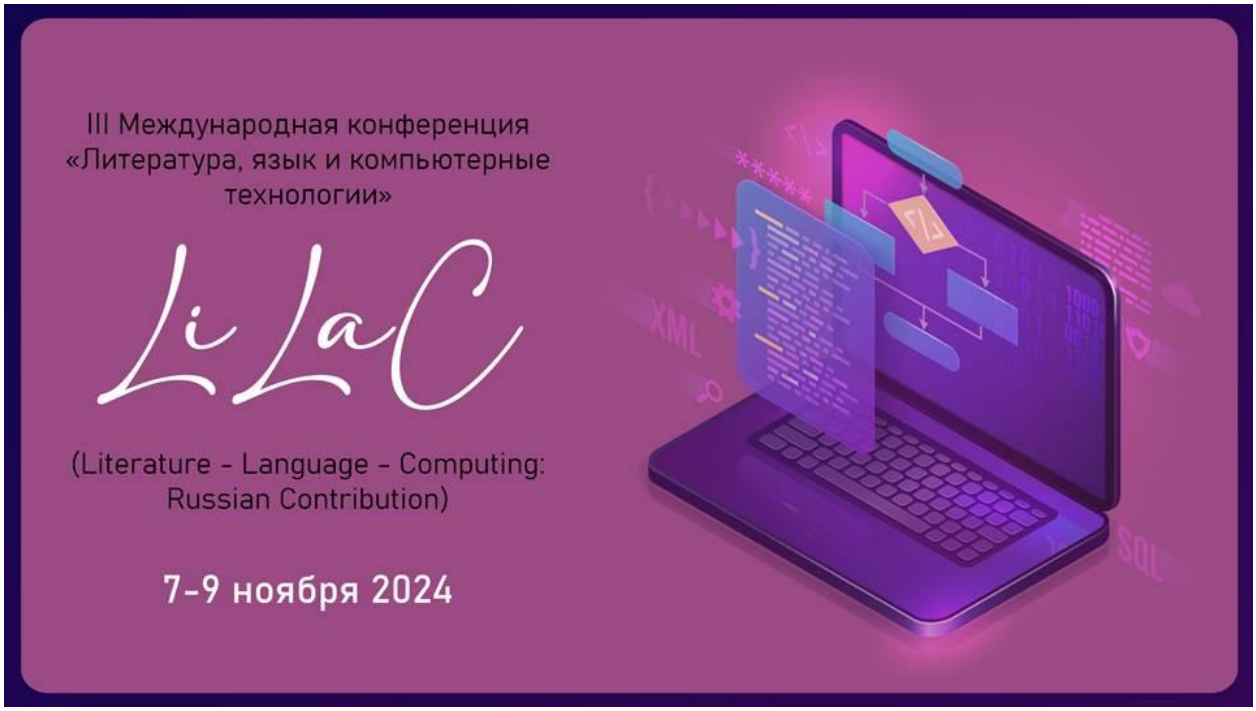
III Международная конференция «Литература, язык и компьютерные технологии»

Санкт-Петербург, Россия

7-9 ноября 2024

Сборник тезисов



Корпус исторических локальных текстов: разметка, архитектура, функционал

Бабина Ольга Ивановна, Орехова Елизавета Владимировна, Южно-Уральский государственный университет, babinaoi@susu.ru, kravtcovaev@susu.ru

Работа направлена на освещение результатов проекта по созданию корпуса текстов «Челябинск: корпус исторических локальных текстов», включающего тексты, посвященных уральскому городу на рубеже XIX-XX веков. Целью проекта является создание концептуально размеченного корпуса текстов и цифрового ресурса (корпусного менеджера), поддерживающего проведение историко-культурных исследований методами корпусной лингвистики и контент-анализа.

Разработка корпуса исторических локальных текстов представляет собой многоаспектное исследование, в ходе которого были поставлены и решены следующие задачи:

1. Осуществить сбор исторических текстов для включения в базовую коллекцию корпуса;
2. Определить модель метаразметки исторических текстов и схемы концептуальной разметки лексических единиц в текстах;
3. Разработать методологию и программную поддержку отбора единиц контент-анализа;
4. Спроектировать структуру базы данных, обеспечивающую хранение размеченных текстов;
5. Создать цифровой редактор для выполнения процедур аннотирования текстов в соответствии с определенной метаразметкой и концептуальной разметкой;
6. Разработать методику разметки текстов, с применением функционала созданного цифрового редактора;
7. Выполнить разметку корпуса исторических локальных текстов;
8. Спроектировать и реализовать интерфейс поиска по размеченному корпусу текстов (корпусный менеджер).

В базовый набор текстов создаваемого корпуса вошли тексты, состоящие из заметок и различных публикаций жителей города Челябинска, а также путешественников-гостей города, в которых они описывали свои впечатления об облике города в целом, архитектуре, инфраструктуре, местных жителях и многом другом. В корпус вошли 15 книг различной жанровой принадлежности (мемуары, травелог, исторический очерк, публицистические тексты), в том числе сборники. Объем корпуса составил около 350 тысяч словоупотреблений.

Процедура проведения контент-анализа предписывает необходимость выделения концептуальных категорий, индикаторы которых могут быть найдены в тексте. В рамках реализации проекта была определены релевантные для исторических исследований категории [Никонова, 2022], положенные в основу схемы концептуальной разметки корпусных единиц анализа (непересекающихся языковых последовательностей любой длины в пределах текста). Для разработки схемы разметки был проведен квантитативный и семантический анализ лингвистических единиц текста с применением ранее разработанного инструментария и процедур [Бабина, 2023; Шереметьева, 2023].

С учетом этих данных была составлена развернутая многоуровневая онтология актуализируемых в исторических текстах концептов. Максимальная глубина иерархии концептов в онтологии составила 4 уровня. Общий объем онтологии составил 160 концептуальных категорий.

Единицей корпуса выступает фрагмент текста на уровне сверхфразового единства (СФЕ) в соответствии с периодизацией описываемых в этих фрагментах событий. В настоящий момент в корпусе содержится 192 СФЕ. Каждой единице уровня СФЕ были

поставлены в соответствие хронологические рамки. Хронология определялась на основе эксплицитно указанной информации в тексте. Элементы корпуса, не включающие явного указания на хронологию, размечались экспертом-историком на основании экстралингвистических сведений, метапараметров источника и прочих критериев. Тексты в историческом корпусе были размечены в соответствии с созданной категоризацией концептуальных меток. В предложениях СФЕ отграничивались минимальные лексические единицы, несущие концептуально-значимую информацию (как правило, слова, словосочетания), и размечались меткой соответствующего концепта.

Для проведения работ по добавлению текстов в корпус, разметке корпуса, а также проведения историко-культурных исследований, был создан корпусный менеджер на базе системы управления контентом WordPress. Создаваемый ресурс воплощается в форме облачного сервиса, обеспечивающего доступ к базе размеченных исторических текстов. Хранение данных осуществляется в базе данных, содержащей сущности *Авторы, Книги, Статьи*. Онтология реализована через справочник *Ключевые слова*, где концепты заданы и организованы в иерархию. Разметка текстов осуществлялась во встроенном текстовом редакторе. Тексты корпуса сопровождаются разметкой на двух уровнях:

- метаразметка: включает параметры *Книга, Автор, Дата описываемых событий, Дата создания текста, Ключевые слова, Размеченный текст*;

- концептуальная разметка: проводится в поле *Размеченный текст*, на основе подгружаемой онтологии посредством заключения текстовой единицы в парные теги-шорткоды с параметром, определяющим id актуализируемых концептов онтологии. Например,

В настоящее время [cprword id="152"]*школьное занятие*[/cprword] *ведется* [cprword id="71,80,81"]*священноцерковнослужителями*[/cprword] [cprword id="80,81"]*монастырской церкви*[/cprword] *и* [cprword id="68,71,80,81"] *послушницами монастыря*[/cprword],

где единицы размечены посредством кодов концептов различных уровней: *Образование* (id концепта 152), *Социальный/профессиональный статус* (71), *Религия* (81), *Монастыри* (80), *Пол* (68).

Интерфейс корпусного менеджера (доступен по адресу http://cheltext.susu.ru) позволяет проводить поиск данных в корпусе, получение статистической информации, составление конкордансов и предоставления результатов пользователю в удобной форме. Поиск в корпусе возможно осуществлять по: а) дате (создания текста, описываемых событий); б) названию книги; в) автору текста; г) ключевым словам (концептам онтологии); д) лексическим единицам. В отобранных по поисковому запросу текстах возможен переход на страницу с полным текстом, где концептуально размеченные единицы имеют вид гиперссылок, позволяющих составлять конкордансы по выбранным концептам.

Проект по созданию корпуса текстов, реализовывался в рамках исполнения *Государственного задания Минобрнауки Российской Федерации* № FENU-2020-0021 (2020070ГЗ) по теме «Изучение региона в контексте глобально-исторических связей с помощью методов цифровой гуманитаристики (на примере Челябинска и Челябинской области)».

Список литературы

1. Бабина, О.И. Концептуальное моделирование предметной области для построения датасета / О.И. Бабина // Актуальные вопросы филологии и лингводидактики: современные тенденции и перспективы развития. – Санкт-Петербург, 2023. – С. 5–13.

2. Никонова, О. Ю. Железная дорога и миграция в уездном городе Челябинске в конце XIX – начале XX в. / О.Ю. Никонова, А.А. Тимофеев // Уральский исторический вестник. – 2022. – №1 (74). – С. 137–146. – DOI: 10.30759/1728-9718-2022-1(74)-137-146

3. Шереметьева, С. О. Концептуальное моделирование лексики предметной области / С. О. Шереметьева, Е. Д. Неручева // Вестник ЮУрГУ. Серия «Лингвистика». – 2023. – Т. 20, № 1. – С. 65–72.

Лингвостатистический анализ художественного текста с учётом его эмоционально-смысловой доминанты

Белянин В.П., Независимый исследователь, psyling@gmail.com

Анализ художественного текста – это его истолкование с помощью терминов и понятий научного языка (литературоведения, лингвистики, психологии, эстетики, социологии, культурологии, философии). Художественный текст изобилует метафорами, сравнениями, образами, недосказанностями; он имплицитен и иносказателен и по своей сути. Именно поэтому предпринимаются попытки привести смысл художественного текста к общему знаменателю – как бы перевести его на язык более или менее однозначных понятий.

Мы предлагаем подходить к художественному тексту ещё с одной стороны – со стороны психолингвистики. Ключевым понятием нашего подхода является понятие доминанты, которая в физиологии представляет собой общий «modus operandi центральной нервной системы» [1: 14]. Понятие доминанты есть и в эстетике, где она понимается наряду с полифонией как конструктивный принцип организации художественного текста [2]. Для выделения доминанты в художественном тексте мы также обратились к эвропатологии, как дисциплине, где развивалась концепция о связи феноменологии гениального человека с симптомами психопатического ряда [3].

При таком подходе художественный текст предстаёт не только отражением действительности, но и выражением средствами языка авторского отношения к миру [4]. Это позволяет рассматривать художественный текст как с позиций нарратологии, так и с позиций когнитивной психологии. Гибридный индуктивный и дедуктивный анализ художественных текстов предполагает сочетание в себе исследовательского характера индуктивного анализа и структуру, обеспечиваемую дедуктивными схемами. Тем самым, использование итеративного процесс, начинается с исследовательского чтения художественного текста, при котором происходит индуктивное кодирование с целью выявления имеющихся там конструктов [5], а затем и дедуктивное кодирование для применения имеющейся гипотезы (теории эмоционально-смысловой доминанты).

В последнее десятилетие одним из распространённых методов анализа текста является сентимент анализ как определение тональности, отношения автора текста к какой-либо теме (объекту). Анализ тональности объединяет в себе методы обработки текстов с привлечением машинного обучения для присвоения взвешенных оценок настроения сущностям, темам и категориям в предложении или во фразе [6]. Мы использовали метод сентимент-анализа для анализа художественного текста в контексте теории эмоционально-смысловой доминанты, взяв за основу типология, в которой тексты разбиваются на определённые типы («весёлые», «печальные», «светлые», «тёмные», «красивые» [7]).

В ходе анализа художественных текстов были составлены лексиконы из слов, которые наиболее часто встречаются в тех или иных типах текстов. Лексиконы были устроены единообразно: все они содержали одинаковые рубрики на всех уровнях, но имели разное наполнение. Если на первом уровне находились сами типы текстов, то на втором уровне были такие классы, как Person, Qualities, Emotions, Relations и др. Третий уровень содержал более дробные рубрики, так, в классе Relations имелись такие группы, как Together, Conflict, Alone, в группе Activity – Job, Politics, Religion и др., в группе Emotions – Fear, Anger, Disgust, Sadness и др., в группе World – Time, Space, Smell, Color, Taste и др. Четвёртый уровень лексиконов составляли собственно лексические элементы. К примеру, в группе Disaster класса Dark были такие слова, как *hurricane, storm, thunder*, а в группе Unreal класса Light – *angel, devil, ghost, illusion, magic, miracle, unreality;* в то же время в классе Merry в группе Unreal содержались такие элементы, как *nymphs, occult, psilocybin*.

Иными словами, лексиконы различались по наполнению и при этом частично пересекались.

Мы провели анализ тональности более 200 художественных текстов, а также скриптов субтитров более 100 фильмов и эпизодов сериалов на английском языке в среде R Studio. Некоторые результаты проведённого анализа оказались таковы:

- теория эмоционально-смысловой доминанты находит в целом своё подтверждение при лингвостатистическом анализе художественных текстов, хотя отдельные положения требуют уточнения;
- часть текстов можно отнести преимущественно к одному типу, большинство же текстов полимодоминантны;
- чем короче текст, тем меньше можно доверять результатам лингвостатистического анализа, и наоборот – чем длиннее текст, тем результаты надёжнее;
- разные тексты разных авторов могут быть близки между собой по лексиконам и, соответственно, по эмоционально-смысловой доминанте (например, Hemingway "The Old Man and the Sea" и E.M. Remarque "All Quiet on the Western Front" (corr = 0.998));
- разные части одного текста (напр., введение и основной текст) могут различаться между собой;
- разные модификации одного текста (напр., книга и научный комментарий, книга и скрипт субтитров фильма) могут различаться между собой;
- психолингвистическая структура некоторых текстов может указывать на их рецепцию читателями (напр., текст Edgar Poe "Raven" является «печальным» и немного «тёмным», а текст Ф. Энгельса «Немецкая идеология» – «светлым», доминанта которого коррелирует с паранойяльностью как склонностью к образованию сверхценных идей).

Конечно же, выделение ключевых слов художественного текста представляет собой большую проблему и метод лексиконов – лишь один из возможных. При этом применение программных методов анализа текста позволяет значительно ускорить анализ текста. Проведённое исследование – продолжение движения по пути анализа продуктов речевой деятельности, которое, как нам представляется, может быть перспективным для исследования художественной коммуникации методами квантитативного анализа и психолингвистики.

Список литературы

1. Ухтомский А.А. Доминанта. Статьи разных лет. 1887-1939.
2. Христиансен Б. Психология и теория познания.– М., 1907.
3. Клинический архив гениальности и одаренности.– Свердловск, 1925-1930.
4. Степанов Г.В. Цельность художественного образа и лингвистическое единство текста // Лингвистика текста. ч. II.– М., 1974.
5. Келли Дж. Теория личности: психология личных конструктов. М., 2000. СПб.: Речь, 2000.
6. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
7. Белянин В.П. Основы психолингвистической диагностики: (Модели мира в литературе).– М.: РАН Ин-т языкознания, Тривола, 2000.

Взаимоперевод и лексикографическое толкование аграрных терминов узбекского и русского языков

Бурхонова Гузал Мухаммадиевна
Ташкентский государственный аграрный университет, Ташкент, Узбекистан

Исследования в области терминологии всегда стояли в центре внимания лингвистов мира, так как с развитием науки и техники происходит постоянное обогащение лексики новыми терминами.

Известно, что XX веке – веке научно-технического прогресса при исторической общности народов бывшего советского союза русский язык сыграл большую роль в развитии терминологии узбекского языка. Большинство терминов, в том числе и аграрных, пришло в лексику узбекского языка из русского (причем многие из них прижились в узбекском языке без перевода), а международные – через русский язык. Об этом Х.Дадабоев [2019] пишет так:"Начиная с 70-годов XX века узбекская терминология вступила в этап развития на основе терминов, заимствованных через русский язык из западноевропейских языков". С.А.Захарова [2012] отмечает, что "у оседлых народов один из базовых пластов словарн ого фонда представляет земледельческая лексика. ... Терминосистема земледелия отражает поступательное движение научно-технического прогресса". Ф.К.Баротов [2012] также пишет, что "...все больше ощущается нехватка исследований, в которых приводилось бы толкование значения тех или иных труднопонимаемых земледельческих терминов". Исследования в области терминологии и ее перевода проводили в республике и зарубежом многие ученые-лингвисты, в т.ч. ВВ..Виноградов [1947], В. П.Даниленко [1977], , АА.Реформатский [1959], М.Ю.Авдонина и другие [2016], О.А.Фомина [2011], Т.Н.Данькова [2010], А.Н.Вороков и др. [2002], Ш.Б. Каримов [2021], О.Усмон, Р.Дониёров [1965, 1972], А.Мадвалиев и др. [2020].

Появление новых терминов в свою очередь требует от переводчиков кропотливой и качественной работы, но при этом в условиях глобализации нельзя стремиться к национализации науки, т.е. к переводу каждого термина.

Искуственное навязывание языку неудачных переводов не гарантирует им широкое использование в активной лексике. Часто такие термины или понятия естественным путем выпадают из языка, или в лучшем случае остаются в печати. А.Мадвалиев также отмечает, что "...несмотря на широкое применение международных терминов большинством языков мира, попытки найти им узбекских аналогов нельзя назвать положительным явлением".

В узбекском языке имеются словари некоторых отраслей сельского хозяйства, однако до сих пор не создан единый словарь агрологических терминов, который охватил бы терминологии всех отраслей сельского хозяйства и стал незаменимым помощником переводчика.

В работе проведен сопоставительный анализ "Толкового словаря узбекского языка" под редакцией А.Мадвалиева и "Толкового словаря русского языка" С.И.Ожегова и Н.Ю.Шведовой, который показал, что в обоих языках есть термины с терминэлементом агро-(агрономия, агрохимия, агротехника, агропромышленность), а в словаре под редакцией А.Мадвалиева имеются также термины с международными терминэлементами – гидро (гидропоника), микро- (микроэлементы), моно-(монокультура), фито-(фитопатология). Сложные термины узбекского и русского языков можно разделить на следующие группы: термиины, обозначающие отрасли сельского хозяйства; названия профессий сельского хозяйства; названия сельскохозяйственных культур; названия орудий труда, заимствованные термины, исконные термины и т.д. В русском языке названия профессий образуются в основном с помощью –**вод** (растениевод, животновод), а отраслей – путем добавления к нему суффикса –**ств-** и окончания –**о** (растениеволство,

животноводство). А в узбекском языке они образуются с помощью словообразовательных морфем –шунос (усимликшунос, чорвашунос) и –лик (усимликшунослик, чорвашунослик).

Также, проведена работа по изучению состояния исследований отраслевой терминологии в мировой и узбекской лингвистике; определению проблем исследования отраслевой терминологии в различных языках; определению системы аграрных терминов узбекского языка; систематизации семантического поля и групп аграрных терминов русского языка; сопоставительному анализу состава, качества и охвата аграрных терминов узбекского и русского языков; изучению опытов взаиперевода аграрных терминов узбекского и русского языков; определению проблем выражения аграрных терминов в толковых словарях узбекского и русского языков и выдвижению предложений по совершенствованию этих толкований; исследованию вопроса эквивалентности аграрных терминов узбекского и русского языков в переводных словарях.

Ключевые слова: международные терминэлементы, сложные термины, простые термины, образование сложных терминов, составные термины, взаимоперевод, толкование, калькирование.

Comparing and cognizing doppelgangers: a quantitative approach

Irina Golovacheva (St. Petersburg State University igolovacheva@gmail.com), Mikhail Zhuravlev (St. Petersburg State University myezhur@gmail.com)

The objective of our research is to view the literary topos *doppelganger* through the critical lense of the mathematical method that allows one to address several issues of comparative literature. We tackle the following questions: A) Is it possible to estimate to what extent the doppelganger topos in a particular text is similar to or different from the one in other texts? B) Does the estimated "degree" of similarity tell us anything new about the texts under scrutiny? C) What do the obtained results tell one about the evolution of the topos? D) How does the presence or the absence of a particular function or functions of the topos affect a doppelganger text in terms of its typicality. Besides providing quantitative/literary analysis proper, we claim that our research is cognitive as well. In particular, we pose the following question: Can the chosen quantitative approach help distinguish "narrative universals" (Hogan, 2020) or "schemata" (Emmot & Alexander, 2014).

To answer the above questions, we employ quantitative analysis, namely, the elementary approach provided by graph theory. Elementary graph theory and matrix theory have been used in literary studies since the 1950s when structuralists first ventured to explore communication between characters in various scenes and episodes, as well as to identify "levels" or "spaces" in which the characters function – primarily in folklore texts (Lévi-Strauss, 1955; Maranda 1973). Lately, graph theory and matrix analysis were applied to different fields of literary studies (Artemova, Komarova & Kretov, 2018; Golovacheva, de Mauny, Zhuravlev, 2021; Neumann & Dion, 2021). In the present work, the quantitative results, though playing an auxiliary role, serve as a trigger and provide a necessary framework for the subsequent literary analysis. In brief, our approach involves three stages. Firstly, we define what features of the chosen texts we will quantify. Then, we apply the most adequate mathematical method to investigate the chosen features. Finally, we assess the obtained results in the framework of literary criticism.

We selected the following canonic texts for our comparative study: "Peter Schlemihl's Miraculous Story" (1814) by A. von Chamisso, "Der Doppeltgänger" (1821) by E.T.A. Hoffmann, "The Nose" (1836) by N. Gogol, "William Wilson" (1839) by E.A. Poe, "Two Actors for One Part" (1841) by T. Gautier, "The Double" (1846) by F. Dostoevsky, "The Shadow" (1847) by H.Ch. Andersen, "The Devil. Ivan's Nightmare" a chapter from *The Brothers Karamazov* (1879–80) by F. Dostoevsky, "The Horla"(1886) by G. de Maupassant, *The Strange Case of Dr. Jekyll and Mr. Hyde* (1886) by R.L. Stevenson, "The Jolly Corner" (1908) by H. James, "The Secret Sharer" (1910) by J. Conrad, and *Despair* (1932-34) by V. Nabokov. The first scholar to describe the attributes of the doppelganger was Otto Rank who authored a seminal book "Der Doppelgänger"/*The Double*, (1925/2009). It is not surprising that Rank primarily highlighted psychological phenomena in doppelganger texts. To Rank's list of psychological attributes of the double, we added a list of specifically literary attributes, i.e. a set of diverse invariant motifs found in the above thirteen texts. We divided the complete set of attributes into four groups, namely, "psychological attributes of doubling," "physical manifestations of doubling," "elements of biography in doubling," and "mysterious attributes of doubling." We constructed an incidence matrix for each group to show in which texts a certain attribute appears. Then we calculated the matrix of similarity indexes for each group. The *similarity* here is the *simultaneous presence* or *simultaneous absence* of an attribute. Finally, we demonstrated the similarities and dissimilarities of texts basing on the complete list of all attributes in the four groups.

The quantitative method led to several important qualitative conclusions: it highlighted the "frontrunners" among doppelganger fictions. Dostoevsky's *The Double*, Conrad's "The Secret Sharer" and Gautier's "The Two Actors for One Part" apparently provide the best working doppelganger imagery "recipes" that appeared to be most adaptive for various artistic goals. One

can assume that any writer, while creating doppelganger imagery, processes certain archetypal phenomena and schemata. Comparing the works of different writers who belong to various literary traditions, we see that the results of such processing reveal a pronounced structure. Among other things, quantitative analysis allows us to survey and objectively compare various outcomes of the writers' reflections of such universal mental processes as defense and splitting.

Acknowledgements: The research is supported by St. Petersburg State University, grant 123042000103-6 [M1_2020_5].

References

Hogan, P.C. (2020). Narrative universals, emotion and ethics. *Poetics Today*, 41(2), 187-204. https://doi.org/10.1215/03335372-8172514

Emmott, C., and Alexander, M. (2014) Schemata. In: Hühn, P., Meister, J. C., Pier, J. and Schmid, W. (eds.) *Handbook of Narratology*, 2nd ed. Series: De Gruyter handbook, 1 (pp. 756-764). de Gruyter.

Lévi-Strauss, Cl. (1955). The structural study of myth. *Journal of American Folklore*, LXVIII, 428–444.

Maranda, P. (1973). Cendrillon: théorie des ensembles et théorie des graphes [Cinderella: set theory and graph theory]. In C. Chabrol (Ed.), *Sémiotique narrative et textuelle* (pp. 122–136). Larousse.

Artemova, O.G., Komarova, E.P., Kretov, A.A. (2018). Svyazuyushchie markemy v britanskoy proze vtoroy poloviny XIX v. [Linking markemes in British prose of the second half of the nineteenth century] *Yazyk i kul'tura* [*Language and Culture*], 53, 70–88. https://doi.org/10.17223/19996195/43/1

Golovacheva, I. V., de Mauny, P., Zhuravlev, M. Ye. (2021). Dvoyniki i matritsy: o novom metode komparativistskogo analiza [Doppelgangers and Matrices: A New Method of Comparative Analysis]. *Filologicheskiy klass* [*Philological Class*], 26 (2), 9–23. 10. https://doi.org/51762/1FK-202126-02-01

Neumann, M. D., Dion, L. & Snapp, R. (2021). *Teaching computational thinking: an integrative approach for middle and high school learning*. MIT Press.

Rank, O. (2009). *Double: A psychoanalytic study*. (Tucker H. trans.) University of North Carolina Press. (Original work published in 1925).

Cutoffs as Hesitation Phenomena in Russian Spoken Discourse of L1 and L2 Speakers: Functions and Quantitative Data on Speakers' Psychological Type

Daria Gorbunova, Independent Researcher, dgorbunova2@gmail.com, Kristina Zaides, Independent Researcher, zaides.kristina@gmail.com, Natalia Bogdanova-Beglarian, St. Petersburg State University, n.bogdanova@spbu.ru

The linearity of the spoken speech flow brings to life such typical ways of organizing the text as cutoffs, repetitions, self-repair, left statements, pause fillers, and silent hesitation pauses, etc. All such approaches arising when organizing a text can be united under the general concept of hesitation phenomena, which equally indicate various kinds of difficulties for the speaker in the process of production of spontaneous speech (the studies of H. Maclay, C.E. Osgood, T.M. Nikolaeva, V.I. Podlesskaya, A.A. Kibrik, H.H. Clark, J.E. Fox Tree, etc.). The reason for such difficulties is the conditions of time shortage, in which, as a rule, spontaneous speech is generated. Hesitation can also be called a form of speech disfluency, in which the smoothness of the speech flow is disrupted. Word cutoffs are very common in spontaneous spoken speech of native speakers of any language (the studies of W. Levelt, E. Shriberg, R. Lopez-Ozieblo). Different structural types of cutoffs can be distinguished:

1) purely hesitation cutoff, with repetition (refinement) of the cut off word (yes… yesli),
2) cutoff without repetition (to ye… vu-u v nachale avgusta),
3) break with correction (obsuzh… obsudit').

