотрим пример «Гости съезжались на дачу. *Были такие, кто решил провести праздничное застолье в одном из двух общих спальен, расставленных в глубоких нишах*». Трудно реконструировать начальную форму слова, которое мы знаем как «спальня»: здесь оно м.р. и обозначает скорее кровать, т. к. может являться «расставленным» в нишах. Та же странность в примере «*…она была … в лагерях и лагерье*». Ошибка в образовании формы слова есть грамматическая ошибка, целые коллекции которых собирают преподаватели. В текстах Порфирьевича встречается неправильное употребление глагола («переругиваясь про старых друзей»), рассогласование местоимений, нарушение условий канонической речевой ситуации (Падучева 1995: 43) и др. типы ошибок. Практически любую из них можно представить в тексте, например, школьного сочинения. Некоторые ошибки настолько неожиданны на общем грамотном фоне, что выглядят как опечатки: «Гости съезжались на дачу. *И вдруг ты как заорет!*» — в том смысле, что их также можно оправдать, счесть случайными.

По всей видимости, искусственную природу текста выдает некоторая необычная сочетаемость «*у меня в носу потекла слеза. А потом в груди возник синий ореол и пошел вверх*» — однако смысловые процессы и эффекты здесь сложны и требуют отдельного исследования.

Наконец, время от времени Порфирьевич порождает текст, в котором наблюдается, например, невозможная последовательность событий: «На свадьбе у Ольги Ивановны *случилось много разных вещей: она погибла, была интернирована в лагерях и лагерье, узнала много нового и познакомилась с таким количеством иностранцев, что у нее создалось ощущение, что она оказалась в экзотической стране со своей собственной культурой».* Ошибки в области невербализуемого знания, представлений об окружающем «по умолчанию», здравого смысла, пресуппозиций и т. п., по-видимому, достаточно ясно отличают (пока) тексты ИИ от текстов человека.

Литература

Гельфанд М. 2008. Четыреста первый способ Остапа Бендера // Троицкий вариант. Выпуск № 13N (839) 30 сентября 2008 г.

Жинкин Н. И. 1998. Язык — речь — творчество. (Избранные труды). М.: «Лабиринт».

Литневская Е. И., Литневская О. А. 2015. К определению текстообразующих признаков когезии и когерентности // Вестник Московского ун-та. Сер. Филология. № 6. С. 116–123.

Падучева Е. В. 1995. В. В. Виноградов и наука о языке художественной прозы // Известия РАН. Серия литературы и языка. Т. 49. № 3. С. 39–48.

Levesque H. 2011. The Winograd Schema Challenge // AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning.

Radford A., Wu J., Child R., Luan D., Amodei D., & Sutskever I. 2019. Language Models are Unsupervised Multitask Learners. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

КОРПУС ДИАЛЕКТОВ И ИЗУЧЕНИЕ ВАРИАТИВНОСТИ ГОВОРОВ В ТАБАСАРАНСКОМ ОНЛАЙН-СЛОВАРЕ

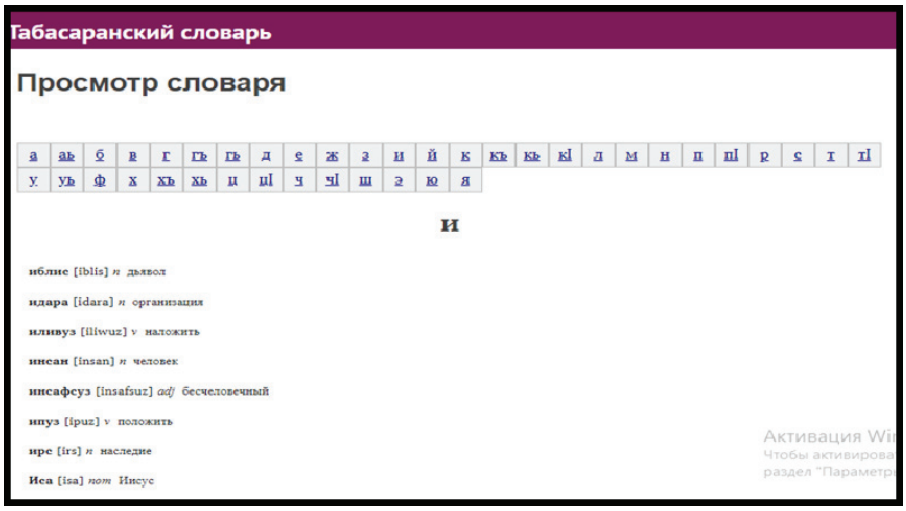
С. Х. Шихалиева

sh_shihalieva@mail.ru

Институт ЯЛИ им. Г. Цадасы ДФИЦ

Документация малых языков России в эпоху глобализации приобретает важное значение, расширяется технологическая сфера разработки различных онлайн-словарей. Для расширения технологической сферы табасаранского языка был создан специальный сайт онлайн-словаря, в данный момент находящийся на стадии разработки (https://webonary.org/tabasaran), см. Таблицу 1.