In relation to cutoffs without repetition and with correction, we can talk about cancellation operations: complete, when the cut off word is simply thrown out by the speaker and (s)he starts the phrase again (2), or partial, when the cut off word is repeated with some correction (3). In cases of a cutoff with correction (partial repair), it is important when exactly the speaker makes this correction: immediately after the broken word (online repair) or only after some time (offline repair).

A cutoff occurs, therefore, when the speaker realizes the error in the starting speech and either, after thinking (hesitation), still finishes the cut off word, or abandons it and chooses another word, or makes a repair. Several other reasons can be noted that cause a cutoff in a spontaneous monologue: a reaction to a visible voice recorder; the phonetic complexity of the word that the speaker intends to pronounce; phonetic influence of neighboring words; grammatical problems (especially in speech in a non-native language). It is clear that in speech in the native and non-native languages, these reasons manifest themselves to varying degrees.

The goal of this study is to investigate all types and features of the cutoff in relation to the psychological type of the speaker, as well as the speaker's language proficiency.

In this study, the speech of two groups of speakers (10 people in each group) is analyzed: (1) Russian monolinguals who speak Russian as their native language (L1R); (2) Chinese speakers who speak Russian as a target language (L2Ch). Both groups of speakers were balanced by gender (5 men and 5 women). In terms of age, all speakers were undergraduate and graduate students from 21 to 28 years old.

All speakers described the comic strip "Hair Elixir" by H. Bidstrup. These texts are part of the corpus Balanced Annotated Text Library, which is the corpus of monological speech created at Saint Petersburg State University. The speech was completely spontaneous since the speakers were given the comic literally before the recording began and did not have time to prepare for its description. The recorded texts were transcribed based on auditory analysis and annotated, marking all types of pauses, both syntagmatic and phrasal and hesitation. All hesitations, interruptions, repetitions, self-repairs, and paralinguistic elements (laughter, sigh, cough, etc.) were annotated as well. Cutoffs in transcripts of monologues were indicated by an ellipsis, not separated by a space from the cut off word. Other ellipses indicate an elongated hesitation by the speaker or a break of the phrase and were not taken into account during the analysis.

The data about the psychotype of each speaker was obtained after recording, using the H.J. Eysenck Personality Questionnaire, which they took in their native language. The group of monolinguals consists of 5 extroverts and 5 introverts; the group of Chinese bilinguals consists of 3 extroverts, 3 ambiverts, and 4 introverts.

In the speech of L1 speakers, the total number of cutoffs found is 24; in the L2 speakers' monologues, the total number of cutoffs is 138, which is more than five times higher. This can be explained by the higher number of difficulties that L2 speakers experience during spontaneous speech production when speaking Russian. However, the functions of these cutoffs can also vary in L1 and L2 speech. The cutoffs that were obtained from the spontaneous monologues belong to several functional types, according to the main reason for their appearance:

1) hesitation: posle etogo: u nego kazhdyi den' *V e-e volosy: sta… stanetsya stanetsya dlinneye i dlinneye (L2Ch);

2) word search and repair (the repair is underlined): nu znachit / on banochku tak lyubovno vzyal kak reb… () voobshche kak mladenca tak / hop / i takoi idyot domoi (L1R);

3) grammatic repair: i: (…) on smotrels… e-e on smotrel na sebe () a v zerkale / i ochen' (…) radovalsya (L2Ch);

4) phrase repair: na pervoi k… () v obshchem tut neskol'ko kartinok // vidimo oni sostavlyayut istoriyu (L1R);

5) word insertion: on ochen' ispugan // no tut (…) o… vnezapno on prosypaetsya v svoei zhe krovati / i: uzhe: smeyotsya () nad svoim (…) *V koshmarom (L1R);

6) pronunciation difficulty: no potom on (…) y uvidel chto / y volos rastyot shs… shlis… s… slishkom () bystro (L2Ch);

7) stipulation: do tekh por poka on shyol *S k zel… k zerkalu vol… volosy vyrosli eshchyo bol'she (L1R).

Grammatic repair is typical for L2 speech and occurs only in the monologues of Chinese Russian speakers where it is the second by frequency. The most frequent type of cutoffs function both in L1 and L2 monologues is hesitation (33 % in L1 and 68 % in L2); word repair is as frequent as hesitation in L1 speech, but not in L2 speech (7 %). The pronunciation difficulties and stipulations take 3-4 % for L1 and L2 speakers separately.

The findings indicate that bilingual speakers (L2Ch) show a significantly higher frequency of cutoffs compared to monolingual speakers (L1R). Specifically, L2Ch females showed the highest percentage of cutoffs at 60.45 %, while L2Ch males had 39.55 %. This suggests that language proficiency and familiarity with the language play crucial roles in speech fluency, with bilingual speakers experiencing more hesitations and cutoffs.

In terms of gender differences, L1R males produced more cutoffs than females (60.87% vs. 39.13 %), whereas in the L2Ch group, females had a higher percentage of cutoffs compared to males (60.45 vs. 39.55 %).

For psychotype analysis L2Ch ambiverts were excluded to maintain clarity and facilitate focused comparisons between extraverts and introverts. It reveals that introverts in both groups have a higher frequency of cutoffs compared to extraverts. Among L1R speakers, introverts accounted for 65.22 % of cutoffs, while extraverts accounted for 34.78 %. Similarly, in the L2Ch group, introverts produced 60.44 % of cutoffs compared to 39.56 % by extraverts. Furthermore, previous studies have shown that introverts typically demonstrate a greater frequency of various types of hesitation, including cutoffs. These findings suggest that introverts may face more challenges in spontaneous speech, potentially due to higher levels of self-monitoring and anxiety.

Hesitation phenomena, such as cutoffs, occur in any speech: monologue and dialogue, any type of speech genre, including speech of native and non-native speakers. Analysis of such phenomena, especially taking into account the social and psychological characteristics of speakers, allows us to obtain many interesting observations, which, on the one hand, clarify our ideas about the mechanisms of spontaneous speech production and on the other hand, make it possible to more adequately model the speaker's speech behavior in different communicative

situations and apply these observations for different purposes, from automatic speech processing to improving artificial intelligence systems.

Keywords: Cutoffs, Introverts, Extroverts, Psychotype, Personality Traits, Speech Patterns, Spontaneous Speech, Psycholinguistics.

Acknowledgments: The presented research was supported by Saint Petersburg State University, project No. 124032900006-1 "Modeling the Communicative Behavior of Residents of a Russian Megapolis in the Socio-Linguistic and Pragmatic Aspects with the Use of Artificial Intelligence Methods".

Частые глаголы в составе лексико-семантических групп и местоименные наречия как средство дифференциации идиостилей классической прозы

Ф. Н. Двинятин (СПбГУ), f.dvinyatin@spbu.ru

Многие исследовательские модели, в остальном совершенно различные, основываются на внимании к лексической частотности и на обработке частотного словаря (Burrows 2002; Шайкевич и др. 2013; Фаустов, Кретов 2017; Скребцова и др. 2021; и мн. др.). Здесь представляется еще один возможный подход, основанный на прослеживании частотностей в очень компактных тематических группах; предполагается, однако, что в силу описываемых закономерностей в распределении ключевых лексических единиц этого может оказаться достаточно для кластеризации идиостилей.

Исследование опирается на НКРЯ. Взяты авторы от ранней середины XIX до поздней середины XX века: Пушкин, Гоголь, Лермонтов, Гончаров, Тургенев, Писемский, Достоевский, Л.Толстой, Щедрин, Лесков, Чехов, Горький, Куприн, Бунин, Белый, А.Н.Толстой, Булгаков, Платонов, Набоков, Газданов, Леонов, Шолохов, Солженицын. Общий объем подкорпуса по 23 авторам — 12 489 650 словоупотреблений, для 20 авторов — более 300 000 у каждого. Для проверки использовался весь массив художественной прозы в корпусе.

Просмотрены ЛСГ глаголов речи; глаголов зрительного и слухового восприятия; глаголов памяти и забвения; а также простые и частотные наречия места и времени с местоименными корнями. Данные типы глаголов не обладают стилистической нейтральностью в общей перспективе стилей языка и типов текстов, но вполне тематически нейтральны (и при этом частотны) для фикциональной повествовательной прозы типа романа или рассказа.

ЛСГ устроены различным образом и это требует разного подхода. Группа глаголов речи практически не структурирована и не иерархизирована (не считая самой частотности), задана почти исключительно как равноправный неразмеченный список конкурирующих лексем (ср. единичные исключения как *отвечать — ответить*). Внутри групп глаголов восприятия и глаголов памяти, напротив, возможна и необходима группировка по большому ряду противопоставлений: собственно грамматические (вид: *слышать - услышать*, возвратность: *помнить — помниться*); б) словообразовательно-стилистические (*поглядеть — глянуть*, *видеть — видать*, *вспомнить — припомнить*); почти полная корневая синонимия (*смотреть — глядеть*); семантическая оппозиция внутри одного поля (*помнить — забыть*); параллельные и соотнесенные подгруппы внутри группы (зрительное — слуховое, пассивное — активное восприятие типа *видеть — смотреть*).

В каждой из рассмотренных групп обнаруживаются отдельные элементы и ряды-подгруппы, обладающие огромным дифференцирующим потенциалом. Доля *здесь* от *здесь+тут* варьирует от 8,5% (Шолохов) до 67,5% (А.Н.Толстой); доля *там* от *здесь+тут+там* от 24% (Пушкин) до 49,9% (Бунин); доля *потом* от *потом+затем* от 48,9% (А.Н.Толстой) до 100 (Пушкин); доля несов. в. от глаголов воспоминания от 15% (Леонов) до 64% (Достоевский); доля *припомнить+припоминать* от *вспомнить+вспоминать+припомнить+припоминать* от 0,5% (Бунин) до 46,2% (Щедрин); доля *глянуть* от *глянуть+поглядеть* от 0 (Пушкин, Лермонтов, Гончаров) или от 1,5% (Куприн) до 68,5% (Бунин); доля *смотреть* от *смотреть+глядеть* от 31% (Тургенев) до 86% (Солженицын) и т.д. В группе глаголов речи подобную роль может играть резкое понижение доли *сказать*, особенно в форме *сказал*, *сказала* (до 8% у Тургенева и 13-15% у Достоевского, Писемского, Лескова при 45% у Л.Толстого и 30-33% у Пушкина, Лермонтова, Гончарова, Чехова); соотношение *сказать* и *спросить* (от 0,98/1 у Тургенева до 7,74/1 у Л.Толстого); активность отдельных диагностических глаголов (461 *промолвил*,

промолвила в шести романах Тургенева, 2 случая в трех романах Гончарова, один случай в романах Л.Толстого, ни разу во включенных в НКРЯ повестях и рассказах Чехова).

Диахронические конфигурации: нейтральность (большинство описываемого материала); однонаправленная эволюция (от 0 форм типа *отвечал* в 1830-н до соотношения *отвечал* и *ответил* как 50/50 к концу XIX века); изменение направления эволюции (возрастание доли форм типа *видал*, *слыхал* почти в 5 раз с 1810-х по 1860-е и уменьшение в 4 раза с 1860-х по 1930-е).

В дальнейшем представляется необходимым еще раз уточнить параметры, приписать каждому из них свой удельный вес и свести воедино в рамках общего алгоритма автоматической кластеризации идиостилей, рассмотренных и новых.

Работа выполнена при поддержке СПбГУ, проект 95434615.

Список литературы

Скребцова Т.Г., Гребенников А.О., Шерстинова Т. Ю. Динамика лексического состава русской художественной прозы (на материале частотных словарей корпуса русских рассказов 1900-1930) // Компьютерная лингвистика и интеллектуальные технологии. 20 (2021). С. 646-659.

Фаустов А.А., Кретов А.А. Понятие маркемы и предварительные итоги маркемного анализа русской литературы // Вестник Воронежского государственного университета: Лингвистика и межкультурная коммуникация. 2017. 4. С. 16-31.

Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. Дистрибутивно-статистический анализ языка русской прозы 1850—1870-х гг. Т. 1. М., 2013.

Burrows, J. ‘Delta’: A measure of stylistic difference and a guide to likely authorship // Literary and Linguistic Computing. 2002. 17(3).P. 267–287.

Семантика образов, вызываемых инструментальной музыкой: экспериментальное исследование на материале русского языка

Дубасова Анжелика Витальевна, Санкт-Петербургский государственный университет, anzhalikad@gmail.com,
Бондаренко Мария Николаевна, Центр Музыкотерапии, г. Минск, Беларусь,
Витко Дарья Леонидовна, Белорусская государственная филармония, г. Минск, Беларусь

Воображение персонажей и ситуаций при прослушивании мелодий представляет собой один из аспектов взаимодействия с музыкой [1: 509]. Семантика таких воображаемых образов и нарративов изучалась в экспериментальных работах [2, 3, 4], в которых, в частности, было обнаружено, что слушатели, принадлежащие к одной культуре, часто формируют схожие или идентичные образы. Исследования, посвященные изучению связей между отдельными аккордами и словами [5, 6], также демонстрируют, что музыкальные композиции способны передавать семантическую информацию и что такая информация воспринимается слушателями сходным образом.

Наше исследование фокусируется на анализе ассоциаций, возникающих у слушателей при восприятии инструментальной музыки. В частности, нас интересовали следующие вопросы:

1) существует ли общность ассоциативных реакций у слушателей – носителей русского языка,

2) оказывает ли субъективная знакомость мелодии влияние на эту общность и на семантику ассоциаций; данное предположение было основано на представлении о том, что предыдущий опыт и количество предъявлений мелодии влияют на ее восприятие (см. обзор в [7: 2–3]).

Данные для исследования были получены в ходе первого этапа комплексного эксперимента по изучению восприятия музыки. На данном этапе в эксперименте приняли участие 100 испытуемых (мужчин и женщин в возрасте от 19 до 72 лет) как с музыкальным образованием, так и без него.

Испытуемые последовательно прослушивали семь произведений (в случае сонат – их первых частей) инструментальной музыки. После прослушивания каждого произведения во время паузы испытуемые оценивали его как знакомое или незнакомое и записывали возникшие ассоциации и их последовательности. В качестве стимулов использовались мажорные и минорные композиции разных эпох, широко известные и малоизвестные, включая два неопубликованных музыкальных произведения.

Для предварительного изучения семантики и проверки гипотезы использовался метод облака слов. Обработка данных и их визуализация осуществлялись с использованием языка программирования Python. Перед генерацией облака слов ассоциации были разделены на отдельные слова и приведены к начальной форме с помощью библиотек nltk и pymorphy3. Список стоп-слов был сформирован на базе nltk и затем доработан. Генерация облака слов осуществлялась с помощью библиотеки wordcloud. Для каждого музыкального произведения было создано три облака слов: общее (ассоциации на произведение в целом), «знакомая мелодия» (ЗМ) и «незнакомая мелодия» (НМ).

В таблице представлены некоторые результаты анализа данных. Для каждой из трех категорий указаны четыре наиболее часто встречающиеся ассоциации. Для ассоциаций общей категории отмечено, присутствуют ли они в категориях ЗМ и НМ (независимо от их частотности). Также указан процент испытуемых, которые классифицировали мелодию как ЗМ или НМ.

Наиболее частотные ассоциации и образы

Произведение	Наиболее частотные образы

	Общие			ЗМ	НМ
		есть в ЗМ	есть в НМ		
№ 1. Бах. Прелюдия до мажор ХТК 1	волна солнце море природа	+ + + +	+ + + +	природа лес рассвет волна	дождь волна вода море
		58%	42%		
№ 2. Бетховен. Лунная соната	человек жизнь ночь дождь	+ + + +	- - - +	человек ночь жизнь расставание	луна поиск девушка дождь
		95%	5%		
№ 3. Моцарт. Соната № 16	ребенок бегать бабочка лето	+ + + +	+ + + +	ребенок цветок бегать бал	ребенок бабочка день игра
		67%	33%		
№ 4. Шопен. Прелюдия до минор	человек жизнь гора расставание	+ + + +	+ + + +	человек путь туча гора	жизнь человек темный гора
		25%	75%		
№ 5. Чайковский. Неаполитанская песенка	танец детский танцевать бал	+ + + +	+ + + +	танец детский танцевать играть	танец танцевать дворец бал
		84%	16%		
№ 6. Витко. Аврора	море ветер лес солнце	+ + - -	+ + + +	море спокойный ветер волна	море лес солнце ветер
		11%	89%		
№ 7. Витко. Ветер в волосах	человек хороший море жизнь	+ - - -	+ + + +	рассвет красивый народ демонстрация	хороший человек море жизнь
		8%	92%		

Из полученных результатов следует, что ассоциации, возникающие при прослушивании как «знакомых», так и «незнакомых» мелодий, в большинстве случаев совпадают, различаясь лишь по частотности и некоторым деталям.

Уникальные ассоциации были зафиксированы для 2-й, 6-й и 7-й мелодий, однако этот результат, вероятнее всего, связан с тем, что 2-ую мелодию большинство испытуемых отметило как знакомую (95%), а 6-ую и 7-ую – как незнакомые (89% и 92%, соответственно), что привело к недостаточному количеству ассоциаций для противоположной группы.

Таким образом, предварительный анализ семантики образов показал, что слушатели воспринимают мелодии сходным образом. Данный результат на материале русского языка совпадает с результатами указанных выше исследований на материале других языков.

При этом не было обнаружено существенного влияния субъективной знакомости мелодии на семантику возникающих ассоциаций.

Ключевые слова: восприятие музыки, ассоциативный эксперимент, образы, семантика

Работа выполнена при поддержке СПбГУ, шифр проекта 94034584 и при поддержке Центра Музыкотерапии, г. Минск, Беларусь.

Список литературы

1. *McAuley, J. D.*, *Wong, P. C. M.,* Bellaiche, L., & *Margulis, E. H.* What Drives Narrative Engagement With Music? *// Music Perception.* 2021. 38 (5). P. 509–521.
2. *Margulis, E. H., Wong, P. C. M.,* Turnbull, C., Kubit, B. M., & *McAuley, J. D.* Narratives imagined in response to instrumental music reveal culture-bounded intersubjectivity // PNAS. 2022. 119 (4).
3. *McAuley, J. D., Wong, P. C. M., Mamidipaka, A., Phillips, N., & Margulis, E. H.* Do you hear what I hear? Perceived narrative constitutes a semantic dimension for music // Cognition. 2021. 212. 104712.
4. *Huovinen, E., & Kaila, A.-K.* The Semantics of Musical Topoi: An Empirical Approach // Music Perception. 2015. 33(2). P. 217-243.
5. *Koelsch, S. Kasper, Sammler, E. D., Schulze, K., Gunter, T. & Friederici, A. D*. Music, language and meaning: Brain signatures of semantic processing // Nature Neuroscience. 2004. 7. P. 302–307.
6. *Painter, J.G. & Koelsch, S.* Can out-of-context musical sounds convey meaning? An ERP study on the processing of meaning in music // Psychophysiology. 2011. 48. P. 645–655.
7. *Freitas, C., Manzato, E., Burini, A., Taylor, M. J., Lerch, J. P., & Anagnostou, E*. Neural Correlates of Familiarity in Music Listening: A Systematic Review and a Neuroimaging Meta-Analysis // Frontiers in neuroscience. 2018. 12. 686.

Дифтонги в текстах византийских грамматиков и их акцентуация в надписях разного рода

Евдокимова А.А., Институт языкознания РАН, arochka@gmail.com

Одним из значимых процессов в истории византийского греческого языка была монофтонгизация дифтонгов, которая согласно последним исследованиям историков греческого языка [*Andriotis 1995, Browning 1983, Holton et als 2020, Tonnet 2003]*, проходила в несколько этапов. Это было связано не только с диалектными особенностями разных регионов византийской империи, но и с процессом итацизма и его спецификой. Несмотря на то, что Мирамбель, изучавший процесс монофтонгизации дифтонга с V по I вв. до н.э., отмечал, что никакой зависимости от ударения не наблюдается [Тронский 1962, с. 115], возникала путанница при простановке ударения на дифтонге, связанная и с разными вариантами монофтонгизации. Один из вариантов монофтонгизации представлял собой консонантизацию конечного элемента дифтонга. О причинах этого процесса независимо друг от друга написали два ученых [Корш 1902, Лучиди 1950]. В таких случаях вполне уместным представлялось сохранение ударения на первом элементе дифтонга, как единственном оставшемся гласном. Другой же вариант, наиболее частотный, был реализован переходом дифтонга в гласный, обычно отличный от первого из них. Что приводило к разнице написания и произношения слова и в каком-то смысле при сохранении старого написания ставило под вопрос постановку ударения в старом стиле с опорой на нисходящий или восходящий характер дифтонга. Для решения этой дилеммы использовались три стратегии: ударение по традиции, по аналогии или по договоренности, в том числе поэтому труды грамматиков, описывающих просодию [см. схолии к Дионисию Фракийскому, *De Prosodia Catholica* Геродиана, эпитомы этого текста Феодосия Александрийского или Псевдо-Аркадия, сочинения Георгия Хировоска VIII-IX вв. или Мануила Мосхопула к. XIII — нач. XIV вв. и т.п.], изобилуют примерами, которые они часто копируют друг у друга. Таким образом, помимо разрыва произношения и написания, существовал еще и разрыв реального узуса языка и грамматических описаний просодии и фонологии, в которых византийские грамматики скорее старались сохранить описания предшественников, цитируя их и составляя различные компиляции из фрагментов их сочинений. Одним из таких частотных цитируемых фрагментов, связанных с дифтонгами, кроме их характеристики, было описание классической функции тремы (διαίρεσις) с примером "παῖς πάϊς".

Поскольку во времена античного койне кроме монофтонгизации дифтонгов параллельно с ней исчезла количественная разница между краткими и долгими гласными, то изменился характер ударения, который из музыкального перешел в динамическое, хотя в некоторых диалектах сохранялись элементы тонической системы (понтийский и ср. современный сербо-хорватский). Отчасти это привело к необходимости изменения александрийской системы акцентуации, которая маркировала движение тона в слове и поэтому каждый знак ударения указывал и на тип тона (акут - восходящий, гравис - нисходящий, циркумфлекс - соединение обоих). Византийская система акцентуации, в том виде, как мы ее знаем из традиционной греческой грамматики, содержала требования к месту ударения, отсюда появление у византийских грамматиков таких терминов, как параксютонон и пропараксютонон, к словам с ударением на предпоследнем и третьем слоге от конца слова соответственно. При этом, как отмечал И.И. Тронский в своем труде, посвященном ударению [Тронский 1962: 17], согласно александрийской системе акцентуации ударение ставилось на первый элемент дифтонга, а по византийской на второй. Таким образом, можно считать эту позицию знаковой при различении систем акцентуации.

Анализ употребления знаков акцентуации над дифтонгами в разных системах акцентуации проводился на материале корпуса византийских акцентуированных текстов

BGAT [Евдокимова 2023]. При этом учитывались палеографические особенности памятников, использование принятых знаков сокращений для дифтонгов и отдельно рассматривались позиции орфографических замен, когда дифтонг был написан вместо ожидаемой одиночной гласной.

В качестве иллюстрации приведем несколько примеров из византийских надписей Грузии: гравис над первой частью дифтонга "χὰιρε" [Kauhchischvili 1999, 165, №140o], а еще с придыханием "χ᾽άερε" и орфографической ошибкой [Kauhchischvili 1999, 164, №140f], при этом придыхание сдвинуто влево и попадает на согласную. В другой надписи этого комплекса на первой части дифтонга стоит акут в "θυγατάιρες", где дифтонг написан вместо ожидаемого "ε". А в первом слове этой надписи вместо дифтонга "αι" использован "ε": "πολλὲ", и гравис стоит над ним, никак не маркируя орфографическую ошибку [Kauhchischvili 1999, 165, №140g]. Как видно из приведенных примеров, для этого региона характерно маркирование первого элемента дифтонга, как в александрийской системе акцентуации. Более детальный анализ акцентуации в этой позиции в надписях на разных материалах из других регионов будет представлен в докладе.

В результате анализа материалов из корпуса BGAT было выявлено, что в ряде памятников существует тенденция иначе акцентуировать начальный дифтонг в сравнении с дифтонгами в других позициях. Это связано с тем фактом, что кроме знака ударения в начале слова над первой гласной ставился знак придыхания, а некоторые авторы надписей предпочитали разводить эти знаки визуально. Одним из вариантов такого разведения становилось последовательное написание знака придыхания над первым элементом дифтонга, а ударения над вторым. В случаях, когда дифтонг в этой позиции был начертан в виде принятой вертикальной лигатуры, то придыхание сдвигалось максимально влево и оказывалось перед ней, а знак ударения, наоборот сдвигался вправо и оказывался после лигатуры. Эта особенность часто сохранялась и в тех случаях, когда в других позициях дифтонга в слове отдавалось предпочтение длинным знакам акцентуации, и тогда они оказывались на обоих элементах дифтонга. При этом существовали тексты, в которых над другими гласными ударение ставилось в рамках византийской системы акцентуации или одной из систем со сдвигами вправо и/или влево, а дифтонг маркировался сдвигом ударения влево на первый элемент, что могло указывать на рефлексы александрийской системы акцентуации или на особенности спеллинга при самодиктовке писца. Как неоднократно указывал в личных беседах Б.Л. Фонкич, некоторые рукописи содержат знаки акцентуации проставленные другой рукой и иногда другого цвета чернилами, что позволило предполагать наличие отдельной категории писцов, которые проставляли знаки уже над написанным текстом. Наш палеографический анализ использования знаков акцентуации в надписях на разных материалах и на папирусах, показал, что и в этом типе памятников неоднократно встречались такие случаи. Вполне вероятно, что проставлявшие знаки писцы руководствовались правилами грамматики, изложенными в трактатах, по которым они в свое время учились, а также виденными ими образцами из других памятников. Сравнение надписей на миниатюрах рукописей и на мозайках в церквях Константинополя показало, что часто знаки и орфография совпадают, и одно могло служить образцом для другого, несмотря на разницу в материале памятника. Таким образом, используя данные акцентуации позиций с дифтонгами, можно получить контрольную выборку для дальнейшего машинного обучения нейронной сети, которая позволит прогнозировать возможные варианты использования разных систем акцентуации, как в памятниках из одного региона, так и на одном материале.

Список литературы

1. *Andriotis N. P. History of the Greek Language. Thessalonica, Greece: Institute of Neo-Hellenic Studies, 1995.*
2. *Browning R. Medieval and Modern Greek. Cambridge, United Kingdom: Cambridge University Press, 1983.*

3. *Horrocks G. Greek: A History of the Language and its Speakers. John Wiley and Sons, 2010.*
4. *Tonnet H. Histoire du grec moderne: la formation d'une langue. L'Asiathèque Langues du monde,* 2003.
5. *Holton D., Horrocks G., Janssen M., Lendari T., Manolessou Io, Toufexis N. The Cambridge Grammar of Medieval and Early Modern Greek. Cambridge University Press, 2020.*
6. Kauchtschischwili T. Korpus der griechischen Inschriften in Georgien. — Tbilissi: Logos, 1999. B.I.
7. Lucidi M. *L' origine del trisillabismo in Greco. Ricerche Linguistiche I, 1, 1950.*
8. Евдокимова А.А. Корпус византийских письменных памятников и методы его разметки // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог» (2023), серия 22, издательство МИИ *(*Москва*)*, том 22, с. 1071-1081 (https://www.dialog-21.ru/media/5873/evdokimovaaa063.pdf).
9. *Корш Ф.Е. Двогласные в греческом языке с физиологической точки зрения. // Русский филологический вестник XLVIII 1902, 281-348.*
10. Тронский И. М. Древнегреческое ударение. — М.; Л.: 1962.

The use of sentiment analysis tools in the study of the evaluation category in the genreof the anglophone analytical review of the foreign exchange market

Zakharova O.S., Saint Petersburg State University,e-mail: st111446@student.spbu.ru

The evaluation category is a linguistic universal that manifests itself at all levels of the language. It occupies an important place in human life, which gives rise to scientific discussions ofrelated issues that open up prospects for additional analysis. Evaluation or value judgment presupposes a value relationship between a subject and an object [Ivin 1970, Wolf 1985, Arutyunova 1988] and from the point of view of cognitive linguistics, it can be defined as the activity of mastering the value of an object, coupled with the process of cognition [Petukhova, Khomyakova 2020; Shutyomova 2022]. Evaluation is related to emotionality, the manifestation of which is called emotivity in the language [Wolf 1985]. V. I. Shakhovsky also draws attention to the fusion of emotivity and evaluation. Summarizing the experience of domestic [Simonov 1966, Dodonov 1978, Kulikov 1997] and foreign colleagues [Titchener 1916, Plutchik 1962, Izard 1977, Ekman 2004], he suggests a classification of emotions with the allocation of three groups based on the semantics of the emotive unit: vocabulary with the meaning of an emotional state (angry, furious, etc.), vocabulary with the meaning of an emotional attitude (to love, to like, etc.), vocabulary with the meaning of an emotional characteristic (baby, imp, etc.). According to the researcher emotions do not manifest themselves singly, but in clusters. The author identifies such basic clusters of emotions as joy, sadness, anger, and fear [Shakhovskij 2010, 2023].

Currently, computational linguistics is developing a direction for determining the emotional coloring (tonality) of an utterance, called sentiment analysis, which implies the use of a set of content analysis methods designed to automatically identify emotionally colored vocabulary in texts and emotional evaluation of authors in relation to the objects discussed in the text [Pang, Lee 2008; Bing 2012, 2015]. **The purpose** of this article is to compare the results of automatic text analysis using some sentiment analysis tools, such as Monkey Learn Sentiment Analyzer and AFINN, available online, VADER, requiring the use of programming language Python, as well asa chatbot with artificial intelligence Perplexity.ai.

The analysis is carried out as part of a study aimed at exploring ways of representation of the evaluation category in the genre of anglophone analytical review of the foreign exchange market. **The relevance** of the chosen topic is determined by the influence that analytical reviews of the foreign exchange market can have on the economic, political and social life of modern society. **The novelty** of the work is explained by the study of the peculiarities of the evaluation category representation in the discursive sphere of foreign exchange market as an important area of modern economics, as well as the use of modern methods of automatic text analysis for this purpose.

The subject of consideration in this article is made up by the means of linguistic representation of the evaluation category in this genre. **The** research **material** comprises the analytical reviews published by British and American analytical agencies (“Reuters”, “IFCMarkets”, “FXStreet”) and business media (“The Economist”, “The Daily Mail”, “Bloomberg”, “Forbes”) for the period of 2022–2024 years. The selection of evaluative contexts was carried out taking into account the criteria of their integrity and coherence, as well as the presence of an evaluative statement in them, and was carried out through directed stratification sampling, using such methods of achieving representativeness as classification and parameterization [Baranov 2007: 488–489]. In general, 995 texts with 403711 word usages were analyzed, which satisfies the requirement of representativeness of linguistic databases. From this array, 1376 evaluative contexts are identified, containing more than 12959 linguistic units representing different types of general and particular evaluation.

Parametric analysis was used in the processing of the material, including morphological,

definitional, contextual, stylistic, linguo-axiological, as well as communicative and pragmatic types of analysis. The analysis algorithm included the selection of eleven parameters (P): P1 "subject", P2 "object", P3 "part of speech", P4 "definition", P5 "contextual meaning", P6 "stylistic marking and expressiveness", P7 "intensification and de-intensification", P8 "type of generalevaluation", P9 "type of particular evaluation", P10 "type of evaluation by the presence of an emotional/rational component", P11 "strategy of speech influence".

Let's summarize the results of the analysis according to our algorithm using the example of an adjective *jumbo* in the following context: *The Fed delivers another **jumbo** rate rise, and it's far from done. As recently as the start of June investors and analysts believed that a **"jumbo"**interest-rate rise for the Federal Reserve meant half a percentage point. How quaint. After four straight increases of three-quarters of a percentage point – the latest on November 2nd – perceptions have changed. Indeed, a stock market rally in the two weeks before the announcement was rooted in the belief that the Fed may scale down to a half-point rate increase at its next meeting in December. What was once **jumbo** is now moderate* [The Economist 2022]. So: P1 "subject": collective, not explicitly expressed; P2 "object": the process of raising Fed rates; P3 "part of speech": adjective; P4 "definition": *extremely large* [CD]; P5 "contextual meaning": large, significant (corresponds to the dictionary meaning); P6 "stylistic marking and expressiveness": a metaphorically reinterpreted adjective that originally referred to a large and clumsy person, but then became known as the name of an elephant from the London Zoo [OD]; P7 "intensification and de-intensification": the function of the intensifier, since in the definition there is an adverb *extremely* with the meaning *to a very high degree* [OD]; P8 "type of general assessment": negative; P9 "type of particular evaluation": emotional, part of a subgroup of psychological evaluations related to the group of sensory [Arutyunova 1988]; P10 "type of evaluation based on the presence of an emotional/rational component": emotional; P11 "strategy of speech influence": introduction to a negatively evaluated context using a metaphor [Baranov 2007], labeling [Shelestyuk 2014, Troschenkova 2016].

Let's compare the results obtained with the data of sentiment analysis tools. Monkey Learn Sentiment Analyzer determined the overall evaluation of the fragment as negative (80.4% out of 100% according to the internal scale of the instrument), which corresponds to our data, while AFINN assigned a positive rating of +4 to the context on a scale from -5 to +5, which contradicts the results of our analysis. The VADER tool has established that the general evaluation belongs to the neutral zone (according to the internal gradation of the instrument, the indicator is 0.959 out of 1) with the presence of a certain proportion of a positive evaluation (0.041 out of 1), which also differs from our conclusions. As for the artificial intelligence model Perplexity.ai, it characterized the overall evaluation of the fragment as mostly neutral, slightly negative, explaining this conclusion by saying that it discusses the Fed's rate hike and how perceptions have changed over time, noting that the text does not express strong emotions or opinions, but rather represents the actual and analytical view on the topic. This result is generally consistent with our conclusion.

Thus, our algorithm allows for a comprehensive analysis of the evaluation units of the language and the context containing them, to identify general and particular types of evaluation, as well as to determine the components of its modal framework, such as the subject and object of evaluation. At the same time, it is possible to characterize the strategy of speech influence, in the implementation of which evaluatively colored lexical units participate. The considered tools of sentiment analysis, meanwhile, determined only the general type of evaluation, assigning it some value in accordance with their internal scale.

К вопросу о создании корпуса устной речи на материале интервью потомков итальянских переселенцев в Крыму

Иванова Екатерина Павловна, д.ф.н., профессор, e.ivanova@spbu.ru; Иванова Елизавета Дмитриевна, инженер-исследователь, elizaveta d.ivanova@spbu.ru; Самарина Марина Сергеевна, д.ф.н., доцент, m.samarina@spbu.ru (СПбГУ)

Полуостров Крым, на протяжении веков являющийся местом пересечения различных культур и народов, представляет собой уникальный пример мультикультурализма и многонациональности. Итальянская диаспора, исторически сложившаяся на полуострове, является ярким свидетельством этого явления. Изучение культурной памяти итальянских переселенцев, запечатленной в их устной речи, позволяет глубже понять как культурно-историческую, так и лингвистическую картину полуострова.

Итальянская диаспора Крыма, сформировавшаяся в результате миграционных потоков из южных регионов Италии в XIX-XX веках, в годы Великой Отечественной войны подверглась трагической судьбе - депортации из Керчи. Послевоенные годы стали периодом длительной борьбы за возвращение на родину и признание несправедливости произошедшего. В 2015 году, после долгого ожидания, был издан указ президента РФ о реабилитации депортированных итальянцев, что стало важным этапом в восстановлении исторической справедливости.

С этого момента главной целью крымско-итальянского сообщества стало сохранение и изучение культурной памяти этого романского народа на территории полуострова. В качестве материала исследования используется библиотека видеоинтервью представителей старшего поколения итальянской диаспоры в Керчи, собранная в рамках проекта, посвященного истории депортации итальянцев Крыма во времена Великой отечественной войны (italiani in Crimea – Memoria Italia) [1]**.** Видеофильмы снабжены транскрипцией, общий объем которой составляет около 300 страниц. Общий объем аудиовизуального материала – около 20 часов, что позволяет говорить о масштабности и ценности собранной информации. Отличительной чертой выбранного для исследования материала является тот факт, что представители итальянской диаспоры Крыма говорят на русском языке, однако в их речи встречаются вкрапления слов и имен, происходящих из итальянского языка.

В настоящее время можно назвать несколько проектов, связанных с созданием корпусов звучащей речи (например, проект «Рассказы о сновидениях и другие корпуса звучащей речи», в котором предлагается система дискурсивного аннотирования устной речи [2]; проекты СПбГУ – аннотированный корпус спонтанной русской речи CoRuss и корпус устной русской речи, предлагающий пользователю акустико-фонетическую транскрипцию [3], проект Высшей школы экономики – устные корпуса русских диалектов, русского языка в различных регионах и различных языков народов России [4].

Данное исследование предлагает создание корпуса устной речи итальянской диаспоры в Крыму, что станет уникальным проектом в рамках изучения культурно-исторического наследия полуострова. Создание подобного корпуса устной речи итальянских переселенцев в Крыму позволит не только задокументировать их личные истории, но и проанализировать процессы сохранения культурной идентичности и передачи традиций последующим поколениям.

В докладе объясняются принципы организации уникального материала и возможности его представления в электронном виде с целью создания электронного корпуса устной речи потомков итальянских переселенцев в Крыму. Корпус позволит не только систематизировать и сохранить ценный материал, но и открыть новые возможности для исследований, направленных на изучение языковых особенностей, диалектных черт, культурных традиций и исторических событий, запечатленных в устной речи представителей диаспоры.

В исследовании применяются описательный метод и многоуровневая разметка, предполагающая выделение иноязычных элементов.

Ключевые слова: корпус устной речи, культурная память, итальянская диаспора, Крым.

Исследование выполнено за счет гранта Российского научного фонда № 24-28-01635.

Электронные источники:

1. https://www.memorial-italia.it/progetti/diritti-umani/italiani-in-crimea/
2. https://spokencorpora.ru/
3. https://russpeech.spbu.ru
4. https://ilcl.hse.ru/corpora?ysclid=m0ms8t4hlu587047803.

Два стиля одного автора: стилеметрический анализ романа М. Варгаса Льосы «Тетушка Хулия и писака»

Борис Вадимович Ковалев, Санкт- Петербургский государственный университет,
b.v.kovalev@spbu.ru

Роман выдающегося перуанского писателя, нобелевского лауреата Марио Варгаса Льосы «Тетушка Хулия и писака» (1977) представляет особый интерес для анализа стилистических особенностей. Этот текст строго делится на две части: нечетные главы написаны от лица самого Варгаса Льосы и посвящены эпизодам из его биографии; четные представляют собой пересказ радиопьес, написанных персонажем по имени Педро Камачо – коллегой главного героя по радиостанции.

Литературный парадокс заключается в том, что читатель не сразу понимает, что четные главы не имеют отношения к реальности нечетных глав, а далее сюжеты радиопьес постепенно смешиваются. Варгас Льоса создает двойную квазиреалистическую конструкцию, призванную травестировать идею правдоподобного и реалистического нарратива на уровне соположения радиопьес Камачо и «автобиографической линии» в параллельных главах.

Стилистические различия параллельных глав изучались средствами традиционного литературоведческого и лингвистического анализа [Fuente González 2005; Oviedo 1983; Mudrovcic 1996]. Многократно исследовалась природа авторского «я» в романе (см., напр.,[Sandoval 2019]), изучалось два авторских голоса [Prieto 1983, Alonso 1991], также анализировалась языковая специфика части Педро Камачо [Andreu 1986]. Представляется полезным и актуальным оценить разницу стилей двух параллельных глав при помощи современных стилеметрических инструментов.

Для решения задач исследования мы обратились к Дельте Берроуза [Burrows 2002] – одному из наиболее надежных и популярных сегодня стилеметрических методов, основанном на средней абсолютной разности между z-оценками наиболее частотных слов в контрольном и атрибутируемом текстах. В рамках нашего исследования использовался частотный словарь объемом от 100 до 300 наиболее частотных слов. Все опыты выполнялись в пакете Stylo.

Для проведения стилеметрических экспериментов был составлен корпус, в которыйвошли все романы Марио Варгаса Льосы (18 единиц), кроме того, по два романа Г. ГарсиаМаркеса и К. Фуэнтеса как сопоставительные.

Было установлено, что на основе стилеметрического анализа романы М. Варгаса Льосы делятся на 4 устойчивые группы: ранние «тотальные» романы, пессимистические романы, документальные романы и легкие романы с изрядной долей эротического элемента. Роман «Тетушка Хулия и писака» атрибутируется группе документальных текстов.

На втором этапе мы разделили текст романа на две единицы корпуса: в первый файл (Vargas_Julia Varguitas) вошли «автобиографические» главы, во второй (Vargas_Camacho) – пересказы радиопьес. В итоге автобиографические главы присоединились к группе легко- эротических текстов, а радиопьесы – к ветви условно документальных текстов.

На третьем этапе мы создали новый корпус, в который отдельными единицами вошли главы исследуемого романа. Все эксперименты показали идентичный результат: автобиографические главы строго отделяются от глав, написанных от лица Камачо. Наконец, на четвертом этапе мы провели контрастивный анализ наиболее частотной лексики «Тетушки Хулии и писаки» при помощи функции opposite(stylo). Среди прочего, тексты Камачо от автобиографических отличает форма используемых глаголов (3 л. vs 1 л.)и лексические указания на высмеиваемые штампы (moral, espíritu, esposa, aguileña) и т.д., что не контринтуитивно и значительно

дополняет и обогащает выводы, полученные
средствами традиционного лингвостилистического анализа.

Таким образом, разница стилей двух параллельных частей подтверждается количественными методами.

Работа выполнена при поддержке СПбГУ,шифр проекта 94033710.

Список литературы

Alonso C. J. “La Tía Julia y El Escribidor”: The Writing Subject’s Fantasy of Empowerment //
PMLA, 1991, 106, no. 1, pp. 46–59.
Andreu A. C. Pedro Camacho: Prestidigitador Del Lenguaje // Modern Language Studies, 1986,16, no. 2, pp. 19–25.
Burrows J. Delta: a measure of stylistic difference and a guide to likely authorship // Literary andLinguistic Computing, 2002, 17, no. 3, pp. 267–287.
Fuente González M. "La tía Julia y el escribidor", de Vargas Llosa, como motivo de acercamientoal estudio de estilos // Tabanque: Revista pedagógica, 1995, no.10, pp.109–121.
Mudrovcic M. E. “La tía Julia y el Escribidor”: algunas lecciones prácticas en torno a la estéticade lo huachafo // INTI, 1996, no. 43/44, pp. 121–134.
Oviedo J.M. Mario Vargas Llosa: estudios críticos. Madrid: Alhambra, 1983. Pp. 200–208.
Prieto R. The Two Narrative Voices in Mario Vargas Llosa’s “Aunt Julia and the Scriptwriter” //Latin American Literary Review, 1983, vol. 11, no. 22, pp. 15–25.
Sandoval J. E. N. Yo autobiográfico y figura(s) de autor en “La tía Julia y el Escribidor” y “El pezen el agua” de Mario Vargas Llosa // Nueva Revista de Filología Hispánica, 2019, 67, no. 2, pp. 545–578.

Статистические характеристики стихотворных фрагментов литературного текста (на материале корпуса русского рассказа XX века)

Колпащикова Евгения Олеговна, Лаборатория языковой конвергенции НИУ ВШЭ в Санкт-Петербурге, jane.kolpashchikova@gmail.com

Исследование посвящено частоте появления и объёму стихотворных фрагментов в прозаических текстах. Материалом выступила выборка в 1000 текстов из Корпуса русского рассказа XX века, из которых по итогам применения моделей машинного обучения и ручной доработки результатов было извлечено 245 стихотворных фрагментов.

Стихотворными фрагментами считались не только цитаты из стихотворений, песен и других ритмизованных текстов, но и подобные авторские произведения, включенные в полотно рассказа. Иными словами, как стихотворный фрагмент был бы расценен и эпиграф из А.С. Пушкина, и песенка собственного сочинения одного из героев рассказа. С теоретической точки зрения это разделение основано на афористичном высказывании М.Л. Гаспарова о сути отличия поэзии и прозы: «стихи печатаются короткими неровными строчками» [Гаспаров 1993].

Основным целью исследования стало изучение характера взаимодействия стиха и прозы в рамках жанра малой прозы XX века: по результатам выявления и анализа фрагментов статистические выводы были сделаны как в статике, так и в динамике. Проблемный вопрос исследования — насколько характерны стихотворные включения для русской прозы в то время и зависит ли его использование от таких внешнетекстовых параметров, как год или декада написания.

Для извлечения фрагментов были использованы модели машинного обучения, признаками для обучения которых стали паттерны чередования ударных и безударных слогов. Предварительная автоматическая расстановка ударений была произведена с помощью пакета ru_accent_poet, разработанного для работы с русскоязычной поэзией.

Суммарная длина найденных 245 отрывков составила 1101 строку. Хотя бы по одному фрагменту встретилось в 136 рассказах 125 уникальных авторов.

Все фрагменты были разделены на три категории: стихотворения, песни и прочее; к третьей относилось всё, что сложно назвать стихотворением или песней, например, цитаты из религиозных источников, лозунги и ритмизованные народные мудрости. Стихотворений оказалось 69 (28% от общего числа фрагментов), песен — 146 (60%), «прочего» — 30 фрагментов (12%).

Среднее количество стихотворных фрагментов на рассказ составило 0.244 при стандартном отклонении 0.786. Большое значение стандартного отклонения свидетельствует о том, что данные неоднородны. Иными словами, количество фрагментов в одном тексте может значительно отличаться, а точнее, как показало исследование, может находиться в диапазоне от 0 до 8.

Что касается подвыборок по десятилетиям, то среднее количество фрагментов по декадам незначительно отличается от среднего по всей выборке и составляет 0.251. Статистически значимы относительно среднего по всей выборке с вероятностью 95% оказываются значения в 1930-х и 1940-х годах. В 1930-е годы все показатели количества стихотворных фрагментов оказываются рекордно высокими, а в 1940-е — рекордно низкими.

Ещё одной рассматриваемой характеристикой стихотворных фрагментов стала их длина в строках. Самой распространённой формой включения стихотворных фрагментов оказались катрены — 63 из 245, около четверти от общего числа. На втором месте идут двустишия — 53 фрагмента.

Наконец, был рассмотрен ещё один параметр стихотворных фрагментов — их суммарная длина в строках. Наиболее высокая концентрация строк в один год приходится

на 1902 год (89 строк), на втором месте — 1935 год (64 строки), на третьем — 1988 (53 строки).

К выводам исследования можно отнести неравномерность распределения стихотворных фрагментов по малой прозе XX века в динамике. Пиковыми с точки зрения количества стихотворных фрагментов и строк оказались 1930-е годы. В 1940-е, напротив, наблюдался значительный спад в количестве стихотворных цитат. Прокомментировать этот факт можно с опорой на исторические события этих периодов: на 1930-е годы приходится пик советской пропаганды, это время, когда молодое советское искусство показывает свою силу и свои перспективы. На 1940-е же приходится Великая Отечественная война и время восстановления после неё: логично предположить, что, несмотря на популярность военных стихотворений и песен постфактум, на тот момент лирика не всегда казалась уместной.

Список литературы:

Гаспаров М.Л. Русские стихи 1890-х-1925-го годов в комментариях. Москва: Высшая школа, 1993.

Документация ru_accent_poet. [Электронный ресурс] URL: https://pypi.org/project/ru-accent-poet/ (дата обращения: 25.04.2024)

Публикация подготовлена в результате проведения исследования по проекту «Текст как Big Data: методы и модели работы с большими текстовыми данными» в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 г.

Verse annotation in the emerging corpus of German catholic liturgical poetry

Mikhail Koryshev (m.koryshev@spbu.ru), St Petersburg State University

Nowadays, text corpora are a prerequisite for linguistic and literature research, with verse studies being no exception. However, whatever the target language, most corpora are collections of texts in contemporary (overwhelmingly literary) languages, considering the still existing challenges in processing texts in earlier languages. Initially missing digital text copies requiring text search, scanning, and recognition imposes extra limitations in building the collection of historical texts. Nevertheless, it is the historical language that can reveal the entire versatility of language, thus providing for full-fledged representative corpora.

Our project is focused on building a corpus of German catholic hymns. Considering that the corpus can be attributed to both historical text and poetry (verse) collection, the annotation shall satisfy both types of corpora to ensure exhaustive description. Our work integrates pioneering ground-breaking corpus tools tailored to the description of liturgical texts with existing principles used in the development of similar corpora (i.e. the Corpus of Russian Poetry incorporated in the Russian National Corpus [1, 2]). Notably, building a corpus of liturgical texts is not our ultimate goal in itself, but rather represents a leap forward in revealing the origin and emergence of the church hymn as a literary genre in 'reverse perspective'.

Our corpus is based on the Gotteslob (1975) hymnbook [3], which is the first combined prayer and hymnbook authorized by German-speaking Catholic dioceses. Currently, the corpus includes songs contained in the common section of the Gotteslob (1975) which are used by all dioceses. The overall size of the corpus is 27,000 words and 1,149 stanzas, representing 232 hymns split by seven time periods – before the 1500s, 1500 to 1599; 1600 to 1699; 1700 to 1799; 1800 to 1899; 1900 to 1950; 1950 to 1975. Haider et al. [4] consider tagging individual language units in poetry texts. The authors propose to rely on the poem's logical structure by and therefore label lines, stanzas, and the text itself. In our case, a similar approach is implemented whereby a hymn means an individual text unit. The implemented annotation strategy includes tags of the following levels: hymn characteristics and stanza characteristics. Texts also have additional meta labels with the following parameters: 1) timeframe; 2) year of hymn origin; 3) musically unaccompanied or accompanied by music composed in the same epoch as the hymn; 4) verse number; 5) year of verse origin (approximately, if available).

Currently, a more extensive meta annotation is under development to account for stanza parameters. As a result, such characteristics as meter, rhyme, stanza arrangement are to be incorporated. The next step is morphological annotation requiring further preparation efforts due to challenges in automated processing of diachronic texts which merits specific scholarly research.

Keywords: liturgical texts, Catholic hymnography, text corpus, German language

References:

1. Grishina E. A., Korchagin K. M., Plungyan V. A., Sichinava D. V. Poeticheskii korpus v ramkakh NKRYa: obshchaya struktura i perspektivy ispol'zovaniya [*Poetic corpus within the framework of the Russian National Corpus: general structure and prospects of use*] // Natsional'nyi korpus russkogo yazyka: 2006—2008. Novye rezul'taty i perspektivy [*Russian National Corpus. 2006 – 2008. New Results and Prospects*]. St Petersburg: Nestor-Istoriya, 2009. P. 71–113.
2. Natsional'nyi korpus russkogo yazyka. 2003—2023. [*Russian National Corpus. 2003 – 2023. [Online resource]. URL: ruscorpora.ru (Date of access: 30.09.2024)*].
3. Gotteslob (1975) Katholisches Gebet- und Gesangbuch. Ausgabe für das Bistum Trier. Trier: Paulinus Verlag. 1054 S.

4. Haider, Th., Eger, S., Kim, E., Klinger, R., and Menninghaus, W. (2020). PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1652–1663, Marseille, France. European Language Resources Association.

Коммуникативные импликатуры робота-компаньона в эмоциональном диалоге

А. А. Котов[1,2,3], А. А. Зинина[1,2], Н. А. Аринкин[1,2], А. А. Филатов[4], З. А. Носовец[1,2]
[1]НИЦ «Курчатовский институт», Москва
[2]МГЛУ, Москва
[3]РГГУ, Москва
[4]ООО «Яндекс.Технологии», Москва

В проекте создания робота-компаньона Ф-2 мы разрабатываем прикладную когнитивную модель, которая должна позволить роботу строить речевые импликатуры из смысла поступивших текстов и использовать их в мультимодальном диалоге, где робот объединяет речевую продукцию с жестами и мимикой. Импликатура – это производный смысл, который непосредственно не содержится в семантике высказывания, но может быть построен при восприятии высказывания адресатом: будем относить к импликатурам пресуппозиции, имплицитные смыслы и результаты естественно-языкового вывода. Существующие системы поддержания диалога обучаются на реальных речевых примерах, при производстве которых человек уже продемонстрировал способность к рассуждению. В отличие от подобных систем, мы стремимся создать модель, которая сама реализует модель рассуждений и конструирует импликатуры для поддержания диалога.

Для конструирования смысла текста робот использует парсер, получающий на вход текст на естественном языке и строящий семантическое представление для каждого предложения (Котов и др., 2021). Механизм коммуникативных импликатур моделируется инвентарём сценариев: единиц типа ‘если–то’ (посылка–следствие). Смысл каждого входящего предложения сравнивается с посылками всех имеющихся сценариев. Наиболее близкий (релевантный) сценарий активизируется и может вызывать на роботе речевую и поведенческую реакцию. Сценарии делятся на два типа: д-сценарии обеспечивают эмоциональную обработку, р-сценарии конструируют импликатуры (Котов, 2021). В нашей модели сейчас используются 367 д-сценариев и 4160 р-сценариев.

В пилотном эксперименте робот слушал высказывания пользователя, после чего с помощью р-сценария конструировал возможные связанные события и предъявлял их в речи. Испытуемые жаловались на то, что (а) эти суждения тривиальны и (б) неясно, «к чему» робот это говорит, какое развитие истории он предлагает. Чтобы преодолеть эти трудности, в данной работе мы представляем модификацию механизма сценариев, позволяющую роботу строить из смысла текста более далёкие импликатруы, а также отбирающую более релевантные выводы, приводящие к эмоциональной оценке.

Работа механизмов сценариев состоит в следующем. Импликатура (следствие), построенная одним из р-сценариев, может оказаться близкой к посылке другого р-сценария, который построит свою имликатуру (Sloman, Chrisley, 2003). Такая модель «рассуждения» может быть представлена в виде семантического графа – см. рисунок 1. В структуре рассуждения робот должен реагировать на релевантные смыслы. Это либо те смыслы, которые непосредственно совпали с д-сценариями (то есть получили эмоциональную оценку), либо те смыслы, импликатуры которых совпали с д-сценариями (то есть смысл получил эмоциональную оценку в ходе естественно-языкового вывода). Если входящий смысл никак не связался с д-сценарием, то смысл считается «не интересным роботу» и не должен вызывать активный коммуникативный отклик. Таким образом, в структуре графа нас будут интересовать пути, соединяющие конкретный смысл с ближайшим д-сценарием.