Таблица 1



Цель настоящей статьи — презентация первичного инструмента русско-табасаранского онлайн-словаря. При первичной документации данных онлайн словаря использовались методы компьютерной лексикографии, позволяющие определить не только задачи грамматического описания текстов, но и правила глоссирования.

Актуальность описания онлайн-словаря табасаранского языка определялась тем, что и поныне нет сайта с параллельными терминами. Поскольку сайт онлайн-инструмента представлялся в концепции художественного перевода, при анализе словоформ использовался принцип хронологизации табасаранских текстов: «Евангелие от Марка, 1997 // Евангелие от Марка, 2010» (посредством программ Paratext 8.3 и Flex). Такой принцип хронологизации выявил задачи классификации терминов в говорах и диалектах:

Евангелие от Мк 1:3 [1997] (кандикский говор нитрикского диалекта южного наречия)

1. Исаяйи	кIура	: « Раббийиз	рякъ	гьязур	апIинай !	
isaja-ji	k'u=ra	rabbi-jiz	rjaq̄	ḥjazur	ap'in=aj	
Исайя	говорить	Господь	дорога	приготовление	делать	
antrop:erg	prs	antrop:dat	nom	s>nom	imp.excl	

«Приготовьте путь Господу!» — говорит Исайя

The voice of one crying in the wilderness: «Prepare the way of the Lord, make his paths straight».

Евангелие от Мк 1:3 [2010] (ляхлинский говор этегского диалекта северного наречия)

2.	ИчIи	чюлиъ	дих	атIабгура	: "	Агъайиз	рякъ	ккабалгай .
	ičʹi	čjul-iʔ	dix	atʹabg-u=ra		aɣa-jiz	rjaq̄	k̄abalg=aj
	пустой	степь	зов	раздается		Господь	дорога	нарядить
	adv	obl	n	prs		antrop:dat	nom	imp.excl

«Голос раздается в пустыне: «Приготовьте путь Господу».

The voice of one crying in the wilderness: «Prepare the way of the Lord».

При первичной документации онлайн-словаря табасаранского языка выявилась проблема систематизации антропонимов «говор ↔диалект ↑ наречие». Представляется, что описание систематизированных данных позволит ранжировать словоформы онлайн-инструмента с основой диалекта литературного языка. Эти концептуальные вопросы онлайн-словаря табасаранского языка будут обсуждаться на форуме.

Литература

Кибрик Е. А. 1982. Табасаранские этюды. М.: МГУ. С. 34–44.

Кибрик А. А. 1999. Три ахиллесовы пяты функционализма // Типология и теория языка: От описания к объяснению. Под ред. Я. Г. Тестельца и Е. В. Рахилиной. М.: ЯРК. С. 36–49.

Кибрик А. Е. (ред.) 2005. Малые языки и традиции: существование на грани. Выпуск 1. Лингвистические проблемы сохранения и документации малых языков. М.: Новое издательство.

Мельчук И. А. 1974. Опыт теории лингвистических моделей «Смысл ↔ Текст». Семантика, синтаксис. М.: Наука. С. 23–35.

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

Теньер Л. 1988. Основы структурного синтаксиса. Редкол.: Г. В. Степанов и др. М.: Прогресс.

ПРОБЛЕМА КАТЕГОРИЙ ТЕКСТА: СЛОЖНОСТЬ, ДИНАМИКА И НЕКОТОРЫЕ ПЕРСПЕКТИВЫ РЕШЕНИЯ

И. А. Щирова
schirova@yandex.ru

А. Г. Гурочкина
i.arkhipov@yandex.ru

РГПУ им. А. И. Герцена

На этапе становления лингвистики текста исследователи столкнулись с задачей вычленения его конститутивных свойств (категорий). Отграничивая текст от иных языковых феноменов, в первую очередь высказывания, категории подтверждали его качественную определенность, а значит, и правомерность признания текста самостоятельным научным объектом. Актуальность проблемы текстовых категорий сохраняется, а её сложность и вариативность решений, в т.ч. по отношению к категориям художественного текста, возрастают. Это объясняется сложностью взаимосвязи концептуально-тематического содержания художественного текста с вербализующей его художественной формой; сложностью реальности, на основе которой конструируется фикциональный мир, и сложностью коммуницирующих сознаний, поскольку именно на их «диалогических рубежах», — воспользуемся фразеологией М. Бахтина, — создается «эстетическое событие» текста (Бахтин 1986: 16).