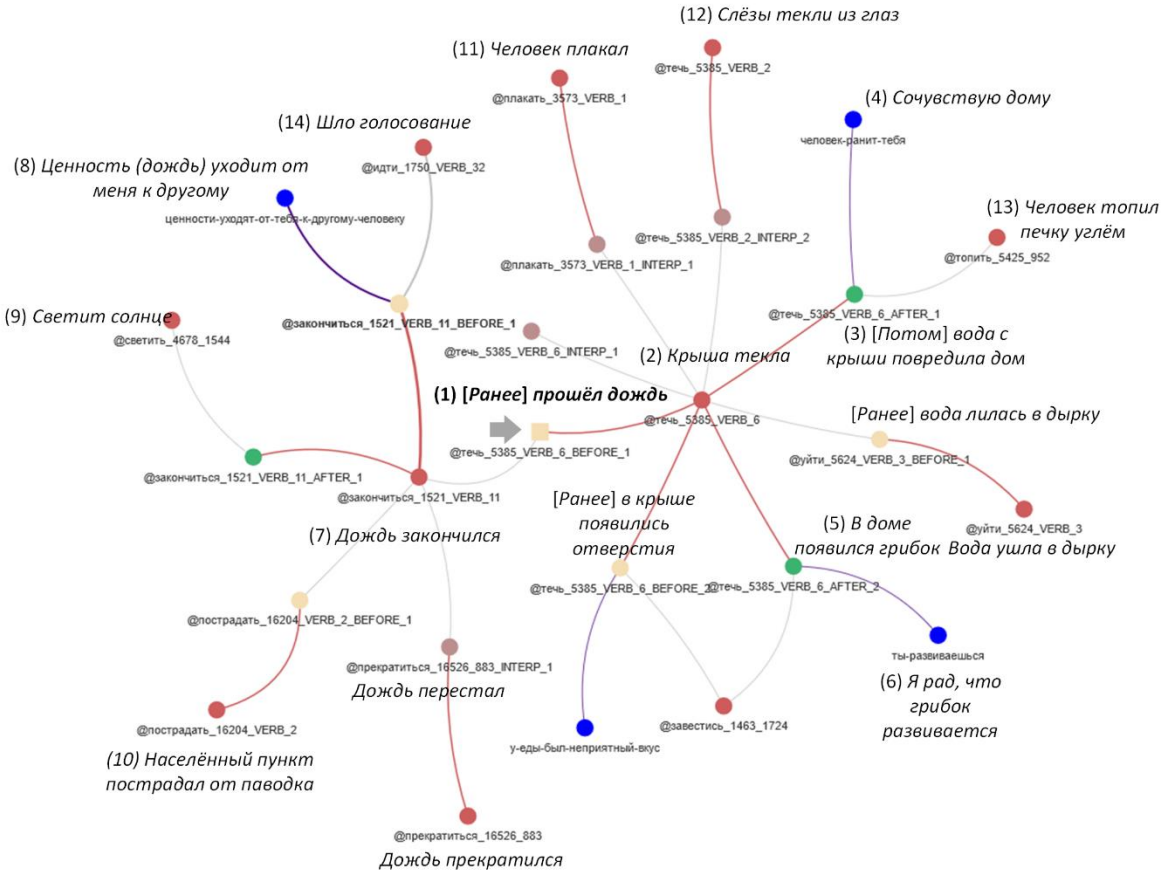


Рисунок 1 – Семантический граф, автоматически построенный от исходной репрезентации ‘дождь прошёл’ на глубину 3.

Высказывание *Идёт дождь* соответствует семантическому узлу (1) [*Ранее*] *прошёл дождь* (отмечен стрелкой) на рисунке 1. Этот узел является возможной пресуппозицией в сценарии (2) *Крыша текла.* Последующая импликатура (3) *Вода с крыши повредила дом* связывается д-сценарием (4) **человек-ранит-тебя** – в этом сценарии субъект (робот) «сочувствует» страдающему актанту, в данном случае – ‘дому’. Таким образом, событие ‘идёт дождь’ может вызывать у робота опасение того, что *крыша может течь* и *дом может пострадать*. Альтернативная импликатура для (2) – событие (5) *В доме появился грибок*. Поскольку у робота нет д-сценария, негативно оценивающего появление грибка, робот сопоставляет этот смысл с позитивным д-сценарием (6) **ты-развиваешься**: робот «радуется», что кто-то (в данном случае – ‘гибок’) растёт и развивается. Таким образом, событие ‘прошёл дождь’ через импликатуры связывается с позитивным следствием ‘в доме вырастет что-то классное’. Другая ветвь импликатур: если (1) *Дождь прошёл*, то (7) *Дождь закончился* или *дождь прошёл.* Этот смысл связывается с д-сценарием (8) **ценности-уходят-от-тебя**: робот воспринимает ‘уход дождя’ как потерю ценности, которая переходит к другому. Он может «завидовать», что другой человек наслаждается дождём, или расстраивается, что дождь больше не рядом.

Граф на рисунке 1 построен на глубину трёх узлов от исходного. Как видно, в этом графе есть импликатуры, которые могут оцениваться человеком как эмоциональные, но на данной глубине ещё не связались с д-сценариями. Например, это позитивное событие (9) *Светит солнце*, негативное событие (10) *Населённый пункт пострадал от паводка*, а также аналоги метафор (11) *Человек плакал*, (12) *Слезы текли из глаз*. Также на графе видна неправильная обработка омонимии в высказываниях (13) *Человек топил печку углём* (*топить* рассматривается как действие, связанное с водой) и (14) *Шло голосование* (ошибочно сопоставленное событию *шёл дождь*). Сейчас такая обработка вызывает

речевые ошибки робота, но потенциально может использоваться для компьютерного юмора.

Граф показывает, что для ситуации *идёт дождь* робот может переживать за дом, радоваться, что в доме вырастет что-то живое, расстраиваться, что ‘дождь куда-то ушёл’ и т. д. – что должно создавать более релевантные импликации для поддержания роботом мультимодального диалога с пользователем.

Исследование выполнено в рамках государственного задания Минобрнауки России НИОКТР - И124091100054-7.

Список литературы

Котов А. А., Аринкин Н. А., Зайдельман Л. Я., Зинина, А. А. Разработка средств семантического анализа текста для управления роботом // Вестник Военного Инновационного Технополиса “Эра,”. – 2021. – 2(2). – С. 115–120.

Котов А. А. Механизмы речевого воздействия. РГГУ, 2021.

Sloman A., Chrisley R. Virtual Machines and Consciousness // Journal of Consciousness Studies. – 2003. – 10(4–5). – Pp. 133–172.

Аудиовизуальные характеристики эмоциональных вариантов иронических высказываний

Кочеткова Ульяна Евгеньевна, СПбГУ, u.kochetkova@spbu.ru, Скрелин Павел Анатольевич, СПбГУ, p.skrelin@spbu.ru, Васильева Полина Евгеньевна, СПбГУ, st076593@student.spbu.ru

Мультимодальность естественной устной речи и потребность в разработке эффективных аудиовизуальных интерфейсов, учитывающих эмоциональный компонент, приводит к необходимости анализа участия различных информационных каналов в передаче оттенков значения и эмоциональной окраски. Особенно важным является комплексный подход к изучению вербальной иронии и сопутствующих ей паралингвистических маркеров, поскольку данная эмоционально-оценочная коннотация может передаваться как исключительно за счет звукового оформления, так и за счет жестов и мимики; последние могут дополнять и усиливать выражение иронического значения в речевом канале, а могут и компенсировать недостаточно яркое выражение иронии за счет фонетических средств.

В некоторых исследованиях используется методика проведения перцептивного анализа, включающего серии экспериментов со звуковыми стимулами, визуальными стимулами без звука и комплексными аудиовизуальными стимулами для оценки каждого из информационных каналов при передаче и восприятии иронического значения.

В настоящей работе используется схожая методика для выявления дополнительных вариантов эмоциональной окраски, которые воспринимаются слушателем и/или зрителем в иронических высказываниях. Дело в том, что и в русском, и в других языках ирония-отрицания, как правило, реализуется совместно с различной эмоциональной окраской, однако до сих пор комплексный акустический, паралингвистический и перцептивный анализ этих сочетаний на материале русского языка не проводился. Не были установлены и границы интонационных единиц, внутри которых выражается подобный кластер эмоциональных оттенков.

В качестве материала исследования были использованы записи 6 лекторов (3 мужчин и 3 женщин), входящие в корпус иронической речи, состоящий из фрагментов с одинаковым лексическим составом, включенных в иронические и неиронические контексты. Длина контекстов вместе с фрагментами составила 2-4 фразы. В корпус вошли 5 наборов контекстов, а также 4 связных текста с ироническими репликами, однако эти реплики, как правило, не имели омонимичных нейтральных аналогов. Аудиозапись осуществлялась одновременно с видеозаписью со скоростью 100 кадров в секунду.

Каждый перцептивный эксперимент включал 24 вопроса, поскольку перед участниками стояла сложная задача: определить, является ли фрагмент ироническим, а затем выбрать ту эмоциональную окраску, которую они слышат в данном фрагменте. Была предложена классификация из 6 эмоций: радость, грусть, страх, удивление, злость, отвращение. Участники могли воспользоваться и специальным полем для свободного ответа, если в списке эмоций не было нужного значения, либо если они хотели сообщить дополнительную информацию. Одни и те же фрагменты присутствовали во всех трех экспериментах: в аудио-формате, в видео-формате и в аудиовизуальном формате. В каждом из экспериментов приняли участие, как минимум, 30 респондентов.

Результаты перцептивного эксперимента были неожиданными: так, для одинаковых стимулов, представленных в разных модальностях, наблюдались прямо противоположные ответы: если в эксперименте со звуковыми файлами чаще всего отмечались негативные эмоции (злость, отвращение), то в эксперименте с видео-фрагментами без звука участники воспринимали, наоборот, положительные эмоции (например, радость) в тех же самых отрывках. Эксперимент с аудиовизуальным сигналом показал результаты, похожие на первый эксперимент: снова негативные эмоции стали преобладать в оценке респондентов, что говорит о не столь однозначном соотношении ролей звукового и визуального каналов

при передаче информации. В отличие от экспериментов по восприятию иронического значения, которые показали превалирование визуального канала, в данном случае, этот канал уступал звуковому.

Соотнесение данных перцептивного и акустического анализа показало, что реализация иронии в речи за счет уменьшения значения просодических параметров (интенсивности и мелодического диапазона) ассоциировалась с грустью и даже отвращением, тогда как увеличение мелодического диапазона, появление изломанного мелодического контура приводило к восприятию удивления и радости.

Сравнение акустических характеристик предъявленных фрагментов и контекстов, в которые эти фрагменты были включены, показало, что акустические характеристики только иронии-отрицания и сопутствующих ей эмоциональных вариантов, наблюдаются не только в целевом фрагменте, но и в окружающем его контексте.

Паралингвистический анализ с помощью программного обеспечения ELAN позволил выявить, что для определения той или иной сопутствующей эмоции важно не столько участие различных мимических мышц и жестикуляторов, сколько направление движения этих мышц. В связи с этим на следующем этапе исследования планируется синхронизация данных акустического анализа с данными автоматической обработки визуального сигнала с помощью нейросети, в частности, для определения траектории перемещения единиц движения (action units), соответствующих лицевым мыщцам.

Между метатекстом и метадискурсом: опыт создания базы данных коннекторов русского языка

Крюкова Анастасия Игоревна, МГУ им. М.В. Ломоносова / Институт Языкознания РАН,
nastyakryukova0077@gmail.com

База данных коннекторов русского языка (https://ruslinkers.github.io/linkers.html) включает описания единиц с различным грамматическим статусом (союзы, вводные слова и словосочетания, частицы и пр.), используемых для связи пропозиций [Инькова-Манзотти 2001: 17]. При ее создании одной из задач было описать функционирование включенных в базу единиц на разных уровнях языка (об уровневых моделях см. [Sweetser 1990, Tsunoda 2018]), в том числе – на метатекстовом уровне, что было затруднено ввиду существования в отечественной лингвистической традиции нескольких интерпретаций понятий *метатекст* и *метадискурс*. Таким образом, необходимо было предложить такие основания для выделения метатекстового уровня употребления коннекторов, чтобы предоставить пользователям наибольшее количество возможностей для изучения материала в рамках уровневого подхода и в сопоставлении с другими характеристиками, включенными в базу.

В базе описываются их фонетические, синтаксические, семантические, дискурсивные, узуальные и сочетаемостные свойства. На данный момент база содержит 696 единиц и 2330 строк (каждая строка – коннектор в новом значении), перечень которых был получен путем анализа существующей литературы и корпусного материала. Выделенные коннекторы были затем распределены по 19 семантическим зонам, внутри каждой из которых также могут быть выделены подзоны.

В англо- и русскоязычных работах выделяются узкий и широкий подходы к пониманию выступающих как наиболее общие терминов *метатекст* и *metadiscourse* соответственно. Поскольку в отечественную лингвистику понятие *метатекста* пришло из литературоведения [Бахтин 1979, 1986] и было введено в обиход в работе [Wierzbicka 1971], большинство русскоязычных работ, посвященных метатексту, написано на материале художественной литературы. В широком понимании метатекст включает в себя единицы разных размеров, которые адаптируют информацию для адресата, уточняют, поясняют, устанавливают внутритекстовые и межтекстовые связи, указывают на отношение автора к высказываемому ([Рябцева 1994; Шаймиев 1996; Бокарева 1999]). Метатекст в узком понимании объединяет различные средства выражения эксплицированной речевой рефлексии [Ляпон 1986]. Хотя в англоязычных работах схожим образом понимают метадискурс в широком смысле ([Hyland 2005]), изучается преимущественно академический дискурс, а метадискурсивные в узком смысле средства выполняют единственную функцию установления связей внутри текста посредством указания на его части, определения последовательности ([Mauranen 1993]). Это может быть связано с тем, что первые работы по метадискурсу были написаны исследователями риторики английского языка [Williams 1981; Vande Kopple 1985] и имели скорее прикладной, нежели теоретический характер.

Было решено предложить две разные классификации метатекстовых коннекторов, отражающие два существующих подхода к определению этого понятия. Выделяются собственно метатекстовые коннекторы и метадискурсивные коннекторы. К первым относятся коннекторы, способные присоединять метатекстовый комментарий, то есть имплицировать предикат глагола речи в главном предложении: *[Я говорю, что] он дурак, **если** так можно выразиться*. Метатекст в данном случае понимается узко. Важно подчеркнуть, что речь идет именно о свойстве коннектора, обеспечивающем ему способность функционировать на уровне, отличном от пропозиционального, сам же он может иметь иную семантику. Две семантические зоны – ПЕРЕФОРМУЛИРОВАНИЕ (*иными словами*) и ОГОВОРКА (*по крайней мере*) – метатекстовые ингерентно, для коннекторов,

относящихся к остальным семантическим зонам, было необходимо проводить корпусное исследование.

К метадискурсивным мы относ им коннекторы, семантика которых так или иначе относится к метатексту в широком понимании. В данном случае мы рассматриваем именно значение коннектора, а значит, коннекторы некоторых семантических зон (ИСТОЧНИК ИНФОРМАЦИИ, ПЕРЕФОРМУЛИРОВАНИЕ, КОМПОЗИЦИЯ ТЕКСТА, ОГОВОРКА, ВЕРИФИКАЦИЯ) метадискурсивны по умолчанию. Метадискурсивное употребление, особенно в сочетании с предикатами речи и интеллектуальной деятельности, развивают также коннекторы семантических зон МЕРЕОЛОГИЯ, ТАКСИС, СЛЕДСТВИЕ, ЗАМЕСТИТЕЛЬНЫЕ ОТНОШЕНИЯ, УСЛОВИЕ, СТЕПЕНЬ.

Проведенное исследование позволило нам сделать следующие выводы:

1) Семантические зоны могут быть упорядочены с точки зрения (не)возможности входящих в них коннекторов вводить метатекстовый комментарий. Некоторые отношения, к примеру, АДВЕРСАТИВНОСТЬ, встречаются в рамках речевой рефлексии чаще, чем ПРИЧИНА.
2) Метадискурсивный комментарий, особенно в сочетании с определенными предикатами, может вводиться коннекторами практически всех семантических зон.
3) Один и тот же коннектор в разных значениях имеет разную сочетаемость и функции, может употребляться на разных уровнях, что подтверждает необходимость анализировать такие единицы с учетом их семантики.
4) Коннекторы, относящиеся к семантическим (под)зонам ДИЗЪЮНКЦИЯ, КОНЪЮНКЦИЯ, ТАКСИС, СТЕПЕНЬ, не имеют метатекстового употребления, поскольку связывают единицы одного уровня.
5) Для ряда семантических зон, например, ЦЕЛЬ и УСЛОВИЕ, характерно, что метатекстовое и метадискурсивное употребление развивают только наиболее употребительные единицы (*чтобы* и *если*), остальные же не обладают описываемым свойством.
6) Вводимый некоторыми коннекторами метатекстовый комментарий имеет фиксированную, идиоматизированную форму, оступлений от которой обнаруживается крайне мало (*чтобы не быть голословным* vs. *чтобы меня не упрекнули в голословности*).

Исследование выполнено при поддержке гранта РНФ «Связь пропозициональных единиц в предложении и в тексте: семантика и пути грамматикализации» № 22-18-00528.

Список литературы

1. Бахтин М. М. Проблемы поэтики Достоевского. 4-е изд. М.: Советская Россия, 1979. 320 с.
2. Бахтин М. М. Эстетика словесного творчества. 2-е изд. М.: Искусство, 1986. 443 с.
3. Бокарева Ю. М. Коммуникативно-синтаксические средства адресации в прозе Н.В. Гоголя и место в них метатекстовых элементов: автореф. дисс. ... канд. филол. наук. СПб., 1999. 22 с.
4. Инькова-Манзотти О. Ю. Коннекторы противопоставления во французском и русском языках (сопоставительное исследование). М.: Информэлектро, 2001. 434 с.
5. Ляпон М. В. Смысловая структура сложного предложения и текст: К типологии внутритекстовых отношений. М.: Наука, 1986. 200 с.
6. Рябцева Н. К. Коммуникативный модус и метаречь // Логический анализ языка: Язык речевых действий: сборник статей / под ред. Н.Д. Арутюновой, Н.К. Рябцевой. М.: Наука, 1994. С. 82–92.

7. Шаймиев В. А. Метатекст и некоторые его признаки // Лингвистический семинар: Межвуз. сб. ст. Вып.1, Пб., 1996. С. 79–84.
8. Hyland K. Metadiscourse: Exploring Interaction in Writing. London & New York: Continuum, 2005. 296 p.
9. Mauranen A. Cultural Differences in Academic Rhetoric: A Textlinguistic Study. Frankfurt: Peter Lang, 1993. 280 p.
10. Sweetser E. From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge Univ. Press, 1990.
11. Tsunoda T. Levels in Clause Linkage: A Crosslinguistic Survey. Berlin, Boston: Mouton de Gruyter, 2018. 892 p.
12. Vande Kopple W. J. Some exploratory discourse on metadiscourse // College composition and communication. 1985. P. 82–93.
13. Wierzbicka A. Metatekst w tekście // O spójności textu / ed. M. R. Mayenova. Wroclaw, Warsawa, Kraków, Gdańsk, 1971. P. 105–121.
14. Williams J.M. Style: Ten Lessons in Clarity and Grace. New York: Harper Collins Publishers, 1981.

Contrastive corpus-based analysis to achieve covert translation: a case study of texts on overview of degree courses in the English and Russian languages

Лекомцева И.А., к.ф.н., доц. кафедры английской филологии и перевода Санкт-Петербургского государственного университета, i.lekomtseva@spbu.ru
Вьюнова Е.К., к.ф.н., доц. кафедры английской филологии и перевода Санкт-Петербургского государственного университета, e.vyunova@spbu.ru
Абдульманова А.Х., к.ф.н., и.о. зав. каф. английской филологии и перевода Санкт-Петербургского государственного университета, a.abdulmanova@spbu.ru

This paper has been growing steadily with the classroom versions of parts of it as teaching materials and papers. The main concern is how to achieve functional, communicative equivalence in translation. The days are long gone when translation was regarded mainly as purely a linguistic transformation. Today, there is a general trend towards functionally oriented approaches both in linguistics and translation studies. What was formerly regarded as cross-linguistic systemic phenomenon has increasingly come to be viewed as the use of this system in texts. In this regard, contrastive functional analysis has much to offer to translation studies, especially in terms of identifying and studying those linguistic means that can be regarded as, cross-linguistically and translationally, "unique items", as Andrew Chesterman put it, i.e. ones that are in some sense specific to the target language and are presumably not so easily triggered by source-language items that are formally different. These unique items, although tend to be under-represented in translations, contribute to more covert translation, in Julian House's terms. In this regard, what underpins our search for unique items are semantic categories, or, in more simple terms, the same meaningness, and their linguistic means of expression in different languages are very much the focus of attention. In other words, we are particularly interested in how the same meaning is expressed in different languages on the level of language use, rather than on the language system, particularly by language specific items.

The aim of the paper is, then, to identify functionally equivalent cross-linguistic correspondences expressing the same semantic categories in English and Russian. Any cross-linguistic comparison presupposes that the compared items are in some sense similar or comparable. The tertium comparationis in our case is a semantic category as invariant categories, i.e. semantic constants, that are expressed by cross-linguistic variants, both lexical and grammatical, i.e. alternative ways of rendering a particular meaning or function in different languages. Closely linked with our focus on the language use, rather than language system, is the fact that linguistic units are firmly placed both on paradigmatic and syntagmatic axis. This emphasis on "the company words keeps" (as J.R. Firth put it), or co-occurrence relations, can lead to a discovery of a wide range of word combinations and multi-word functionally complete units which vary in fixedness and idiomaticity in cross-linguistic level. These functionally complete units are functionally complete in a way that they have all the components that are necessary for them to function. In this regard, the starting point is corpus of texts matched in terms of text type, subject matter and communicative function. Such comparable corpus represents natural language use with the genre and are unaffected by various translation effects. As a result, these cross-linguistic functional correspondences, especially if they are unique items, can be used to achieve functional, communicative, natural equivalence in translation, or covert translation.

The methodology we use in the paper is the following:

1. First, we compiled a corpus of texts in English and Russian macheted in terms of the genre (official and science genre) and subject matter (texts describing degree courses/ programmes) by using the T-Lab software;

2. Second, we identified the main semantic categories in this subject matter by using the function of identifying the themes on the basis of the key words by these tools; the tertium comparisionis is the meaning and how it is rendered in the two languages;
3. Third, we identified the linguistic means expressing of these semantic categories (themes), both in their paradigmatic and syntagmatic dimensions, and matched them cross-linguistically;
4. Finally, we tested our hypothesis whether these functional cross-linguistic correspondences can be used as translation solutions to achieve functional, communicative equivalence in translation of texts describing degree programs.

Following this algorithm of contrastive corpus-based analysis, we identified the following linguistic items expressing the same meaning in English and Russian texts about academic programmes, or courses. All these cross-linguistic correspondences are functionally complete units and unique items in translation:

Все программы характеризует междисциплинарный подход / The structure of all these programmes encourages a cross-disciplinary approach.

Выпускники могут работать в преподавательской сфере, проводить научные исследования / Our graduates enter professions such as teaching or go on to do research.

В процессе обучения студенты изучают … / The degree spans such key topics as: … .

Основной особенностью программы является практико-ориентированный междисциплинарный подход к обучению / There is emphasis placed on a practice-oriented interdisciplinary approach within the course.

Фонетические характеристики речи при когнитивной нагрузке

Максимова Мария Романовна, СПбГУ, st076821@student.spbu.ru
Евдокимова Вера Вячеславовна, СПбГУ, v.evdokimova@spbu.ru

Когнитивная нагрузка – это нагрузка на рабочую память человека. Рабочая память обеспечивает временное хранение и обработку информации. Так как объем рабочей памяти ограничен, для эффективной умственной работы нужно контролировать когнитивную нагрузку [8].

В сфере образования выявление когнитивной нагрузки по речи позволяет адаптировать учебный процесс под возможности обучающихся. В области информационных технологий отслеживание когнитивной нагрузки делает голосовые помощники более удобными для использования, помогает корректировать работу системы под психофизиологическое состояние пользователя [8].

При когнитивной нагрузке увеличивается напряжение голосовых складок. Это приводит к повышению подсвязочного давления и, соответственно, к повышению частоты основного тона (ЧОТ) [5]. Когнитивная нагрузка проявляется в сужении диапазона ЧОТ [4, 5], повышении интенсивности голоса [5], изменении темпа речи и скорости артикуляции [4, 5, 6, 9].

Темп речи и скорость артикуляции могут повышаться или понижаться в зависимости от задач и условий их выполнения. Например, исследование К. Мюллера [6] показало, что при ограничении по времени скорость артикуляции повышается, а при выполнении дополнительной задачи - понижается. В эксперименте Т. Ф. Япа [9] было выявлено, что темп речи повышается при высокой когнитивной нагрузке. При низком и среднем уровне когнитивной нагрузки не наблюдается изменений темпа речи. Участники эксперимента Ф. Н. Ли [5] определяли среднюю когнитивную нагрузку по медленному, но неравномерному темпу речи. Показателем высокой когнитивной нагрузки был быстрый, но равномерный темп.

Целью данной работы является сравнение фонетических характеристик речи при наличии когнитивной нагрузки, при ее отсутствии и при предпаузальном удлинении. Предпаузальное удлинение – это увеличение длительности звуков, расположенных рядом с границей синтагмы или фразы [1].

Материалом для исследования послужили аудиозаписи речи семи студенток кафедры фонетики и методики преподавания иностранных языков СПбГУ. В качестве задачи информантам было предъявлено 20 названий цветов, где цвет шрифта отличался от значения слова (тест Струпа [7]). Испытуемые должны были назвать цвет шрифта. Дополнительной задачей были вопросы, связанные с фонетикой английского языка. Речь без когнитивной нагрузки была реализована информантами при чтении фонетически представительного текста «Был тихий серый вечер» [3].

Был проведен акустический анализ полученных записей в программе Praat. Было проведено сравнение длительностей безударных гласных, щелевых согласных и сонантов при когнитивной нагрузке, при отсутствии размышлений и при предпаузальном удлинении. Также были сравнены диапазоны ЧОТ в интонационных контурах при перечисленных условиях.

Результаты показали, что при когнитивной нагрузке увеличивается длительность гласных и согласных. Соответственно, скорость артикуляции понижается. Кроме того, понижается интенсивность голоса, становятся частотными восходящие и нисходяще-восходящие контуры ЧОТ, расширяется диапазон ЧОТ. Увеличение диапазона ЧОТ и понижение интенсивности в случае затруднений при выполнении задания противоречит результатам некоторых предшествующих работ. Следовательно, влияние когнитивной нагрузки на эти параметры нуждается в дополнительных исследованиях.

В докладе будет представлен мультимедийный корпус речи при когнитивной нагрузке. В нем содержатся видео- и аудиозаписи. На видеозаписях фиксируется мимика. Информанты проходят симулятор вождения в городе и одновременно отвечают на вопросы на общие знания (кругозор). Вопросы нужны для создания дополнительной нагрузки. С помощью этого корпуса можно будет обучать системы распознавания когнитивной нагрузки в речи, которые могут использоваться в навигаторах с голосовым управлением. При выявлении когнитивной нагрузки по речи водителя система будет выдавать уведомление о необходимости отдыха [2].

Список литературы

1. Качковская Т. В. Взаимодействие сегментных и просодических факторов, влияющих на степень и локализацию предпаузального удлинения в русском языке: дисс. ... канд. филол. наук. СПб., 2015.
2. Сапрыкин, Я. Д., Рязанцев В. И., Смирнов А. А. Обзор подходов к распознаванию усталости водителя и существующих технических решений // Известия МГТУ МАМИ 3 – 2020 – С. 48-58.
3. Степанова, С. Б. Фонетические свойства русской речи: реализация и транскрипция: Дис. ... канд. филол. наук. Л., 1988.
4. Berthold A., Jameson A. Interpreting Symptoms of Cognitive Load in Speech Input // UM99, User modeling: Proceedings of the seventh international conference, Springer, Vienna – 1999 – p. 235-244.
5. Le, P. N. The Use of Spectral Information in the Development of Novel Techniques for Speech-Based Cognitive Load Classification: PhD thesis. Sydney, 2012.
6. Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., Wittig, F. Recognizing time pressure and cognitive load on the basis of speech: An experimental study // International Conference on User Modeling, Berlin, Heidelberg – 2001 – p. 24–33.
7. Stroop J. R. Studies of interference in serial verbal reactions // Journal of Experimental Psychology, 18 (6) – 1953 – p. 643-662.
8. Sweller, J. Cognitive load during problem solving: Effects on learning // Cognitive Science, 12 (2) – 1988 – p. 257-285.
9. Yap T. F. Speech Production Under Cognitive Load: Effects and Classification: PhD thesis. Sydney, 2012.

Поэма Кутба «Хусрау уа Ширин»: словарь, транскрипция, перевод

Мамырбек Гульфар Мажитовна
Институт языкознания имени А.Байтурсынова
Казахстан, Алматы
gulfar76@mail.ru

Первое упоминание о поэме Кутба «Хусрау уа Ширин» было сделано французским тюркологом Жаном Дени (Jean Deny) в его труде 1921 года «Grammaire de la langue turque» («Грамматика турецкого языка»). С тех пор произведение Кутба привлекло внимание ученых и стало объектом исследований. По мнению исследователей, поэма «Хусрау уа Ширин» была написана в 1342 году в Золотой Орде, а в 1383 году ее переписал Берке Факих, выходец из племени кипчаков, в Египте. Известно, что этот экземпляр хранится в Национальной библиотеке Парижа и является основой для всех последующих исследований.

Впервые упомянутое в одной из частей «Шахнаме» Фирдоуси, это произведение позже было расширено Низами до полноценной поэмы. Персидская поэма Низами «Хусрау уа Ширин» вдохновила множество поэтов в иранской и тюркской литературе на создание новых версий под названием «Фархад уа Ширин» или «Хусрау уа Ширин». Основной темой всех версий произведения является любовь и привязанность. На сегодняшний день известно около двадцати иранских и пятнадцати тюркских поэтов, которые писали на эту тему. В иранской литературе известны такие поэты, как Низами (XII в.), Амир Хусрау (XIII в.), Шихабуддин Абдулла, Хатифи (XIV в.), Касими (XVI в.), Кевсери (XVI в.), Шапур (XVI в.), Асаф Хан (XVI в.), Шариф Каши (XVI в.), Рухул-Эмин (XVII в.), Машрики (XVII в.), Хинду (XVII в.), Ибрахим Едхем (XVII в.), Хизри (XVII в.), Нами (XVIII в.), Шихаб-и Туршизи (XVIII в.), Шуле (XVIII в.) и др. В тюркской литературе «Хусрау уа Ширин» создавали поэты: Кутб (XIV в.), Фахри (XIV в.), Шейхи (XV в.), Ахмед Ридван (XVI в.), Садри (Гайдар II) (XVI в.), Хаяти (XVI в.), Джалили (Абдулджелил) (XVI в.), Ефшанжи (XVI в.), Сахири (XVI в.), Имамзаде Ахмед (XVI в.), Халифа (XVI в.), Идрис-бек (Махви) (XVI в.), Фасих Ахмед Деде (XVII в.), Салим (XVIII в.), Мустафа Насыр (XIX в.), Самед Вургун (XX в.).

Таким образом, в тюркской литературе поэма «Хусрау уа Ширин» неоднократно переосмыслялась и переписывалась в традиции назира на протяжении XIV-XX веков, в течение семи веков.

В этом году в рамках грантового проекта «Теория и практика лексикографирования языка письменных памятников Золотой Орды» мы опубликовали словарь, полную транскрипцию и перевод поэмы Кутба. Рассмотрим структуру данного труда:

1. Словарь. В словарь включены все слова поэмы. Для объяснения отдельных слов были использованы работы М. Кашгари («Диуан лугат ат-турк», XI в.), В. В. Радлова («Опыт словаря тюркских наречий», 1888-1911, 4 тома), Л. З. Будагова («Сравнительный словарь турецко-татарских наречий», 1863), Э. И. Фазылова («Староузбекский язык. Хорезмские памятники XIV века», 1965), Древнетюркский словарь (1969), Э. В. Севортяна («Этимологический словарь тюркских языков», 1974).

2. Перевод. При выполнении перевода были учтены исследования ученых Э. Наджипа, А. Ибатова, М. Сабыра, А. Тагирджанова, Г. Алиева, Х. Миннегулова, Н. Хажиэминоглу, Ю. Демирчи, А. Керимулы, Т. Кыдыра.

3. Транскрипция. Процесс транскрибирования текста, записанного другим алфавитом, в систему второго алфавита имеет свои трудности, которые решаются с помощью транскрипции и транслитерации. Существует несколько видов транскрипции: фонетическая, фонематическая, международная и практическая. Для передачи текста поэмы использована практическая транскрипция. В ряде случаев транслитерация

предпочтительнее для сохранения рифмы, как в случае с арабскими и персидскими словами. Например, если передать слово «рәнж» как «рениш», то нарушится рифма с «гянж», поэтому в некоторых случаях использовалась транслитерация, а не транскрипция. Некоторые слова встречаются в поэме в нескольких вариантах, такие как агач/йыгач, болмасун/болмасу, булут/бұлұт и др.

4. Структура поэмы. Поэма Кутба «Хусрау уа Ширин» состоит из 233 страниц, 91 главы и 2364 строф (куплетов). Для удобства чтения транскрипции и перевода были пронумерованы страницы и строки.

Формульность немецкого языка как трудность в машинном переводе на русский язык (на примере нейронных сетей DeepL)

Манёрова Кристина Валерьевна
Санкт-Петербургский государственный университет
k.manerova@spbu.ru

В докладе представлены возможности перевода немецких формульных структур на русский язык с помощью машинного перевода, предоставляемого немецкой компанией DeepL – нового онлайн-переводчика на основе нейронных сетей. (https://www.deepl.com/ru/translator). Формульность структур понимается в докладе как свойство коллокаций и фразеологических словосочетаний разной степени идиоматичности, перевод которых представляет собой смысловую и стилистическую трудность в паре языков немецкий-русский переводчика DeepL.

Материалом исследования послужили 100 фразеологизмов немецкого языка разного структурного типа, имеющие формульный характер, с примерами, отобранными для перевода из Цифрового словаря немецкого языка (Digitales Wörterbuch der deutschen Sprache, dwds.de). Методы, используемые при анализе перевода формульных конструкций: сопоставительный метод, метод анализа переводческих ошибок по Вилару и соавторам (Vilar 2006).

Директор и основатель DeepL Ярослав Кутыловски, обладатель докторской степени по информатике. Компания возникла в 2016 г.: в рамках компании Linguee GmbH команда под руководством DeepL Ярослава Кутыловски начинает работу над первой версией DeepL Переводчика. Команда имеет возможность использовать массив данных системы контекстуального поиска по переводам Linguee. В существующие нейронные сети вносятся многочисленные усовершенствования, что позволяет добиться высочайшего качества перевода. Поддерживается перевод на 30 языках. Русский язык как поддерживаемый и (контролируемый) язык вводится в 2018 г.

На сайте компании заявлено: «Цель DeepL – использовать нейронные сети для того, чтобы расширить человеческие возможности, устранить языковые барьеры и сблизить культуры» (https://static.deepl.com/files/press/companyProfile_RU.pdf). Нейронный подход к МП (der neuronale MÜ-Ansatz (NMÜ)), используемый переводческими системами DeepL на основе нейронных сетей призван воспроизводить когнитивные модели человеческого мышления, чем отличается от используемых ранее подходов к машинному переводу. С помощью нейронного подхода к машинному переводу можно избежать грамматических ошибок, ошибок в порядке слов переводного текста, однако такие критерии как точность и адекватность, пропуски слов и не намеренная редукция в переводе, стилистические ошибки, ошибки сочетаемости, перевод формульных конструкций являются недоработками систем НМП.

Недоработки подобного рода можно отнести к качеству контролируемого языка (в нашем случае направления перевода это - немецкий язык). Контролируемый язык (КЯ) понимаем вслед за Зузанной Гепферих как "подсистемы естественных языков, словарный запас и допустимые грамматические конструкции которых представляют собой подмножество словарного запаса или возможных грамматических конструкций неконтролируемого естественного языка, из которого они получены" (Göpferich 2008). Подробное исследование немецкого языка как КЯ находим в монографии Шаймаи Марзук из Университета г. Майнц (Marzouk 2022).

Так, у идиомы в машинном переводе DeepL может переводиться только буквальное значение ее предшественника, т.е. словосочетания, которое послужило основой для возникновения фразеологизма (die 1. Lesart): nah am Wasser gebaut haben – 1. Построить дом у воды; 2.Быть плаксивым / плаксивой: *Sie hat nah am Wasser gebaut - Она построила дом недалеко от воды* (но не как «*Она - плаксивая женщина*»), адъективный фразеологизм

eine alte Schachtel - 1. Старая коробка, 2. Старая перечница (о пожилой, сварливой женщине) в предложении *Sie ist eine alte Schachtel* переводится как *Она- старая коробка, она – старый ящик* (но не как «*Она - старая перечница*»). В DeepL через опцию «Глоссарий» для фразеологизма можно eine alte Schachtel добавить верный вариант перевода *старая перечница* (04.07.2024 добавлено). Подобные примеры будет предложены и проанализированы в докладе.

Стилистический ошибки встречаются в переводе комплексных идиом: einen Narren an jmdm gefressen haben (любить кого-то, нравиться кому-л., полюбиться кому-л.) - *Die Oma hat einen Narren an ihrem Enkel gefressen - Бабушка увлеклась своим внуком* (но не *«Бабушка души не чает в ее внуке») / Am tüchtigen Mann seiner Nichte, Fritz Schacher, hatte Paul Schmidt einen Narren gefressen - Паулю Шмидту приглянулся трудолюбивый муж его племянницы, Фриц Шахер* (но не *"Паулю Шмидту полюбился трудолюбивый муж его племянницы, Фриц Шахер»).*

Библеизмы также переводятся в их прямом значении: *Er ist ein (einsamer) Rufer in der Wüste - Он - одинокий путник в пустыне* вместо *«Глас вопиющего в пустыне».*

В то же время, нейронная сеть переводчика может предвосхищать наличие компонентов формульного характера в немецких коллокациях и давать верный перевод: Massnahmen treffen (принять меры): E*r hat alle nötigen Massnahmen* и *Er hat alle nötigen Massnahmen* ***getroffen*** переводится одинаково: *Он принял все необходимые меры.*

Анализ машинного перевода формульных немецких конструкций с помощью нейронных сетей DeepL выявил следующие типы ошибок: подмена смысла выражения, буквальный перевод выражения, стилистические ошибки, неточное распознавание структуры формульной конструкции и коллокационной сочетаемости в языке оригинала, ошибочный подбор эквивалента в русском языке. Выявленные ошибки – свидетельство проявления языковой неоднозначности в современном машинном переводе.

Источники

DeepL https://www.deepl.com/ru/translator (дата обращения 4.07.2024)

Digitales Wörterbuch der deutschen Sprache https://www.dwds.de (дата обращения 4.07.2024)

Список литературы

1. Göpferich, Susanne. Textproduktion im Zeitalter der Globalisierung: Entwicklung einer Didaktik des Wissenstransfers. 3. Aufl. Tübingen: Stauffenburg Verlag. 2008. S. 366

2. Marzouk Shaimaa. Sprachkontrolle im Spiegel der Maschinellen Übersetzung: Untersuchung zur Wechselwirkung ausgewählter Regeln der Kontrollierten Sprache mit verschiedenen Ansätzen der Maschinellen Übersetzung (Translation and Multilingual Natural Language Processing 20). Berlin: Language Science Press., 2022 .S. 6

3. Vilar, David, Jia Xu, Luis Fernando D'Haro & Hermann Ney. 2006. Error analysis of machine translation output. In LREC-2006: Proceesings of the 5th international conference on language resources and evaluation, 697–702. Genoa: European Language Resources Association (ELRA). http : / / www . lrec - conf . org /proceedings/lrec2006/pdf/413_pdf.pdf.

Выражение «языка травмы» в русском языке: количественный и качественный анализ речи людей, переживших травмирующий опыт

Екатерина Евгеньевна Межорина, eemezhorina@edu.hse.ru
Виолетта Владимировна Наумоваvvnaumova_2@edu.hse.ru
Варвара Максимовна Каличева, vmkalicheva@edu.hse.ru
Татьяна Ивановна Попова, НИУ ВШЭ (Санкт-Петербург), tipopova@hse.ru

Доклад **посвящен** исследованию особенностей речи людей, переживших насилие, в частности, изучению влияния травмирующего опыта на лексический и просодический уровни языка. **Целью** настоящей работы является выявление и дальнейшая систематизация лингвистических коррелятов, которые бы отражали изменение языка вследствие травматического опыта. Из-за случившегося в организме человека могут происходить неконтролируемые процессы, при которых пережитый опыт все-таки проявляется, как правило, бессознательно, и его проявления могут быть как на уровне поведенческих реакций, так и на уровне языка [Уоллин 2024].

В основе анализа лежит **гипотеза** о существовании в речи людей, подвергшихся насилию, лингвистических коррелятов, которые формируют многоаспектное понятие «язык травмы». Материалом исследования послужили интервью с людьми, пережившими домашнее и/или сексуальное насилие. 10 интервью (общей продолжительностью 3 часа) с девушками, пережившими насилие. Объем текстовых расшифровок — 15 474 токена.

Разметка полученных расшифровок из пользовательского корпуса проводилась в программе ELAN, после чего были проанализированы используемые знаки аннотации, чтобы с помощью систему разметки измерить эмоциональную нагруженность повествования на разных временных участках расшифровки. Затем полученные данные были лемматизированы в программе Mystem, на основе чего были составлены частотные словари. Далее мы провели лексический анализ расшифровок в программе AntConc.

Среди аспектов, формирующих сам термин “language trauma”, встречаются, например, артикуляционные нарушения, такие как психогенная афония и афазия, бессознательное отклонение от семантической составляющей произносимого в сторону его эмоционального содержания, выраженное в чрезмерном употреблении экспрессий, современных жаргонизмов и концептуальных метафор [Mitchell 1998]. **Результаты**, характеризующие изменения на просодическом уровне языка, были представлены в ходе анализа полученных в программе ELAN расшифровок. Среди них можно выделить следующие:

1. Заминки, как и краткие паузы хезитации, часто встречаются прямо в середине синтагмы, а не только в начале и в конце, как в любом привычном контексте;
2. Можно обнаружить значительное количество вдохов и выдохов преимущественно в начале и в конце синтагм, однако в рамках всего расшифрованного в процессе исследования языкового материала количество вдохов значительно преобладает над количеством выдохов;
3. При произнесении некоторых фраз говорящие делают краткие и длинные паузы хезитации после каждого последующего слова, как бы прерывая единство синтагмы несколько раз, но в тоже время сохраняя ее просодическое свойство;
4. Количество кратких и длительных пауз хезитации значительно больше в начале и в конце речи каждого информанта. В середине повествования их количество сокращается;
5. Затяжки звука часто встречаются в середине синтагмы, увеличивая ее протяженность.
6. На первых минутах повествования среднестатистически вздохов почти в два раза больше, чем выдохов, а на последних минутах повествования — наоборот. Также было выявлено, что краткие паузы хезитаций намного чаще встречаются прямо в середине синтагм, что не характерно для длительных пауз хезитаций.

Результаты, отражающие изменения на лексическом уровне, были представлены в виде таблиц. Их анализ позволил сделать вывод, что в речи людей переживших насилие, можно обнаружить схожие по семантике тематические лексические группы, преимущественно подчеркивающие специфику социальной среды, в которой информанты находились в момент случившегося болезненного опыта, описывающие их внутренние переживания в момент говорения и направленные на поиск ответов на экзистенциальные вопросы, связанные с течением жизни и времени. Среди выделяющихся наблюдений:

1. Относительная частота 0,32% лексемы *папа* в сравнении с показателем 0,46% у слова *отец*;
2. Часто используются существительные с абстрактным значением *(жизнь, время, детство);*
3. Среди глаголов преобладают леммы, обозначающие внутреннее состояние человека *(бояться)* и характеризующие бытовую обстановку, окружающую большинство информантов в момент переживания ими негативного опыта, например, *пить* со значением пристрастия к алкоголю.

Таким образом, **результаты** исследования позволяют говорить о том, что лингвистические корреляты, которые могут свидетельствовать о формировании особого “языка травмы”, возможно обнаружить не только на уровне лексики. Лексический состав речи напрямую связан с темой и определяется ею, но исследования в области “языка травмы” говорят о том, что травма начинает определять язык, то есть определять то, как человек говорит.

Явления, выявленные на просодическом уровне, во многом характерны для эмоциональной спонтанной устной речи в целом, но эти особенности могут задать вектор для будущих исследований в контексте языковой травмы. Кроме того, методы современной компьютерной лингвистики позволяют разбирать структуру речи, опираясь на количественные данные, что в свою очередь способствует объективности выводов.

Использование интеллектуальных систем при переводе французских полифункциональных лексем на русский язык

Овсейчик Юлия Владимировна, Грищенко Ратмир Олегович;
Минский государственный лингвистический университет, Минск, Беларусь

Настоящее исследование посвящено установлению особенностей использования интеллектуальных систем при переводе французских полифункциональных лексем на русский язык. В качестве примера используется единица *alors*. Русскоязычными коррелятами исследуемой лексемы для адвербиального способа употребления являются *тогда/в то время/в ту пору/тогда-то*, для коннективного – *тогда/в таком случае/то*, для дискурсивного – *ну/ну как/ну так/ну и как/ну так что/ну и что/ну что же* (Новый французско-русский словарь).

Источником материала для исследования послужил французско-русский параллельный подкорпус Национального корпуса русского языка (далее – НКРЯ), общий объем которого составляет 7 631 430 словоупотреблений. Количество вхождений исследуемой единицы составило 5 102. В выборку включены контексты из аутентичных французских текстов и их переводы, выполненные профессиональными переводчиками, на русский язык. Случаи употребления лексемы *alors* в составе сложного союза *alors + que* ‘тогда как/не смотря на то, что/в то время как’, конструкций *Prép. + alors*, где *Prép.* – предлог, и *Conj. + alors*, где *Conj.* – союз, исключены из выборки. Эмпирический материал исследования составил 248 контекстов, в которых изучаемая единица демонстрирует адвербиальный способ употребления (43,55 %), дискурсивный (34,68 %) и коннективный (21,77 %), установленные на основании разработанного нами алгоритма определения функционально-дистрибутивных свойств единицы *alors*.

В НКРЯ совпадения перевода лексемы *alors* с русскоязычными коррелятами имеют место в 27,01 % от всех контекстов, наибольшее количество которых приходится на контексты с ее адвербиальным способом употребления (64,18 %). Несовпадения с русскоязычными коррелятами наблюдаются в 43,15 % контекстов. К опущению единицы *alors* профессиональные переводчики прибегают в 29,84 % контекстов, треть составляют контексты с ее адвербиальным способом употребления.

Каждый из отобранных контекстов обрабатывался в наиболее распространенных среди пользователей интеллектуальных системах: *PROMT.One*, *Google Переводчик*, *Яндекс Переводчик* и *Reverso*. Наибольшее количество совпадений с переводами, представленными в НКРЯ, наблюдается у интеллектуальной системы *Reverso* (23,79 %), далее следуют *PROMT.One* (22,18 %), *Яндекс Переводчик* (21,37 %) и *Google Переводчик* (20,56 %). Опущение изучаемой лексемы при переводе наблюдается в 2,83 % переводов, выполненных *Яндекс Переводчиком*, в 6,45 % – *Google Переводчиком*, 10,08 % – *PROMT.One*, 17,74 % – *Reverso*.

Наибольшее количество совпадений с русскоязычными коррелятами в переводах интеллектуальных систем приходится на контексты с адвербиальным способом употребления единицы *alors* (*PROMT.One* – 72,22 %, *Яндекс Переводчик* – 56,48 %, *Google Переводчик* – 51,85 % и *Reverso* – 37,04 % от всех контекстов с таким способом употребления), при этом при коннективном способе употребления исследуемой единицы совпадения наблюдаются в 64,81 % контекстов, переведенных *PROMT.One* в 44,44 % – *Яндекс Переводчиком*, в 27,78 % – *Reverso*, в 20,37 % – *Google Переводчиком*, при дискурсивном способе употребления – в 17,44 % контекстов, переведенных *PROMT.One,* в 12,79 % – *Reverso*, в 10,47 % – *Яндекс Переводчиком*, в 6,98 % – *Google Переводчиком*. Наибольшее количество случаев опущения исследуемой лексемы приходится на контексты с адвербиальным способом употребления (*Яндекс Переводчик* – 4,63 %, *PROMT.One* – 10,19 %, *Google Переводчик* – 11,11 %, *Reverso* – 32,4 %).

Полученные количественные данные свидетельствуют о высокой степени распознавания интеллектуальными системами адвербиального способа употребления единицы *alors*, о трудностях в определении ее коннективного и дискурсивного способов употребления, о низком проценте случаев опущения изучаемой полифункциональной единицы. Разработанный нами алгоритм определения функционально-дистрибутивных свойств единицы *alors*, переведенный в дальнейшем на язык программирования, призван улучшить разметку по способам употребления полифункциональных единиц, тем самым способствуя более точному переводу и совершенствованию интеллектуальных систем перевода.

Аффективное моделирование нарратива о страшном (на материале сборника рассказов «Темный карнавал» Р. Брэдбери)

Персидская Н.В., Национальный исследовательский университет «Высшая школа экономики», г. Санкт-Петербург, nvpersidskaya@edu.hse.ru

Тема страха в литературе является одной из центральных и часто попадающих во внимание исследователей, так как страх — фундаментальное человеческое чувство, которое может проявляться в различных формах и контекстах социального взаимодействия [1]. Особый интерес в этой связи представляет литература жанра ужасов и фантастики, в рамках которого писатели используют страх как прием для стирания границ между реальным и «ужасным» сверхъестественным. В настоящем исследовании рассматриваются аффективные особенности нарратива в рассказах Рэя Брэдбери — американского писателя и сценариста XX–XXI веков, считающегося одним из мастеров «страшного рассказа». Целью работы является сравнение плана лексического выражения эмоции «страх» с сюжетными особенностями произведений автора на материале сборника «Темный карнавал», относящегося к его раннему творчеству [2]. Кроме того, обсуждаются особенности стиля писателя и их воздействие на читательское восприятие.

В ходе исследования были получены эмоциональные арки [4] для всех 22 рассказов сборника (см. иллюстративные примеры на Рисунке 1). Тексты обрабатывались на языке оригинала, т. е. на английском. Для каждого предложения определялась его лексическая эмоциональность с применением метода SentiArt [3], реализующего подход на основе векторных моделей для определения семантической близости слов в тексте той или иной эмоции (в нашем случае – лейблу «страх»). Затем было проведено сопоставление эмоциональных пиков, обнаруживаемых на графиках, и ключевыми эпизодами в соответствующих рассказах.

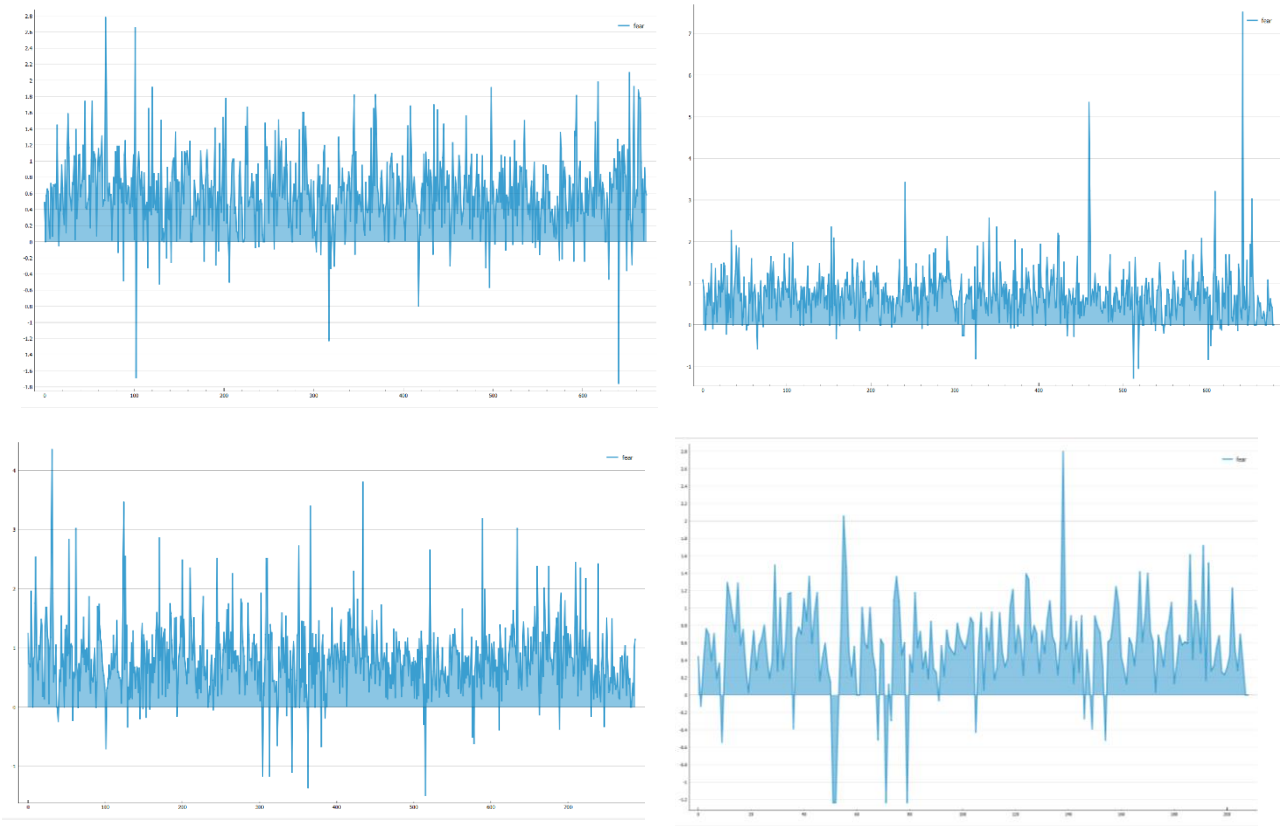


Рисунок 1. Распределение эмоции «страх» в рассказах «The Homecoming» (верхний ряд слева), «Skeleton» (верхний ряд справа), «The Small Assasin» (нижний ряд слева), «The Lake» (нижний ряд справа).