«Одним из самых сложных понятий науки вообще» исследователи называют разум человека (Кубрякова 1996: 153). Оправданно предположить, что текст как продукт уникальных способностей *творческого* разума априори является сложным исследовательским феноменом. В нем объективируется взаимодействие чувственного восприятия и разума в рамках единого когнитивного цикла; памяти, регулирующей взаимоотношения автора с реальностью, авторского воображения и авторской фантазии, которые трансформируют воспоминание в многообразии художественных форм. Текст рождается в «подвалах сознания» (фразеология В. В. Налимова), предполагает участие подсознательных и бессознательных феноменов, например, интуиции, важной для креативного субъекта, но с трудом укладывающейся в понятийно-терминологические рамки лингвистического описания. Традиционно сложными признаются и художественные формы изображения всех упомянутых процессов (ср. модернистские текста «потока сознания»).

Сложность художественного текста как системного образования и иерархически организованной структуры, в свою очередь, регулярно подтверждается многочисленными описаниями текстовых «измерений»: актуализируемых в тексте авторских смыслов, обрастающих семантическими обертонами по мере их интеграции смысловым целым; различных форм структурной организации текста (тематической, повествовательной, синтаксической и пр.); системы интегрируемых в него интердискурсивных включений; уникальной иерархии образов, в которой реализуется специфика текста как индивидуально-авторского варианта концептуализации мира и т. д.

Понятийные и методологические акценты в подобных описаниях меняются, коррелируя с парадигмальными установками науки на отдельных этапах ее развития. Примером «революционной» динамики в развитии взглядов на свойства текста может служить отказ лингвистики текста от чисто формальных устремлений. В результате обретения коммуникацией и диалогом статуса базовых философских категорий описание текста как системы взаимозависимых элементов формы вытесняется построением коммуникативно-когнитивных моделей его представления. В тексте видят коммуникативное событие, многокомпонентную и неразрывно связанную с реальностью абстрактную коммуникативную ситуацию; механизм, ‘запускающий’ когнитивные процессы восприятия и декодирования (В. Е. Чернявская). Как следствие, меняется направление изучения категорий текста. Разработки категории связности уступают место анализу содержательных взаимосвязей, ассоциирующихся с категорией

целостности. Целостность текстового смысла соотносится с взаимодействием когнитивно-коммуникативных субъектов: автора и читателя. Вырастает интерес к текстовым категориям, эксплицирующим идею «сотворчества» этих литературных коммуникантов, как то «интертекстуальность» или «адресованность». Бахтинские идеи диалога, на которых они основываются, становятся методологическим стимулом для формирования принципиально новых перспектив в понимании текста и общающихся в нем сознаний как целостных позиций.

Проблема категорий текста и сегодня не утрачивает своей сложности, а лакуны в их описании отчасти предопределяют перспективы её решения. Отсутствует единое мнение о статусе текстовой категории, недостаточно четко сформулировано её соотношение с понятием, концептом и иными видами категорий. Например, с философской, лингвокультурной или коммуникативной, не всегда ясно место категории в категориальной иерархии: в разряд категорий попадают свойства разной степени абстракции и универсальности. Сложность систематизации текстовых свойств усугубляется открытостью их списка, которая закономерно предопределяется созданием нового научного знания, вычленением новых свойств текста и поиском новых концептуальных, методологических и ценностных ориентаций для их изучения. Фрагментарность описанию категорий придает и сама специфика метода моделирования: выдвигая на первый план свойство, релевантное для решения конкретных задач, исследователь абстрагируется от иных свойств текста, не менее значимых для его понимания.

Отчасти эта фрагментарность описания может быть компенсирована разработкой холистических принципов текстообразования и текстовосприятия и, в частности, акцентуацией холистического видения текста как ингерентного компонента литературной коммуникации, а его свойств как когнитивных параметров, зависящих от интерпретативной ситуации. Изучение непрямолинейного взаимодействия текстовых категорий может быть включено в контекст идей лингвосинергетики: так, фундаментальная категория целостности материализуется в формальной связности текста, но является при этом качественно новым образованием, а не простой совокупностью составляющих текст частных смыслов. Перспективной представляется реализация в исследовании гибких теоретических и методологических подходов, например, моделирование категориального пространства по прототипическому образцу, а также придание описанию междисциплинарного характера, преимущества которого едва ли нуждаются в обосновании. Методологически перспективным в этом отношении представляется и привлечение возможностей искусственного интеллекта, учитывающих присутствие человека, формой существования которого можно назвать художественный текст.

Библиография

Бахтин М. М. 1986. Эстетика словесного творчества. М.: Искусство.

Кубрякова Е. С., Демьянков В. З., Панкрац Ю. Г., Лузина Л. Г. 1996. Краткий словарь когнитивных терминов. М.: Филологический факультет МГУ им. М. В. Ломоносова.

Научное издание

Международная конференция
«Лингвистический форум 2020: Язык и искусственный интеллект»
12–14 ноября 2020 г. Институт языкознания РАН, Москва.
Тезисы докладов