В результате компьютерного и сравнительного литературоведческого подходов к анализу удалось сделать ряд выводов о соответствии эмоционального и сюжетного, особенностях конструирования аффекта «страх» лексически и стилистически, а также сгруппировать рассказы сборника и используемые в них приемы следующим образом:

1. Эмоция страха ярко выражена и движет сюжет

В рассказах «Skeleton», «The Lake» и «The Crowd» наблюдается точное совпадение лексических эмоциональных пиков и поворотов сюжета. Отрывки текстов, оказывающие наибольшее влияние на читателя, состоят из наибольшего количества негативной эмоционально окрашенной лексики. Страх героев чаще всего связан с физической болью.

2. Страх выражен через несколько категорий эмоций

В некоторых рассказах на читателя воздействует не буквальный страх, но подкрепляющие его эмоции «злость» и «отвращение». Так, в «The Smiling People», «The Emissary» и в «The Cistern» соответствующий эффект создается через обильное использование лексики, выражающей отвращение, а в «The Tombstone» больше всего вызывает испуг злость призрака.

3. Эмоциональная лексика не воздействует на читательское восприятие

В рассказах лексика, связанная со страхом, может не совпадать с воздействием на читателя за счет особенностей ее подбора. Так, в некоторых фрагментах рассказов автор использует нейтрально окрашенную лексику, которая в контексте на самом деле оказывается конструирующей определенный эмоциональный фон. В этом, на наш взгляд, проявляются психологические приемы, используемые автором, и, соответственно, не обнаруживаемые при автоматическом анализе текста. Например, в рассказе «The Dead Man» читателя ужасают подробности течения жизни главного героя несмотря на то, что лексика нейтральна и не может быть ассоциирована со страхом вне контекста.

4. Уровень страха волнообразен и пересекается с сюжетной линией, но не имеет явного лексического выражения

К этой группе следует отнести все рассказы про семью Эллиотов («The Homecoming», «The Traveller», «Uncle Einar», «The Next in Line»). Автор погружает нас в параллельный мир и знакомит с обычаями сверхъестественных существ, мрачность для которых является нормой. За счет этого в рассказах этой группы также менее ощутимы контрасты между пугающим и обыденным. Постоянная мрачная атмосфера, наоборот, с меньшей силой воздействует на читателя, при этом эмоционально окрашенной лексики в рассказах микроцикла гораздо меньше, чем в других.

Таким образом, в раннем творчестве Р. Брэдбери наблюдаются такие способы эмоционального воздействия на читателя, как изображение буквального страха и конструирование страха через схожие негативные эмоции (имеют лексическое выражение), а также через пугающее, неприятное «послевкусие», достигаемого за счет детального описания событий, использование синтаксического параллелизма и приемов психологизма.

Публикация подготовлена в результате проведения исследования по проекту «Текст как Big Data: методы и модели работы с большими текстовыми данными» в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 г.

Список литературы

1. *Сагитова Л.Ф., Мухина Н.Б.* Эмоция «страх» как объект филологического исследования // Перспективы лингвистического знания: молодёжь и наука. 2020. С. 121-126.

2. *Bradbury R., Albright D*. Dark Carnival. – Sauk City, WI: Arkham House, 1947. [Электронный ресурс]. Режим доступа: https://www.ysk-books.com/en/show/book/dark-carnivalpdf (дата обращения 25.04.2024)

3. *Jacobs, Arthur M.* Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. Frontiers in Robotics and AI, 6, 2019. [Электронный

ресурс]. Режим доступа: https://doi.org/10.3389/frobt.2019.00053 (дата обращения 25.04.2024)
4. *Pretnar A.* Detecting Story Arcs With Orange. [Электронный ресурс]. Режим доступа: https://orangedatamining.com/blog/detecting-story-arcs-with-orange/ (дата обращения 25.04.2024)

Проект создания информационно-поисковой системыпо рукописям Пушкина

Перцов Николай Викторович, Сорокина Елизавета Дмитриевна, Эрлих Лев Исаакович
Институт Русского Языка имени В.В.Виноградова РАН, Москва

Задачи исследования: Представление рукописей Пушкина в удобном, обозримом и легко доступном виде с помощью современных компьютерных технологий и средств визуализации.

Методология: разработка и размещение в Интернете информационно-поисковой системы, содержащей как факсимиле рукописей, так и разные варианты их транскрипций. Предполагается разработать удобный пользовательский интерфейс, делающий наглядным соответствие между страницей рукописи и страницей транскрипции. Этот интерфейс позволит параллельно просматривать страницу факсимиле и страницу транскрипции с установлением точного геометрического соответствия между позиций в факсимиле и позицией на листе транскрипции, что особенно важно для рукописей, содержащих строки текста, расположенные в разных направлениях и содержащих многочисленные зачёркивания и вставки, располагающиеся часто между строк. Предполагается также составление средствами базы данных словарей-тезаурусов, содержащих все встречающиеся в рукописях начертания каждой словоформы, что обеспечит возможность быстрого поиска страниц рукописей, содержащих указанные пользователем словоформы. Будет также предоставлена возможность просмотреть все варианты написания каждой словоформы, встречающиеся в рукописях. Кроме того, будет разработана структура информационного описания страниц рукописи, включающая дату написания, названия произведений, фрагменты которых имеются на странице, дату первой прижизненной публикации, издание, в котором текст был впервые опубликован, и т.д., причём для каждого поля этой структуры будет автоматически формироваться словарь, содержащий все возможные значения, встречающиеся в этом поле в разных единицах хранения. Такие словари облегчат формирование запросов на поиск нужных страниц в системе и сбор статистики (для каждого значения будет указана его частота; кроме того, легко будет выбрать все страницы, содержащие это значение или совокупность значений).

Предполагаемые результаты исследования: каталогизация всех разночтений, связанных с трудночитаемыми местами рукописей, допускающими разные варианты прочтения. Возможное уточнение редакций текстов, не публиковавшихся при жизни Пушкина.

Критический обзор литературы: Нам не известны отечественные работы схожей тематики. Однако существует ряд зарубежных разработок сходной направленности (проекты по рукописям Бентама https://blogs.ucl.ac.uk/transcribe-bentham/, по рукописям Линкольна https://papersofabrahamlincoln.org/, Annotated Books Online https://abo.annotatedbooksonline.com/, Bess of Hardwick's letters https://www.bessofhardwick.org/, Colonial Despatches: The Colonial despatches of Vancouver Island and British Columbia 1846-1871 https://bcgenesis.uvic.ca/index.html, DIY History https://diyhistory.lib.uiowa.edu/, Manuscripts Online https://www.manuscriptsonline.org/, Electronic Beowulf https://ebeowulf.uky.edu/, Goethe und Schiller Archiv https://ores.klassik-stiftung.de/ords/f?p=401:1), но их задачи были несколько иными, кроме того, они еще не решают проблем, поставленных нами.

Выводы: Намечены принципы разработки пользовательских интерфейсов для быстрой навигации по массиву рукописей и параллельного представления транскрипций вместе с листами факсимиле.

Динамика поэтического перевода / переперевода в библиометрической перспективе: Ф. Петрарка и Ф. Гарсиа Лорка в России

В. С. Полилова (МГУ) vera.polilova@gmail.com, Б. В. Орехов (НИУ ВШЭ), nevmenandr@gmail.com, А.С. Белоусова (МГУ), nastassja.belousova@gmail.com, О. В. Полосина (независимый исследователь)

И Данте со своей Петраркой
И Рилька с Лоркою своей
Небесной триумфальной аркой
Мерцают из страны теней
И русскоземный соловей
Когда пытается расслышать —
Молчит и слышит только медь
Да что он может там расслышать?!
И он предпочитает петь
Д. А. Пригов

Переперевод (фр. retraduction, англ. retranslation) — это перевод текста, который ранее уже был переведен на тот же язык (Gambier 1994: 413). Этот феномен начал привлекать внимание исследователей с 80-х и 90-х годов прошлого столетия (Левин 1981; Berman 1990; Gambier 1994) и в последние десятилетия стал одним из центральных объектов изучения в теории и истории перевода. Развитие количественных и цифровых подходов значительно преобразило методы изучения динамики переводной литературы и переперевода. В частности, большой потенциал продемонстрировал количественный анализ на основе библиографических данных (Cheesman et al. 2017; Berk Albachten & Tahir Gürçağlar 2019).

Его успешно применяли при анализе динамики перевода и переперевода крупных прозаических и драматических произведений, но при изучении перевода поэтических текстов возникает практическая сложность использования библиометрического подхода. Она заключается в том, что международные библиографические индексы обычно не содержат детализированных сведений и включают лишь описания целых изданий. Например, книга избранных произведений поэта или переводы одного переводчика будут иметь одну библиографическую запись. Этот факт объясняет отсутствие библиометрических работ, исследующих переводы лирики.

В представляемой работе мы сумели применить количественный библиометрический подход к материалу русской переводной поэзии, использовав в качестве источника данных уникальные библиографические указатели, составленные советскими библиографами. Среди подобных указателей переводов выдающихся иноязычных авторов в качестве пробного материала мы выбрали два: указатель, посвященный советским переводам Федерико Гарсиа Лорки, составленный Э. В. Брагинской (1971; содержит информацию о переводах текстов Лорки, опубликованных в книгах и журналах в период с 1936 по 1969 г.), и указатель русских переводов Ф. Петрарки, подготовленный В. Т. Данченко (1986; содержит информацию о переводах и материалах, опубликованных в 1598, 1762–1985 гг.). Превратив эти указатели в базу данных, мы получили инструмент, позволивший нам впервые ответить на ряд важных вопросов о переводе и переперево де европейской поэзии в России.

В нашем исследовании мы ограничились рассмотрением переводов италоязычной лирики Петрарки и лирических стихотворений Гарсиа Лорки. В результате анализа библиографических данных была определена динамика переводов по годам и ключевые

этапы рецепции творчества двух поэтов в России (Петрарка: 1890–1915 и с 1950-х; Гарсиа Лорка: 1939–1946 и с 1956 г. и далее), выявлен круг самых активных и влиятельных переводчиков, выделены литературные поколения, определившие специфику их восприятия.

Рассматривая динамику перевода, переперевода и повторных публикаций русских переводов Петрарки и Лорки, мы описали, как складывались “русский Петрарка” и "русский Лорка", как ранние переводы заменялись новыми и как сформировалась группа классических переводов. Кроме того, мы определили самые переводимые на русский язык тексты Петрарки и Гарсиа Лорки и описали динамику их повторных переводов и публикаций, выявив количественными методами те их произведения, которые стали каноническими в русской культуре.

Использованный библиометрический метод подтвердил предыдущие критические наблюдения о восприятии Петрарки и Гарсиа Лорки в России, но также выявил множество особенностей процесса освоения их творчества, которые оставались вне поля зрения историков литературы и критиков. Мы сумели продемонстрировать применимость библиометрического метода для изучения динамики перевода и переперевода лирики и их механизмов. В будущем необходимо будет сравнить полученные данные с тем, как складывался процесс рецепции этих поэтов в других культурах. Еще одной актуальной задачей является сбор сопоставимых данных о переводах других иностранных поэтов в России.

Работа выполнена в НИУ ВШЭ при поддержке Российского научного фонда (проект № 23-28-01201).

Список литературы

BERK ALBACHTEN, Ö., TAHIR GÜRÇAĞLAR, Ş. (2019). “The Making and Reading of a Bibliography of Retranslations”. En Ö. Berk Albachten, Ş. Tahir Gürçağlar (ed.), *Perspectives on Retranslation: Ideology, Paratexts, Methods* (pp. 212-230). Routledge. S.l.

BERMAN, A. (1990). “La Retraduction comme espace de traduction”, *Palimpsestes 13*(4), pp. 1-7.

CHEESMAN, T., FLANAGAN, K., THIEL, S., RYBICKI, J., LARAMEE, R. S., HOPE, J., ROOS, A. (2017). “Multi-Retranslation Corpora: Visibility, variation, value, and virtue”. *Digital Scholarship in the Humanities, 32*(4), pp. 739-760. https://doi.org/10.1093/llc/fqw027

GAMBIER, Y. (1994). “La retraduction, retour et détour”. *Meta*, *39*(3), pp. 413-417. https://doi.org/10.7202/002799ar

БРАГИНСКАЯ, Э. В. (Сост.) (1971). *Федерико Гарсиа Лорка: Библиографический указатель*. Москва: Книга.

ДАНЧЕНКО, В. Т. (Сост.) (1986). *Франческо Петрарка: Библиографический указатель русских переводов и критической литературы на русском языке*. Москва.: Книга.

ЛЕВИН, Ю. Д. (1981). К вопросу о переводной множественности. В *Классическое наследие и современность* (с. 365-372). Ленинград: Наука.

Дискурсивные формулы Прагматикона VS речевые формулы корпуса ОРД

Наталья Викторовна Богданова-Бегларян, Татьяна Ивановна Попова n.bogdanova@spbu.ru, t.i.popova@spbu.ru
Санкт-Петербургский государственный университет

Интерес к устойчивым единицам устного дискурса в лингвистике неизменно высок и уже привел к появлению не только большого числа конкретных исследований, но и к созданию специализированных баз данных, главные из которых – Русский Конструктикон и Прагматикон. Объектом фиксации в этих базах являются не слова, а более крупные единицы, находящиеся на границе между словарем и грамматикой и фактически стирающие эту границу.

В центре внимания в настоящей работе находятся дискурсивные формулы (ДФ), которые собраны в Прагматиконе и рассматриваются как одна из разновидностей конструкций – ответные реплики в диалоге, ставшие устойчивыми формулами (*Ничего себе! Да ты что?! И не говори!* и под.) (авторы называют их *прагматическими конструкциями*). Источником материала для Прагматикона стали драматические произведения, богатые такого рода единицами (Рахилина и др. 2021). Очевидно, однако, что такие ДФ представляют собой лишь квазиспонтанную речь, имитацию реальной разговорной, придуманную и продуманную авторами пьес, мастерами художественного слова. Это и привело к мысли сопоставить словник Прагматикона, созданный в рамках кодифицированного литературного языка и собравший, условно говоря, то, *что может быть*, – с речью действительно реальной, по-настоящему спонтанной, реализованной в повседневном общении на русском языке – так же условно говоря, то, *что есть на самом деле*. Именно такую речь фиксирует звуковой корпус «Один речевой день» (ОРД) (Богданова-Бегларян и др. 2019).

В ходе первичного ручного аннотирования подвыборки в 300 тыс. словоупотреблений (195 речевых эпизодов) из корпуса ОРД было отобрано 1088 устойчивых неоднословных единиц (УНЕ) разного типа: от форм-идиом (ИД) (*по барабану, не вопрос, без проблем, в принципе*) до конструкций (КС) (<X Y-ка знает>,<X в том, что>) и прецедентных текстов (ПТ) (*это я удачно зашёл; алё, гараж!*). ТипологияУНЕ включает в себя 8 классов (Bogdanova-Beglarian et al. 2024), в том числе речевые формулы (РФ), в наибольшей степени близкие к ДФ Прагматикона. Показалось интереснымсопоставить эти два словника (единицы спонтанной и квазиспонтанной речи) и установить,как именно соотносятся РФ и ДФ, включая их функции, или речевые акты (РА), которые они формируют.

После проведенной коррекции результатов аннотирования в словнике УНЕ осталось 928 единиц; в таблице показано соотношение выявленных классов и место РФ среди них.

Таблица Количественное соотношение разных классов УНЕ в аннотированном подкорпусе ОРД

№№	**Класс УНЕ**	**Абс. кол-во**	**Отн. кол-во (%)**
1	ФК	297	32,0
2	НК	248	26,7
3	**РФ**	**122**	**13,1**
4	ИД	89	9,6
5	КС	82	8,8
6	ПМ	64	6,9
7	ПТ	20	2,2
8	ОК	6	0,6
Всего		**928**	100,00

Словник РФ включает 122 единицы; этот класс имеет ранг 3 в общем списке УНЕ (13,1 %) гапакс (количество единиц, которые встретились в материале всего один раз) – 67 (55 %). Этот список РФ был соотнесен со словником ДФ Прагматикона, включающим 607 единиц. Оказалось, что 43 РФ из 122 (35,2 %) входят и в состав ДФ, т. е. совпадают в обоих списках. Иными словами, две трети РФ не попали в словник Прагматикона, хотя активно функционируют в нашей повседневной речи. Из 13 основных РА, зафиксированных в Прагматиконе для ДФ, 10 обнаружились и среди контекстов с РФ. Не встретилось РА вопроса, уговоров и предложения/совета/просьбы. Расширение материала из корпуса ОРД наверняка выявит и эти функции.

На первом месте в обоих словниках находится формула *ничего себе*(31 употребление в ОРД), Прагматикон приписывает ей РА удивления. Именно эта функция чаще других характеризует формулы, общие для двух словников (10 случаев из 43 РФ), ср.:

- [Ж1] *сколько ? полчаса стоим уже / да ? даже больше ??* [И60] ***ничего себе*** (ОРД);
- [И72] *притом трактор сначала сам начал у... увя... / *П ну в общем это былоужасно //* [Ж1] @ ***ничего себе*** (ОРД);
- [И65] *один входной билет //* # [Ж2] *ва(:)у /* ***ничё себе*** (ОРД)**.**

В основу типологии УНЕ положен ряд принципов: грамматический (ИД), семантический (фразеологизированные коллокации, ФК) и прагматический (РФи прагматические маркеры, ПМ). Это позволило точнее фиксировать специфику выявленных единиц, поэтому часть ДФ, которые в принципе совпали с УНЕ, находится не только в классе РФ. Так, 16 ДФ в предложенной типологии УНЕ имеют другую трактовку:

- 6 отнесены к классу КС, поскольку слоты в их инвариантной структуре могут иметь различное заполнение, ср.: **ДФ**: *фиг/чёрт/Бог знает* – **КС** (*...) X (Y-а) знает; **ДФ**: *богс ним/с тобой* – **КС** (*…) X с Y-ом;
- 3 – к классу фразеологизированных коллокаций (ФК) (за счет переноса значения): *с ума сошёл*, *чушь собачья*;
- 2 – к классу ИД: *без вопросов, без проблем* (на грамматическом основании, хотя в качестве второй трактовки они имеют и статус РФ);
- 3 – к классу прагматических маркеров (ПМ): *вот так вот, как вам сказать, как сказать* (на прагматическом основании, хотя в качестве второй трактовки они имеюти статус РФ);
- 2 – к смешанному классу нефразеологизированных коллокаций (НК): *в самом деле, какая разница* (единицы этого класса пока не поддаются однозначной трактовке).

Таким образом, только 59 единиц из словника УНЕ (и отнюдь не только РФ) встречаются и в Прагматиконе, что составляет 9,7 % от всего списка ДФ.

Анализ словника РФ выявил основные особенности тех единиц, которые не попали в состав ДФ: преимущественно это слова стилистически сниженные или вообще табуированные. Например, в словнике Прагматикона зафиксировано много единицс корнем *Бог-* (*бог миловал, боже упаси, бог тебе судья*) и лишь одна – с корнем *господ-*(*а господь его знает*); тогда как в ОРД эти корни представлены почти одинаково часто (*недай Бог! прости Господи! ой Господи Боже мой! Ради Бога*), а кроме них, в словник РФ входят еще и такие единицы, как *К чёрту! Ни фига себе! Ни х*ра себе! Да ну на фиг!* и под.В число ДФ из сниженной лексики попали только *фиг/чёрт его/тебя знает* и *чёрт с ним/с тобой*. Такая стилистическая разница двух словников вполне, впрочем, объяснима, если помнить об источниках материала для фиксации и анализа.

Думается, что сопоставление двух баз данных, в том числе на уровне формульных единиц, или прагматических конструкций (ДФ и РФ), позволит сделать еще множество интересных наблюдений, полезных в разного рода системах автоматической обработки речи или искусственного интеллекта.

Исследование выполнено при финансовой поддержке гранта РНФ (проект № 22-18-00189 «Структура и функционирование устойчивых неоднословных единиц русской повседневной речи»).

Исторический подкорпус казахского языка: структура и лингвистическая разметка (Об опыте первого периода)

Айнур Аташбековна Сейтбекова[1], *Асель Сейдамат* [2]

[1] *Институт языкознания имени Ахмета Байтурсынулы, Алматы, Казахстан*
Ainurseit@mail.ru
[2] *Институт языкознания имени Ахмета Байтурсынулы, Алматы, Казахстан*
assel.seidamat@gmail.com

В мировой лингвистике вопрос создания больших языковых корпусов с каждым годом становится актуальнее. Корпусная лингвистика способствует развитию новых языковых теорий, основанных на анализе больших массивов данных, которые бросают вызов существующим ортодоксиям в прикладной лингвистике. В этом контексте создание исторического подкорпуса казахского языка представляет собой более сложный процесс, чем разработка других подкорпусов НККЯ, так как тексты раннего периода сохранились на разных графических системах, в различных списках и каллиграфических стилях.

Все тексты исторического подкорпуса должны быть представлены пользователю сайта НККЯ в конкордансе контекстов, где по каждому контексту даны наиболее подробные сведения. Параметры поиска, аннотаций подкорпуса исторических текстов должны разрабатываться для удовлетворения потребностей ученых лингвистов, специалистов других отраслей гуманитарной науки, учителей и преподавателей казахского языка и пользователей широкого круга. Учитывая многогранность и обширность потребностей современного пользователя, на первом этапе разработки подкорпуса исторических текстов были учтены лишь такие самые важные на первый взгляд параметры как оснащенность каждого текста, как элементы конкорданса, факсимиле, переводом на современный казахский язык, транскрипцией, метразметкой, лингвистичекой разметкой. Статья описывает разработку исторического подкорпуса Национального корпуса казахского языка (НККЯ), целью которого является систематизация и сохранение письменного наследия казахского языка с XI по XX век. Проект основывается на создании текстовой базы с письменными памятниками, представленными в различной графике, что обеспечивает доступность для лингвистических и культурных исследований.

Особенностью исторического подкорпуса является его уникальная метатекстовая разметка, отличающаяся от других подкорпусов НККЯ. Поскольку сбор и обработка письменных памятников и рукописей с полными и точными данными представляет собой сложный процесс, разработчикам пришлось учесть условия неопределенности, как в периоде создания ряда текстов, так и в отсутствии данных об издательстве и дате издания для некоторых рукописей. Исходя из опыта других исторических корпусов, таких как церковнославянский корпус, для исторического подкорпуса казахского языка была разработана детальная метатекстовая разметка. Она включает следующие параметры: краткую аннотацию, автора текста, графику, первую публикацию, стиль и жанр текста, источник и списки, автора транскрипции и перевода, количество страниц и словоупотреблений, сроки введения текста в корпус и примечания.

Создание лингвистической разметки в историческом подкорпусе оказалось одной из самых сложных задач, поскольку многие элементы текстов на арабской графике трудно распознать из-за разнообразной каллиграфии авторов и особенностей письменности. В процессе распознавания возникли следующие сложности: нечёткость и отсутствие

диакритических знаков (фатха, кесра, сукун) и графических символов, а также проблемы со слитным и раздельным написанием слов. Поэтому на первом этапе разметка применялась к транскрибированным текстам, что позволило решить ряд проблем, связанных с поиском и отображением слов в корпусе.

Для транскрипции арабографичных текстов на кириллицу разработчики столкнулись с особой проблемой передачи гласных звуков. В старописьменном тюркском языке было всего шесть букв для гласных (три для длинных и три для кратких), тогда как в древнем казахском языке представлено более девяти гласных звуков. В результате были определены следующие параметры лингвистической разметки:

Варианты: графические варианты слов в различных источниках.

Лемма: исходное слово или корень слова.

Лексический пласт: этимология лексической единицы.

Значение: значение лексической единицы в контексте.

Форма: основное и производное слово.

Аффиксы: словообразующие и формообразующие аффиксы.

Эти параметры разметки применяются только к тем словам, которые включены в реестр первого этапа разработки, уделяя внимание менее знакомым для современных пользователей словам. Для удобства поиска аффиксы, которые раздельно написаны в источниках, были переданы слитно, что упрощает процесс поиска и лемматизацию.

Ключевые задачи первого этапа включают:

Оцифровка и аннотирование: перевод исторических текстов на современный казахский язык, транскрибирование арабографических текстов на кириллицу и лингвистическая разметка транскрибированных текстов.

Разработка метаразметки и механизмов поиска: создание параметров метаразметки, которая позволяет учитывать жанр, стиль, графику и подстиль текста; и внедрение программного обеспечения для поиска лексико-грамматических элементов.

Параллельное представление: использование трехкомпонентного формата (факсимиле, транскрипция и перевод), что позволяет пользователям видеть оригинал, транскрипцию и перевод текста. Этот подход важен для анализа текстов на арабской графике, где часто встречаются трудности с обозначением диакритических знаков и написанием слов. Транскрипция и перевод облегчают понимание содержания оригинала.

Таким образом, разработчики исторического подкорпуса не только переняли опыт зарубежных исторических корпусов, но и сформировали эффективные подходы к решению специфических задач, связанных с транскрипцией и оцифровкой казахских письменных памятников. Данный проект способствует укреплению статуса казахского языка как значимого элемента межэтнической и межкультурной коммуникации, создавая основу для дальнейших исследований древнетюркского наследия.

Acoustic features of prosodic timbre

Скрелин Павел Анатольевич, skrelin@phonetics.pu.ru
Титюшина Анна Олеговна, anna.tityushina@yandex.ru
СПбГУ

Human emotion perception is a complex process, but it can be associated with some measurable characteristics such as pitch, energy and duration change as well as spectral qualities [1]. Speech timbre which is related to spectral characteristics of speech is one of the main prosody factors of an utterance. Prosodic timbre reflects emotional connotations of speech, while individual timbre is associated with personal characteristics of speaker's voice.

Nowadays there are methods of timbre cloning based on neural networks. However, these models require a reasonable amount of data, such as parallel corpus of the same utterances pronounced in different emotional contexts. The problem of transmitting and controlling prosody using utterances with different text transcriptions remains opened and requires formalization [2, 3].

In this work we propose a method that allows to check the significance of acoustic parameters of an utterance and detect the interconnection between spectral characteristics of sound and the emotional context. It is based on resynthesis of a signal, a method of acoustic characteristics modification of an audio sample aimed to reproduce a target signal with a high level of perceptual similarity. Experiments based on analysis through synthesis based on changing F0 frequency, allophone and syllable durations and intensity were held in different studies [4, 5]. However, a method of altering spectral characteristics in order to achieve required prosodic timbre still needs development.

The method proposed in this work is based on detecting the most significant frequencies of a vowel period spectrum and reconstructing a new fully synthesized vowel using found frequencies, target vowel duration and pitch contour. A resynthesized utterance was constructed of synthesized sonorant units based on target utterance spectrum and the rest allophones taken from the source utterance.

For testing the proposed algorithm performance it was used a set of utterances with identical transcription and different emotional backgrounds read by the same speaker. Auditory experiment verified that the developed method allows to achieve a high level of similarity between the target utterance and the resulting one. The synthesized utterance brings the target utterance emotional context and loses it’s previous tone.

The obtained results allow to research which parts of vowel spectrum have the strongest relevance to the speech prosody and to build prosodic timbre models related to different emotional aspects. The proposed method opens up new possibilities to analyze the perceptual effects of spectral changes and can be used in emotion recognition and speech synthesis fields of study.

References:

1. *Светозарова Н. Д.* Интонационная система русского языка. — Л.: Изд-во Лен. ун-та, 1982.
2. *Sisman, B.; Yamagishi, J.; King, S.; Li, H.* An overview of voice conversion and its challenges: From statistical modeling to deeplearning. IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 132–157.
3. *Wang, Y.; Stanton, D.; Zhang, Y.; Skerry-Ryan, R.J.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; Saurous, R.A.* Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In Proceedings of the 35th International Conference on Machine

Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; Dy, J.G., Krause, A., Eds.; PMLR: Mc Kees Rocks, PA, USA, 2018; Volume 80, pp. 5167–5176
4. *Kochetkova U., Skrelin P., Evdokimova V., Novoselova D.* Using audio signal modification to validate the acoustic features relevance, 2023, Записки научных семинаров Санкт-Петербургского Отделения Математического Института им. В. А. Стеклова РАН. — p. 86-101, 16 p.
5. *S. Gonzalez-Fuente, P. Prieto, I. Noveck,* A fine-grained analysis of the acoustic cues involved in verbal irony recognition in French, 06 2016.

Pragmatic units in contemporary poetry: a corpus-assisted analysis

Olga V. Sokolova (Institute of Linguistics, RAS, olga.sokolova@iling-ran.ru), Vladimir V. Feshchenko (Institute of Linguistics, RAS, vladimirfeshchenko@iling-ran.ru)

Key words: poetic discourse, corpus linguistics, pragmatics, anomalization

The paper examines the features of the pragmatic dimension of poetic discourse (hereinafter referred to as PD), which actively interacts with ordinary language in the modern era of new media (last decades). The role of pragmatics in contemporary poetry increases due to new strategies of subjectivation and addressing in the context of digital technologies and the transformation of the traditional model of communication (according to Roman Jakobson) and poetic auto-communication (according to Yuri Lotman). Although studies in the field of poetic pragmatics have been under way (Literary Pragmatics 1991; Person, Wooffitt, Rae 2021, Kovtunova 1986; Radbil 2012; Zoljan 2013; Sokolova 2024 etc.), a general analysis of this dimension of poetic discourse is still in demand. The existing studies of corpus pragmatics (e.g. Corpus Pragmatics 2014, Rühlemann 2019) have not yet paid attention to poetic corpora and poetic pragmatics.

As part of the research, a poetic corpus was compiled, including three sub-corpora in Russian, English, and Italian. The total volume of the corpus is about 3 million words. The corpus includes poetic texts from the 1960s to the 2020s.

To conduct a corpus-discourse analysis of contemporary poetry, a step-by-step analysis was developed, including quantitative and qualitative methods. A list of pragmatic units was compiled, which were assigned appropriate indices (for example, 111 - assertives, 12 - personal deixis markers, etc.). Based on the developed classification of linguopragmatic parameters, the corpus of poetic texts was marked using the AntConc and Taguette programs.

The study includes a quantitative analysis of the features of the use of pragmatic units in PD against the background of their conventional use in colloquial national corpora (in the spoken subcorpus of the National Corpus of the Russian Language; COCA (Corpus of Contemporary American English, Blogs), KIParla (L'italiano parlato e chi parla italiano), as well as in corpora of texts from the Internet, including text materials from the blogosphere, social networks and news resources).

The general conclusions that were made based on the calculations are related to the difference in the frequency of use of pragmatic units in PD and in colloquial speech. In general, the actualization of the pragmatic dimension often triggers a parallel anomalization at the level of semantics, grammar, and syntax.

Along with the common features of poetic pragmatics in different languages, the differences characteristic of poetic texts in different languages are also highlighted. For example, in Russian and Italian poetry, discourse markers are used more often than in American poetry, which is due to the more developed system of these units in the Russian and Italian languages in terms of quantity and variety. In addition, correlations have been identified between differences in the linguistic structure and the features of poetic pragmatics, which affect the specifics of the pragmatic experiment.

References

Corpus Pragmatics: *A Handbook* Ed. Aijmer K. & Rühlemann C. Cambidge University Press, 2014.

Kovtunova I.I. Pojeticheskaja rech' kak forma kommunikacii // Voprosy jazykoznanija. 1986. Vyp. 1. S. 3–13.

Literary Pragmatics. Ed. by R. D. Sell. Routledge, 1991.

Person R. F., Rae J. P., Wooffitt R., Bridging the Gap Between Conversation Analysis and Poetics. Routledge, 2021.

Radbil' T.B. Jazykovye anomalii v hudozhestvennom tekste: Andrej Platonov i drugie. M.: Flinta, 2012.

Rühlemann Ch. *Corpus* Linguistics for *Pragmatics* A guide for research. Routledge, 2019.

Sokolova O.S. New Technologies and Pragmatic Techniques in Contemporary Poetry // Slovo.ru. Vol. 15, 2024.

Zoljan T.S. Semantika i struktura pojeticheskogo teksta. M.: URSS, 2014.

Особенности просодического оформления лекторской речи

Титова Рената Руслановна, СПбГУ, renata3201@mail.ru;
Кочеткова Ульяна Евгеньевна, СПбГУ, u.kochetkova@spbu.ru

Целью настоящей работы является анализ просодического оформления речи лекторов – преподавателей вузов, а также проведение пилотных экспериментов по выявлению перцептивной релевантности отдельных просодических параметров при оценке эффективности коммуникации. Несмотря на большое количество работ, посвященных данной тематике, до сих пор недостаточно изучено мелодическое оформление в связи с темпоральными и динамическими характеристиками лекторской речи, особенности синтагматического членения; в работах отсутствует сравнительный анализ различных частей лекции. Зачастую можно встретить рекомендации для лекторов, а также некоторые общие сведения (без подробного акустического анализа) об интонационных конструкциях и/или оформлении интонационного центра высказывания. При этом на современном этапе развития речевых технологий усовершенствование диалоговых систем, используемых в том числе и в образовательных целях, требует уточнения информации о просодическом оформлении речи преподавателей вузов. Кроме того, важно определить и перцептивную релевантность отдельных просодических параметров.

В настоящей работе рассматриваются лекции двух преподавателей вузов (мужская и женская речь соответственно) для обучающихся гуманитарных направлений разных курсов. В ходе акустического анализа с использованием программного обеспечения Praat и WaveAssistant были рассмотрены следующие параметры: артикуляторный темп речи (на основе реальной транскрипции), частота и длительность пауз, средняя длина синтагмы, средняя длительность гласных, частота удлинения гласных, диапазон ЧОТ, среднее значение ЧОТ. Для проверки статистической значимости различий использовался t-критерий Стьюдента. Перцептивные эксперименты включали оценку участниками (56 человек) характеристик звучания голоса лектора, его эмоциональной выразительности и заинтересованности.

Сопоставление результатов перцептивного и акустического анализа вводной части лекции показало, что аудиторы давали более высокую оценку лектору-женщине (по параметрам «приятности», «интереса» и желания обучаться у данного преподавателя), чья речь отличалась по сравнению с лектором-мужчиной более высоким темпом (5,6 против 5,3 слогов в сек соответственно), более дробным членением (1,5 слова в синтагме против 3,5 слов), меньшей суммарной длительностью пауз (2,9 сек против 5,8 сек) и меньшим разнообразием пауз (наблюдались паузы молчания – синтагматические и выделительные, тогда как у лектора-мужчины присутствовали паузы молчания на границе синтагм, заполненные паузы хезитации, затягивание гласных и ларингализация в различных местах высказывания, а также паузы вдоха и выдоха). Кроме того, речь лектора-женщины отличалась не только более высокой средней ЧОТ (что естественно), но и большим диапазоном ЧОТ. Для выявления перцептивной значимости паузации из речи обоих дикторов были удалены различные виды пауз. Большинство аудиторов оценили предъявленные отрывки так же, как и в предыдущем эксперименте. Однако около трети участников поставили более высокие оценки как лектору-женщине (26%), так и лектору-мужчине (28%). Полученные результаты могут быть учтены как в дальнейших экспериментах на большем количестве материала, так и при создании корпуса речи, сбалансированного не только по социолингвистическим признакам, но и с учетом фонетических параметров, что особенно важно для обучения нейросетевых моделей.

Список литературы

1. *Ефремова Е.П. и др.* Исследование временного аспекта речи преподавателя на занятиях по иностранному языку. Фундаментальные исследования №9, 2012
2. *Риехакайнен Е.*И., Браташ В.С., Зубов В.И., Сергоманов П.А. Методика аннотирования корпуса устной речи учителей // Вопросы образования, №2, с.251-285, 2024
3. *Савинова М.С.* Профессия как часть социального статуса и ее отражение в просодических характеристиках речи. Вестник КГУ им. Н.А. Некрасова № 4, 2009
4. *Фрейдина Е. Л., Ковпак Н. А., Королева Ю. П., Пчелина Т. М., Сейранян М. Ю., Смирнова О. Н.* Просодия публичной речи: Прометей; Москва; 2013
5. *Järvinen, K.; Kähkönen, A.-L.; Nieminen, P.; Mäntylä, T.* Talking Like a Teacher—A Study of Pre-Service Teachers' Voice and Speech Characteristics in Learning and Teaching Situations. *Educ. Sci. 14*, 210; 2024

TranscribePro: Enhancing linguistic analysis through mobile technology

Victoria Firsanova, Saint Petersburg State University

In recent years, numerous large language models (LLMs), such as Gemma (Mesnard, Hardin, Dadashi et al. 2024), GPT-4 (Achiam, Adler, Agarwal et al. 2023), or Phi-3 (Abdin, Jacobs, Awan et al. 2024), demonstrate their capacity on a wide range of tasks from the field of theoretical linguistics. For example, LLMs proved their efficiency in semantic role labeling (Li, Kazeminejad, Brown et al. 2023), syntax tree generation (Altıntaş, Tantuğ 2023), and morphological segmentation (Pranjić, Šikonja, Pollak 2024). However, applying LLMs in academic research might be challenging.

Firstly, despite the accelerating development of LLM user interfaces, such as ChatGPT, most artificial intelligence models are not designed for non-technical users. Utilizing state-of-the-art neural network models requires a high level of computational expertise, which makes them inaccessible for many experts in Arts and Humanities. Secondly, the computational resources needed to run LLMs can be substantial. The deployment of LLMs is challenging on low-resource hardware commonly found in academic settings. Thirdly, the LLM development is typically focused on large-scale industrial applications rather than on specialized linguistic research. Consequently, LLMs lack support for essential linguistic tools and standards, such as the International Phonetic Alphabet (IPA), which is crucial for detailed phonetic and phonological analysis.

This paper introduces a novel mobile application TranscribePro, designed to make LLMs accessible and useful for linguistic research. The application integrates large-scale state-of-the-art neural networks into a single, minimalistic, user-friendly interface, providing a powerful tool for phonetic and phonological analysis, as well as language education. The research focuses on exploring approaches to LLM compression, adapting existing models to the specific needs of the linguistic community, and making advanced LLM technology accessible for both theoretical linguistics and language education through efficient design.

Most LLM compression techniques are based on the intrinsic dimensionality concept (Aghajanyan, Gupta, Zettlemoyer 2021), demonstrating that significant training data features are represented by a relatively small number of the overall model parameters. Such techniques, as quantization (Gholami et al. 2022), palletization (Cho et al. 2021), pruning (Hoefler et al. 2021), knowledge distillation (Gou et al. 2021), and low-rank adaptation (Hu et al. 2021), are often used for LLM compression. This study implies exploring the advantages and limitations of each method in the context of mobile development.

Fig. 1 illustrates the TranscribePro application user interface. The project is open source, the code base is provided at GitHub (https://github.com/vifirsanova/TranscribePro). TranscribePro is an Android application that uses Kotlin programming language and XML layouts. The machine learning models are hosted on user device. The application allows for loading pre-recorded audio file supporting a wide range of file extensions. The application automatically creates a linguistic transcription file using a markup language, inspired by such annotation tools as ELAN and Praat.

The study focuses on exploring LLMs for foundational linguistic research. The research method implies training, fine-tuning and adapting various multilingual machine learning models for text-to-speech synthesis in the theoretical linguistics context. The study consists of technical and theoretical parts. The technical part of the study compares two different baselines: (1) converting OpenAI Whisper to TensorFlow Lite format and integrating the model to an application; (2) using C++ version of OpenAI Whisper and Java Native Interface to run quantized model on a mobile device. The theoretical part of the study implies the research and development of annotation for phonetic and phonological studies based on empirical material.

The technical part of the study compares inference speed, model size, and performance based on user feedback. The theoretical stage resulted in transcription markup development

validated using empirical linguistic data from multiple languages. The study demonstrates the potential of applying LLMs in foundational linguistic research, highlighting the advantages and trade-offs of different deployment strategies. Both TensorFlow Lite and C++ models maintained high accuracy levels, suitable for linguistic research and language learning. User feedback indicated a preference for the C++ model, while the TensorFlow Lite model is recommended for its ease of integration. The annotation framework developed in the theoretical part of the study provided a robust tool for phonetic and phonological studies, contributing to the field of theoretical linguistics.

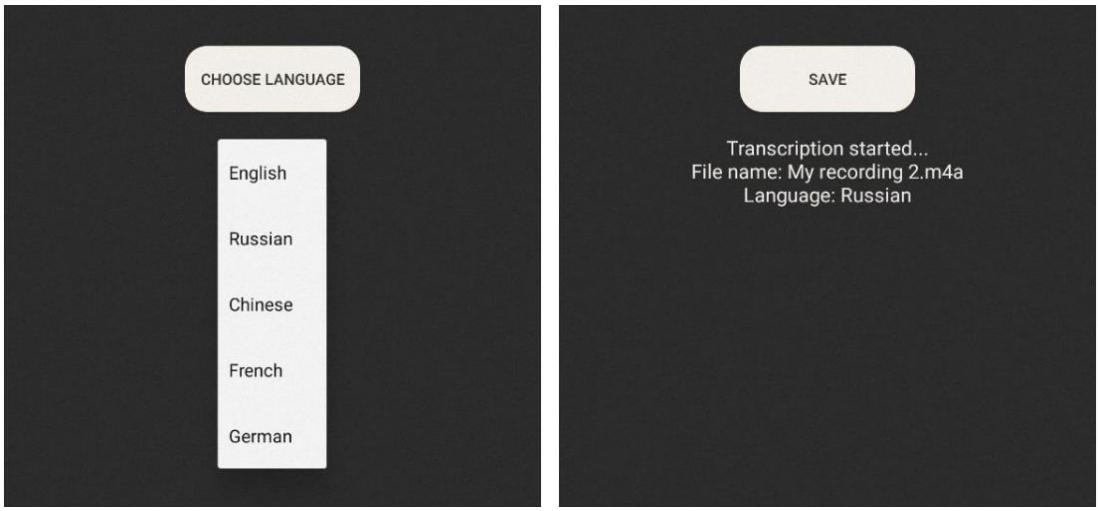


Fig. 1. TranscribePro application user interface example

References

Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H. and Benhaim, A. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Aghajanyan A., Gupta S., Zettlemoyer L. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021. P. 7319-7328.

Altıntaş M., Tantuğ A.C. Improving the performance of graph based dependency parsing by guiding bi-affine layer with augmented global and local features. Intelligent Systems with Applications 1 (18), 2023.

Cho M. et al. DKM: Differentiable k-means clustering layer for neural network compression. arXiv preprint arXiv:2108.12659, 2021.

Gholami A. et al. A survey of quantization methods for efficient neural network inference. Low-Power Computer Vision. Chapman and Hall/CRC, 2022. P. 291–326.

Gou J. et al. Knowledge distillation: A survey. International Journal of Computer Vision, 2021, 129 (6). P. 1789-1819.

Hoefler T. et al. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research, 2021, 22 (241). P. 1-124.

Hu E.J. et al. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

Li T., Kazeminejad G., Brown S.W., Srikumar V., Palmer M. Learning Semantic Role Labeling from Compatible Label Sequences. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023. P. 15561-15572.

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J. and Tafti, P.. Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.

Pranjić M., Robnik-Šikonja M., Pollak S. LLMSegm: Surface-level Morphological Segmentation Using Large Language Model. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024. P. 10665-10674.

Paraphrasing Tools in Uzbek Language: Characteristics of Uzbek Journalistic Style and Distinguishing Features from Other Styles

Zarnigor M. Khayatova
Tashkent State University of Uzbek Language and Literature named after A. Navoi
Tashkent, Uzbekistan
e-mail: khayatovazarnigor@gmail.com
Khamroeva Shahlo Mirdjonovna
Tashkent State University of Uzbek Language and Literature named after A. Navoi
Tashkent, Uzbekistan shaxlo.xamrayeva@navoiy-uni.uz

Introduction: The research focuses on the growth and utilization of paraphrasing tools adapted for the Uzbek language, with special focus on journalistic style. In many linguistic fields such as translation, education in a foreign language, and natural language processing, paraphrase instruments have gained increasing significance. The aim of this study is determination of unique Uzbek journalistic style characteristics as well as its distinction from other stylistic forms by meansof advanced paraphrase tools.

Problem Statement: Paraphrase tools face distinct challenges when working with the rich culturaland literary heritage found within the Uzbek language which contains diverse stylistic nuances. The language of journalism specifically has certain linguistic features that differentiate it from other sciences such as technical, creative and conversational languages used in everyday life. This research aims to identify them so that they can be included into rephrase programs that are fit for use among speakers of Uzbek where accuracy also matters.

Literature Review: A critical review of existing literature on paraphrasing tools and their application in various languages reveals a gap in resources specifically designed for the Uzbek language. Notable contributions in the field include works by Djavdet Suleiman and Dr.Fridman, who have extensively studied the stylistic features and linguistic nuances of the Uzbek language. Suleiman's research highlights the complexity of Turkic languages' syntax and morphology, particularly (Suleiman, 2020), while Fridman emphasizes the importance of semantic accuracy intranslation tools (Fridman, 2019).

Additionally, research by other scholars, such as Karimov and Tursunov, has underscored the necessity of developing specialized tools that can address the unique stylistic elements of less commonly studied languages like Uzbek (Karimov, 2021; Tursunov, 2022). These studies collectively highlight the need for paraphrasing tools that can handle the subtleties of different styles, particularly in the journalistic domain, which requires a balance of formality, clarity, and informativeness.

Methodology: The study uses a mixed-methods approach, through qualitative and quantitative analysis aiming at identifying key characteristics of Uzbek journalistic style. In this regard, linguistic examination is done on an Uzbek journalistic text corpus to extract features like syntactic constructions, morphology, and semantic structures which are unique only to journalists' writing. This will lead us to designing a prototype paraphrase tool that would capture these aspects while ensuring accuracy.

Results: nitial findings show that there is marked improvement in terms of capturing the essence of the Uzbek journalism in comparison with generic paraphrasing tools; thus, highlighting its successful development. The tool actually maintains stylistic integrity without inaccuracies and it does this by relying on correct contextual information it has captured from original materials.

Conclusion: The study underscores the importance of developing language-specific paraphrasing tools that consider the unique characteristics of different styles. For the Uzbek language, incorporating the distinctive features of the journalistic style into paraphrasing tools not only enhances their functionality but also contributes to the preservation and promotion of Uzbek journalistic traditions. Future research should explore the extension of this approach to

other stylistic forms and languages, fostering the development of more sophisticated natural language processing tools.

Палитра эмоций в германской прессе: опыт автоматического анализа

Хохлова М.В. (m.khokhlova@spbu.ru), Корышев М.В. (m.koryshev@spbu.ru), Санкт-Петербургский государственный университет

За последние два десятилетия методология филологических и лингвистических исследований претерпела значительные изменения, которые связаны с появлением новых методов анализа текстов и их описания. Возможности, которые открывают современные методы, привели как к зарождению новых прикладных направлений, так и к проявлению интереса к уже существующим. Так, методы автоматической обработки текстов стимулировали развитие междисциплинарных направлений, находящихся на стыке культурологии, политологии, социологии, психологии, с одной стороны, и лингвистики и филологии, с другой.

Частотность лексических единиц в текстах соотносится со значимыми событиями, представляющими читательский интерес, поэтому при помощи количественных методов можно выявить ключевые лексемы и описать взаимосвязь с характеризующим их историко-политическим контекстом. При помощи методов компьютерной лингвистики прослеживаются диахронические изменения в тематике печатных изданий, что находит отражение в ряде исследований. Так, в статье [Yang, Torget, Mihalcea 2011] авторы рассматривают американские газеты в периоды с 1865 по 1930 годы и используют алгоритм MALLET для выявления наиболее употребительной лексики. Прослеживается связь с историческими событиями того или иного периода (например, связь слова “cotton” ‘хлопок’ с возникновение партии Популистов в Техасе, полностью зависящим от производства хлопка после Гражданской войны).

Несмотря на то, что англоязычный материал преобладает в исследованиях, тем не менее, существуют работы, посвященные применению рассматриваемых методов к текстам на немецком языке. Анализ парламентских дебатов в немецком Бундестаге по вопросам, связанным с углем, был выполнен в работе [Müller-Hansen et al. 2021]. Представления о южных странах (Португалия, Италия, Греция, Испания) анализировались при помощи структурного тематического моделирования (STM) в немецкоязычной прессе с 1946 по 2009 гг. на материале газеты “Die Welt” [Küsters, Garrido 2020].

Материалом нашего исследования послужили 46 879 текстов немецкоязычных изданий “Spiegel” и “Zeit” общим объемом 29 499 497 токенов. Отбор статей осуществлялся автоматически по следующим тематическим рубрикам: *Familie*, *Gesellschaft*, *Kultur*, *Politik*, *Wissen*, *Ausland*, *Leben* и *Panorama*. При помощи словаря [Mohammad, Turney 2013] была произведена автоматическая разметка текстов по эмоциям и классам положительно и отрицательно окрашенной лексики. Полученные результаты позволили проследить связь между темами, которые затрагиваются в упомянутых изданиях, и значимыми лексемами, которые могут являться маркерами определенных настроений.

Исследование выполнено за счет гранта Российского научного фонда № 24-28-00937, https://rscf.ru/project/24-28-00937/.

Ключевые слова: эмоции, лексика, немецкая пресса, автоматическая разметка, квантитативный анализ

Список литературы

Küsters A., Garrido E. Mining PIGS. A structural topic model analysis of Southern Europe based on the German newspaper Die Zeit (1946-2009). In Journal of Contemporary European Studies 28(4). 2020. P. 477-493.

Mohammad S., Turney P. Crowdsourcing a Word-Emotion Association Lexicon. In Computational Intelligence, 29 (3), 2013. P. 436-465.

Müller-Hansen F., Repke T., Baum Ch. M., Brutschin E., Callaghan M. W., Debnath R., Lamb W. F., Low S., Lück S., Roberts C., Sovacool B. K., Minx J. C. Attention, sentiments and emotions towards emerging climate technologies on Twitter. In Global Environmental Change, Volume 83, 2023, 102765, https://doi.org/10.1016/j.gloenvcha.2023.102765. URL: https://www.sciencedirect.com/science/article/pii/S0959378023001310

Yang Tze-I., Torget A.J., Mihalcea R. Topic modeling in historical newspapers. In Proceedings of the 5th ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2011. Association for Computer Linguistics, 2011. P. 96–104.

Система интерактивных заданий в рамках модели "Перевёрнутый класс" (начальный этап обучения РКИ)

Цянь Цияо, Санкт-Петербургский государственный университет, 1685632927@qq.com
Анциферова Ольга Васильевна,, Санкт-Петербургский государственный университет, olga.anciferova.23@mail.ru

В настоящее время, система преподавания русского языка как иностранного сложилась ситуация, когда устоявшиеся методы, приемы и формы обучения требуют осмысления, коррекции и новых педагогических решений. Это обусловлено прежде всего повсеместным внедрением и широким использованием информационно-коммуникационных технологий (ИКТ). Употребление ИКТ в сфере преподавания иностранного языка позволяют моделировать аутентичную языковую среду, оснащать учебные курсы интерактивными материалами, увеличивать долю самостоятельной работы учащихся, повышать их мотивацию и эмоциональную вовлеченность в процесс обучения [Ермакова, 2021; Shetzer, Warschauer, 2020; DurraniU, 2022; Березняцкая, Денисенко, Калинина, 2023]. Целый ряд объективных причин позволяет сегодня говорить о приоритете педагогической модели «перевёрнутый класс» и использования интерактивных заданий в процессе обучения, позволяющего эффективно формировать коммуникативную компетенцию в условиях продуктивной самостоятельной работы учащихся.

Модель «перевернутый класс» является одной из составляющих смешанного обучения и определяется как инновационный подход к организации обучения, при котором прямое обучение перемещается из пространства группового обучения в пространство индивидуального обучения, а результирующее групповое пространство трансформируется в динамическую интерактивную среду обучения с помощью педагога [Афзалова, 2022; Золотых, Цю С, 2018; Тихонова, 2018; Ян Яньцзе, 2022]. Модели «перевернутый класс» посвящено значительное число зарубежных исследований: M. Ronchetti, M. Kalogiannakis, U.K. Durrani, J. Bergmann, G. Sprint, O. Hujran. В Китае этой проблемой активно занимаются, Jon Chao Hong, Ming Yueh Hwang, Yi Hsuan Liu, Kai Hsin Tai. Среди российских ученых следует выделить исследования Н. В. Тихонова, Ж.И. Трафимчик, О. В. Востриковой и другие.

В методике обучения иностранному языку накоплен опыт использования модели «перевернутый класс» в системе высшего образования как в России, так и за рубежом. В Российском университете дружбы народов (РУДН) разработано учебное пособие «Русская литература на экране» в целях интеграции технологии «перевернутый класс» в практику преподавания русского как иностранного [Денисенко, 2023]. В Китае, в таких университетах, как Сычуаньский [Золотых, Цю, 2018], Университет Внутренней Монголии [Ян Шуфан, 2017; Ян Яньцзе, 2022], технология «перевернутый класс» успешно используется в учебном процессе на продвинутом этапе обучения. Однако следует отметить, что в этой области недостаточно исследован вопрос эффективности обучения русскому языку как иностранному на начальном этапе с помощью модели ПК с применением ИКТ-технологий.

Внедрение модели «перевернутый класс» в процесс обучения иностранному языку инициирует пересмотр основополагающих и выделение доминирующих принципов, учитывающих характеристики данного субъекта образовательного процесса [Ридная, 2014]. Специфика обучения иностранному языку на начальном этапе (комплексный урок) обоснована рядом основополагающих принципов: принцип доступности; принцип посильности; принцип системности; принцип мотивации; принцип учета родного языка учащихся и принцип интерактивности. Учитывая вышеуказанные принципы и реальные условия обучения, нами разработана адаптированную модель «перевернутый класс» для элементарного уровня обучения РКИ с использованием интерактивных заданий.

Под интерактивном упражнением понимается творческое учебное задание, которое требует от учащихся не простого воспроизводства информации, а содержит больший или меньший элемент неизвестности и имеет, как правило, несколько подходов [Ковалева, 2019]. Интерактивное задание способствует установлению неформальной обстановки на занятии и тем самым созданию условий для информационно и эмоционально полноценного общения [Белик, 2020; Яшина, Чернякова, 2021, Бабаева, 2023]. Существуют такие ученые, как Бабаева Д., Белик, В.А., Кудрявцева, О.А., Блохин К.А., Ковалева А.В., Сопова Е.,Чичерина Н.В., которые трудились в этой области исследования.

Структурный компонент разработанной методической модели обучения китайских учащихся элементарного уровня владения русским языком трёхчастный: подготовительный этап, аудиторный этап, постаудиторный этап. В соответствии со трёхэтапной моделью ПК мы выделили систему интерактивныхзаданий, которая включает эвристическая беседа, кластера, вопросно-ответные упражнения, ответить вопросы на основе аутентичных аудиовидео – материалов,собери диалог в нужном порядке, игровые формы упражнения и другие. Задания распределены по трём этапам: этап знакомства с изучаемым явлением языка; этапотработки автоматизма в употреблении языковых единиц; этап формирования умений свободного пользования усвоенными единицами в речи. Каждый этап имеет свою цель. Логика построения учебных материалов и системы упражненияправлена на формирование и развитие всех видов речевой деятельности.

Система интерактивных заданий позволяет всем обучающимся активно участвовать в процессе обучения, быть активными, сотрудничать совместно друг с другом для достижения положительного результата, соответствует определенным стадиям педагогического взаимодействия. Она создает своего рода естественную ситуацию общения, цель которой – «поставить обучаемого в такие условия, в которых осуществление акта общения на иностранном языке становится для него насущной необходимостью» [Алхазишвили, 1988]. Каждая группа интерактивных упражнений подкреплена определенным видом продуктивной деятельности, который предполагает общение учащихся для достижения совместного результата и, следовательно, способствует развитию речи на изучаемом языке.

Список литературы

1. Алхазишвили А.А. Основы овладения устной иностранной речью: Учеб. Пособие для студентов пед. ин-тов. М.: Просвещение, 1988. – 128 с.

2. Афзалова А.Н., Тренды образовательных технологий в современном мире / Педагогико-психологические и медико-биологические проблемы физической культуры и спорта (электронный журнал). 2022, т.17, в.4, с.165-170

3. Бабаева, Д. Использование интерактивных технологий при обучении английскому языку / Д. Бабаева// Матрица научного познания. – 2023. – № 10-2. –С. 364-366.

4. Белик, В.А. Использование интерактивных технологий обучения в преподавании иностранного языкав вузе / В.А. Белик // Вопросы педагогики. –2020. – № 11-2. – С. 44-48.

5. Березняцкая М.А., Денисенко, Ю.М., Калинина Ю. М. Информационно-коммуникационные технологии в практике преподавания русского языка, литературы и культурологии иностранным студентам-билингвам нефилологических специальностей // Полилингвиальность и транскультурные практики. 2023. Т. 20. № 1. С. 179–187

6. Блохин К.А. Применение интерактивных диалоговых тренажёров в обучении иностранному языку. Университетские чтения – 2021: материалы научно- педагогических чтений ПГУ. Пятигорск, 2021; Ч. 2: 67–75. Available at: https://www.elibrary.ru/item.asp?id=46258537

7. Ермакова О.Б. Цифровое сопровождение процесса обучения РКИ (на примере виртуального класса Google) // РКИ: Лингвометодическая образовательная платформа: Сб. трудов Междунар. научно-практич. конф., посв. 145-летию Белгородского

государственного национального исследовательского университета. Белгород, 2021. С. 100–106.

8. Золотых Л.Г., Цю С. «Перевернутый класс» как новый метод преподавания русского языка в практике китайских вузов: опыт Сычуаньского университета. Русистика. 2018. Т. 16. № 4. С. 451—463.

9. Ковалева А.В. Интерактивные педагогические технологии при обучении иностранных учащихся лексике русского языка: автореф. дис. … канд. пед. наук.М., 2015. – 23 с.

10. Кудрявцева, О.А. Интерактивные интерент-технологии в практике обучения иностранным языкам /О.А. Кудрявцева // Студенческая наука и XXI век. – 2018. –

№ 2-2. – С. 280-283.

11. Ридная Ю. В. Принципы обучения иностранному языку магистрантов технического профиля // Приоритетные задачи развития системы профессионального образования на современном этапе. 2014. № 2. С. 89-94.

12. Сопова Е. Технологии интерактивного обучения на уроках иностранного языка // Просвещение. Иностранные языки: Интернет-издание для учителя. [Электронный ресурс], 2013. Режим доступа: http://iyazyki.ru/2013/03/interactivelanguage , свободный. – Загл. с экрана. – Яз. рус.

13. Тихонова Н.В. Технология «перевернутый класс «в вузе: потенциал и проблемы внедрения / Н.В. Тихонова // Казанский педагогический журнал. - 2018.

- №2. - С. 74-78.

14. Трафимчик Ж.И. Информатизация образования с позиции ее позитивных и негативных сторон / Ж.И. Трафимчик // Проблемы здоровья и экологии. – 2017. –

№2 (52) [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/informatizatsiya-obrazovaniya-s-pozitsii-ee

15. Чичерина Н.В. Интерактивные методы обучения в контексте интеграции языкового и медиаобразования // Вопросы современной науки и практики. Университет им. В.И. Вернадского. 2008. – №3(13). – Том 1. – С. 76-82.

16. Ян Шуфан, Исследование о внедрении модели ПК в инновационной практике преподавания русского языка в вузах, Университет Внутренней Монголии. 2017.

17. Ян Яньцзе и др. Влияние перевернутого класса на эмоции и мотивацию студентов // Научное исследование. 2022. № 4.

18. Яшина Т. Р., Чернякова Ю. С. Интерактивные методы обучения чтению на уроках английского языка в средней школе // Научно-методический электронный журнал «Концепт». – 2021. – № 4 (апрель). – С. 34–45. – URL: http://e- koncept.ru/2021/211020.htm.

19. DurraniU. K. “Gamified flipped classroom versus traditional classroom learning: Which approach is more efficient in business education?” The International Journal of Management Education, vol. 20, no. 1, Mar. 2022, doi: 10.1016/j.ijme.2021.100595.

20. G. Sprint and E. Fox, “Improving student study choices in cs1 with gamification and flipped classrooms,” Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE, 2020, pp. 773–779, doi: 1145/3328778.3366888.

21. Jeong J.S., González G.D. Students’ perceptions and emotions to ward learning in a flipped general science classroom // Journal of Science Education and Technology. 2016. № 25 (5). P. 747-758.

22. J. Bergmann and A. Sams, Flip your classroom: reach every student in every class every day. International Society for Technology in Education, 2012.

23. J.-C. Hong, M.-Y. Hwang, Y.-H. Liu, and K.-H. Tai, “Effects of gamifying questions on English grammar learning mediated by epistemic curiosity and language anxiety,” Computer Assisted La

24. J. Zhao, G.-J. Hwang, S.-C. Chang, Q. Yang, and A. Nokkaew, “Effects of gamified interactive e-books on students’ flipped learning performance, motivation, and meta- cognition tendency in a mathematics course,” Educational Technology Research and Development, vol. 69,

no. 6, pp. 3255–3280, Dec. 2021, doi: 10.1007/s11423-021- 10053-0.

25. J.-C. Hong, M.-Y. Hwang, Y.-H. Liu, and K.-H. Tai, “Effects of gamifying questions on English grammar learning mediated by epistemic curiosity and language anxiety,” Computer Assisted Language Learning, vol. 35, no. 7, pp. 1458–1482, Sep. 2022, doi:10.1080/09588221.2020.1803361.

26. Kalogiannakis M. “Gamification in Science Education. A Systematic Review of the Literature,” Education Sciences, vol. 11, no. 1, p. 22, Jan. 2021, doi: 10.3390/educsci11010022.

27. Lamarca A. Longo Addressing student motivation, self-regulation and engagement in flipped classroom to decrease boredom // International Journal of Information and Education Technology. 2017. № 7 (3). P. 230.

28. Ronchetti M. ―Using video lectures to make teaching more interactive, International Journal of Emerging Technologies in Learning (IJET), June 2010.

29. U. Durrani, O. Hujran, and A. S. Al-Adwan, “CrossQuestion game: A group-based assessment for gamified flipped classroom experience using the ARCS model,” Contemporary Educational Technology, vol. 14, no. 2, Jan. 2022, doi:10.30935/cedtech/11568.

30. Shetzer, H. and Warschauer, M. (2000). An Electronic Literacy Approach to Network-Based Language Teaching in Network-Based Language Teaching: Concepts and Practice, M. Warschauer and R. Kern (ed.), New York, Cambridge University Press,pp. 171–185.

Устойчивые неоднословные единицы русской устной речи: методы пополнения словника и статистический анализ

Т.Ю.Шерстинова, tsherstinova@hse.ru, Т.И.Попова. tipopova@hse.ru Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург; Санкт-Петербургский государственный университет

Изучение единиц устной речи, которые выполняют особые дискурсивные функции, давно проводится на материале различных корпусов. Роль таких дискурсивных единиц, конструкций или формул лингвисты описывают с разных сторон: в грамматическом, семантическом и прагматическом аспекте.

Анализ особенностей прагматических маркеров устной речи на материале корпуса «Один речевой день» позволил обнаружить некоторые закономерности: единицы устной речи зачастую многокомпонентны, и их функция в дискурсе фиксируется прежде всего за счет употребления единицы в определенном виде. Так было сформировано понятие устойчивой неоднословной единицы (УНЕ) [Bogdanova-Beglarian et al. 2023]. Эти элементы подразделяются на следующие основные классы[1]:

1. фразеологизированные конструкции;
2. форма-идиома;
3. конструкция;
4. речевая формула;
5. прагматический маркер;
6. окказиональные единицы;
7. фрагменты прецедентных текстов;
8. нефразеологические коллокации.

В результате пилотной экспертной разметки пользовательского подкорпуса объёмом 300 тыс. словоупотреблений был сформирован базовый словник УНЕ, состоящий из 1088 единиц. В результате дальнейшей коррекции, выявления наиболее частотных ошибок и перевода некоторых единиц в другой класс в словнике осталось 928 единиц всех типов.

Была получена предварительная статистика употребления этих единиц на экспериментальной выборке. В полученный список попали самые частотные и распространенные устойчивые неоднословные единицы русской повседневной речи, которые можно считать «ядром» системы таких элементов для русского дискурса. Такие, например, как В ПРИНЦИПЕ (ранг 1) – лексикализованная предложно-падежная форма (форма-идиома или прагматический маркер – вербальный хезитатив или разграничитель, прежде всего навигационный, в зависимости от контекста), ЭТО САМОЕ (ранг 2) (вербальный хезитатив, маркер самокоррекции, разграничительный маркер всех трех типов (стартовый, навигационный и финальный) и – редко – маркер-ксенопоказатель), НА САМОМ ДЕЛЕ (ранг 5) (вербальный хезитатив) и И ТАК ДАЛЕЕ (ранг 7) (маркер-заместитель). По предвартельным подсчетам, около половины наиболее частотных УНЕ нашей устной коммуникации – это прежде всего прагматические маркеры [Bogdanova-Beglarian, Filyasova 2018], не попадающие в этом статусе ни в традиционные толковые словари, в том числе словари русской разговорной речи, ни в «Русский Конструктикон», ни в «Прагматикон».

Но, очевидно, список УНЕ, полученный на ограниченной выборке, на этом не исчерпывается, поэтому задачей настоящего исследования является расширение этого списка – русского словника устойчивых неоднословных единиц. Пополнение словника может идти двумя путями: 1) продолжение эмпирической разметки путем сплошного аннотирования расшифровок звукозаписей и выделения новых единиц с указанием их

[1] Случаи, когда единица относится к двум разным классам, встречаются, но только дискурсивный контекст сможет разрешить коммуникативную неоднозначность.

контекста и класса и 2) привлечение автоматических методов обработки корпусных данных. Материалом для анализа являются расшифровки корпуса «Один речевой день», при этом экспертная разметка выполняется на объеме в 700000 словоупотреблений, а автоматическая – на всем объеме существующих расшифровок корпуса, в том числе и полученных автоматически, общим объемом более 1,5 млн словоупотреблений.

Методы автоматического выявления устойчивых неоднословных единиц опираются на существующие словники, однако ввиду омонимичности языковых единиц требуют последующей ручной коррекции. Специальные скрипты создаются для поиска новых форм конструкций [Рахилина 2010] по заданным правилам на основе записи УНЕ этого класса в их инвариантной структуре. Более того, ведется работа о подготовке скриптов для выявления на корпусном материале дискурсивных элементов, представленных в других словарях и электронных ресурсах, посвященных дискурсивным элементам устной речи.

В докладе описаны методы автоматической обработки корпуса, а также представлены количественные данные, полученные в результате ее применения: получена статистика о частоте реализации каждой устойчивой неоднословной единицы на представительном речевом материале, выделены наиболее частотные дискурсивные единицы, представлены модели их образования и вариативность компонентов. Полученные данные лягут в основу Словаря коллокаций и других устойчивых неоднословных единиц повседневной русской речи.

Исследование выполнено при поддержке гранта РНФ № 22-18-00189 «Структура и функционирование устойчивых неоднословных единиц русской повседневной речи».

Использованная литература и ресурсы:

1. Bogdanova-Beglarian N. V., Blinova O. V., Khokhlova M. V., Sherstinova T. Yu. Towards the Description of Multiword Units in Russian Everyday Speech: State-of-the-Art and the Methodology of Further Research. In: Mukhamediev R., Pereira R., Mityagin S., Bolgov R. (eds.) Springer Geography, Springer Nature, 2023. Pp. 129-139
2. Bogdanova-Beglarian N. V., Filyasova Yu. A. Discourse vs Pragmatic Markers: A Contrastive Terminological Study // 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts, SGEM 2018. Vienna ART Conference Proceedings, 19—21 March, 2018. Vol. 5, Iss. 3.1. P. 123—130.
3. Рахилина Е. В. Лингвистика конструкций / Отв. ред. Е. В. Рахилина. М.: Издательский центр "Азбуковник", 2010.
4. Корпус «Один речевой день» https://ord.spbu.ru/
5. «Русский Конструктикон» https://constructicon.github.io/russian/
6. «Прагматикон» https://pragmaticon.ruscorpora.ru

Hypallage in Machine Translation

Natalia Shutemova[1], *Natalia Sokolova*[2]
[1,2]*Saint Petersburg State University*
[1]*n.shutemova@spbu.ru*
[2]*n.y.sokolova@spbu.ru*

Our research considers how Machine Translation (MT) copes with linguistic creativity. This phenomenon characterises various kinds of discourse, is subject to representation in translation, influences its accuracy and quality, which explains the relevance of the issue for Translation Studies, applied, corpus and comparative linguistics [1–3].

The research is focused on hypallage and is aimed at studying its representation in machine translation of literary texts. Belonging to alogisms, hypallage is a challenge for comprehending, on the one hand, and verbalising in target languages (TL), on the other. Its translation is a nontrivial task presupposing a heuristic, rather than a standard, solution. We are going to divide this paper into two main parts corresponding to the following key questions: 1. What is hypallage? 2. How is it represented in MT?

To answer these questions we use semantic, syntactic, derivational, cognitive, stylistic and comparative analysis of source hypallages (SH) in the novel in verse "Eugene Onegin" by A. S. Pushkin [4] and their English versions produced by the neural machine translation services "DeepL Translator" (advertised as the world's most accurate tool), "Google Translate" and "Yandex Translate" in difference from variants created by J. Falen [5] (one of the most famous "human" translators of the original).

Hypallage is a semantic and syntactic shift based on reinterpreting static and dynamic characteristics of an object:

1) a property of an object may be regarded as a feature of its action and expressed by means of an adverb rather than an adjective ("*gryazno taet* Na ulitsakh razrytyy *sneg*" /lit. 'The scattered *snow melts muddily* in streets'/ [4: 219]);

2) a property of an action may be perceived as an object's feature and verbalised abnormally by an adjective rather than an adverb ("To stan sov'et, to razov'et I *bystroy nozhkoy nozhku b'et*" /lit. 'Now she bends, now she unbends, and *beats her leg with a rapid foot*'/ [4: 23]).

Thus, the shift results in a contradiction between semantic and syntactic relations in an utterance. In the first example the adverb "gryazno" ('muddily') semantically depends on the noun "sneg" ('snow'), while syntactically – on the verb "taet" ('melts'). In the second case the adjective "bystroy" ('rapid') semantically depends on the verb "b'et" ('beats'), but syntactically – on the noun "nozhkoy" ('with/by a foot'). This contradiction allows the writer to accentuate his ideas in a laconic form.

Although hypallage is a much rarer phenomenon than metaphor or metonymy, it is similarly a mental procedure that characterises an individual way of thinking and interpreting common reality. Ideally, the search for hypallage could be optimised through corpus analysis. However, corpora are not provided with appropriate stylistic annotation and fail to select hypallages. Moreover, we have checked how corpora tag hypallages manually collected by us. For example, the research which was conducted using The Russian National Corpus (ruscorpora.ru) has revealed inconsistencies in their morphological annotation. If hypallages of the second kind are correctly tagged as adjectives, those in the first group are incorrectly tagged as adjectives rather than adverbs.

The discrepancy between semantic and syntactic relations determines difficulties for "human" translation of hypallage. For instance, James Falen deconstructed the majority of Pushkin's hypallages (70%) and managed to reconstruct 30%. It means that the translator attempted to convey hypallage as a feature of Pushkin's style but in some cases could not do it because of rhythmic requirements (metre, foot, rhyme).

Comparative analysis of source and target contexts allows us to differentiate four ways in which MT represents SH, irrespective of the ST's rhythm. Firstly, SH is formally restored, with sense being unchanged: "*Glazami beglymi chitaet* Prostuyu nadpis'" [4: 164] – *With fluent eyes reads* a simple inscription ("DeepL"). This method was used by "DeepL" in 70% of SH, by "Google" – 50% and "Yandex" – 40%. Secondly, SH is formally reproduced, but the ST's sense turns into nonsense: "To stan sov'et, to razov'et I *bystroy nozhkoy nozhku b'et*" [4: 23] – Then the camp will grow, then it will develop And *beat the leg with a quick foot* ("Yandex"). This method was revealed in "Google" (10%) and "Yandex" (10%). Thirdly, SH is destructed, but the ST's sense is conveyed: "I *neotvyazchivyy lornet On obrashchaet* pominutno Na tu, chey vid napomnil smutno Emu zabytye cherty" [4: 205] – And he *turns his lorgnette to the one* whose appearance reminded him vaguely of forgotten features ("DeepL"). This method was used by "DeepL" (30%) and "Yandex" (20%). Finally, SH is destructed, while the ST's sense is replaced with nonsense: "No tak i byt' – *rukoy pristrastnoy Primi* sobran'e pestrykh glav" [4: 11] – But so be it – *with a partial hand, accept* the collection of motley chapters ("Google"). This method characterises "Google" (40%) and "Yandex" (30%). Thus, regardless of the ST's rhythm, "DeepL", "Google" and "Yandex" can restore both the form and meaning of SH. "DeepL" rarely reduces the former, "Google" – the latter or both. "Yandex" covers all the four options.

References

1. Biber, D.: Representativeness in Corpus Design. In: Zampolli, A., Calzolari, N., Palmer, M. (eds.) Current Issues in Computational Linguistics: In Honour of Don Walker. Linguistica Computazionale, vol. 9, pp. 307–407. Springer, Dordrecht (1994). doi: 10.1007/978-0-585-35958-8_20, https://link.springer.com/chapter/10.1007/978-0-585-35958-8_20, last accessed 2021/04/28.

2. Hvelplund K.T. Digital Resources in the Translation Process – Attention, Cognitive Effort and Processing Flow // Perspective Studies in Translatology. Vol. 27, Issue 4. Taylor & Francis, 2019. P. 510–524.

3. Qin, Y., Shang, J., Lu, X.: The gap between NMT and professional translation from the perspective of discourse. In: ACM International Conference Proceeding Series. 3rd International Conference on Innovation in Artificial, pp. 50–54. Association for Computing Machinery, New York (2019). doi:10.1145/3319921.3319936, https://www.researchgate.net/publication/333183912, last accessed 2021/05/27.

4. Pushkin A.S. Evgeniy Onegin // Pushkin A.S. Sobranie sochineniy. V 8 t. T. 5. M.: Khudozh. lit., 1969. 320 s.

5. Pushkin A. Eugene Onegin. A Novel in Verse. Translated with an Introduction and Notes by James E. Falen. UK, Oxford World's Classics, Oxford, 1995. 288 p.

К вопросу об использовании моделей дистрибутивной семантики в исследованиях синонимических рядов в современном русском языке

Щукина К.А., СПбГУ, kira_a@list.ru

Последние годы на кафедре русского языка как иностранного Санкт-Петербургского государственного университета появляется значительное количество бакалаврских и магистерских работ, выполненных в рамках функционального подхода. Понятно, что при использовании этого подхода выпускники обращаются к Национальному корпусу русского языка, но, как нам представляется, используя при этом не все возможности данного ресурса и связанных с ним проектов-спутников, в частности, сервиса RusVectōrēs.

Возьмем для примера темы работ за последние несколько лет, посвященные синонимам и синонимическим рядам: «Синонимический ряд глаголов с доминантой «исчезнуть»: функционально- семантический аспект (на фоне корейского языка)», «Семантика и функционирование глагольных синонимов с доминантой «бояться» (на фоне китайского языка)», «Синонимический ряд прилагательных с доминантой «вежливый»: лексико-семантический, лингвокультурологический и лексикографический аспекты», «Синонимические ряды русских глаголов с доминантами «бездельничать» и «хитрить» на фоне вьетнамского языка: функционально-семантический аспект», «Семантика и функционирование русских глагольных синонимов с доминантой «убеждать» (на фоне китайского языка)», «Синонимический ряд русских прилагательных с доминантой «бездушный»: функционально-семантический аспект» и др. Материалом для подобного рода работ служат данные словарей синонимов и толковых словарей русского языка, а также материалы сайта «Национальный корпус русского языка». Традиционно по словарям определяется семантика, а по материалам корпуса выявляются особенности функционирования лексических единиц.

Нам представляется интересным некоторое расширение спектра работ в область дистрибутивной семантики, возможности которой явно недооцениваются в лингвистических исследованиях. Этому вопросу, в частности, посвящены две работы М.К. Тимофеевой: «Возможности использования сервиса RusVectōrēs для выявления семантических ассоциатов глаголов русского языка» и «Типология семантических отношений, выявляемых посредством инструмента RusVectōrēs» [Тимофеева, 2018а, 2018б], где автор предлагает классификацию семантических отношений и подчеркивает, что «сервис RusVectōrēs выявляет довольно большое количество синонимов, однако сравнение с данными словаря синонимов русского языка (…) показывает, что доля обнаруживаемых словарных синонимов не очень велика (17,61 %)» [Тимофеева, 2018а], а также отмечает, что «при идентификации семантической связи между входным словом и словом, выявленным посредством RusVectōrēs, рассматриваются следующие возможности: выявленное слово может быть по отношению к заданному слову синонимом, антонимом, гипонимом, гиперонимом, холонимом, меронимом, признаком, операцией, ситуационно связанным понятием, словообразовательным вариантом» [Тимофеева, 2018б]. Мы можем предположить, что в равной мере эта же информация может относиться и к Национальному корпусу русского языка, на материале которого, собственно, и работает RusVectōrēs (хотя, по нашим наблюдениям, данные несколько отличаются).

Возвращаясь к работам, выполненным на кафедре русского языка как иностранного СПбГУ, обратимся к выпускной квалификационной работе бакалавра Ю.В. Володы «Синонимический ряд прилагательных с доминантой «вежливый»: лексико-семантический, лингвокультурологический и лексикографический аспекты» [Волода, 2021]. В своей работе Ю.В. Волода анализирует данные словарей синонимов и на их основе составляет сводный синонимический ряд с доминантой «вежливый» из прилагательных, входящих в 3 и более словаря: *вежливый* (5), *учтивый* (5), *любезный* (5),

обходительный (5), *корректный* (4), *деликатный* (4), *галантный* (3), *предупредительный* (3). (там же) (см. Таблица 1).

Данные, полученные из Ю.В. Володой, мы сравнили данными сервисов «Похожие слова» Национального корпуса русского языка и проекта-спутника RusVectōrēs, в которых отображаются ближайшие семантические ассоциаты выбранного слова. В первом столбце таблицы мы приводим данные из НКРЯ, ранжировав их по значению коэффициента семантической близости.

Таблица 1. Синонимический ряд с доминантой «вежливый» в сравнении с корпусными данными

«Похожие слова» (НКРЯ)	**«Похожие слова» (RusVectōrēs по НКРЯ)**	**Данные Ю.В. Володы**
учтивый (0.884902)	учтивый 0.69	учтивый (5)
обходительный (0.734349)	обходительный 0.61	обходительный (5)
тактичный (0.714415)	тактичный 0.64	тактичный (Гаврилова)[2]
приветливый (0.709105)	приветливый 0.61	приветливый (Гаврилова)
предупредительный (0.708281)	предупредительный 0.61	предупредительный (3)
-	деликатный 0.61	деликатный (4),
дружелюбный (0,70118)	дружелюбный 0.60	отсутствует в словарях синонимов
доброжелательный (0,684435)	доброжелательный 0.59	отсутствует в словарях синонимов
-	корректный 0.59	корректный (4)
-	сдержанный 0.59	отсутствует в словарях синонимов
-	-	любезный (5)
-	-	галантный (3)

Из таблицы видно, что данные словарей лишь отчасти совпадают с корпусными данными, так, в частности, лексемы «тактичный» и «приветливый» встречаются лишь в Словаре синонимов и антонимов А.Н. Гавриловой, лексема «деликатный» есть в 4 словарях синонимов, но НКРЯ не показывает ее семантическую близость со словом «вежливый». С лексемами «любезный» и «галантный», на наш взгляд, ситуация немного другая, поскольку для «галантный» словари выдают помету «устаревший»; для лексемы «любезный» такой пометы нет, но, с нашей точки зрения, в современном русском языке она постепенно переходит в разряд устаревших, что и подтверждается корпусными данными. Лексемы «дружелюбный» и «доброжелательный» с точки зрения словарей не являются синонимами слова «вежливый», и тут, вероятно, стоит говорить о других типах семантических отношений, вероятнее всего, о гипонимии, что, разумеется, требует дальнейших исследований.

Таким образом, можно сказать, что при анализе данных словарей и сопоставлении их с корпусными данными мы можем получить довольно интересные результаты, отражающие состояние современного русского языка не только в отношении прилагательных, взятых

2 Гаврилова — здесь и далее: Гаврилова, А.С. Словарь синонимов и антонимов современного русского языка. 50000 слов / Под ред. А.С. Гавриловой // М.:«Аделант», 2014. – 800 с.

нами для примера, но и для других частей речи. Понятно, что корпусные результаты сгенерированы автоматически и в них возможно появление ошибок и неточностей; тем не менее, исследования, задействующие методы и модели дистрибутивной семантики, представляются нам перспективными.

Список литературы

1. Володa Ю.В. Синонимический ряд прилагательных с доминантой «вежливый»: лексико-семантический, лингвокультурологический и лексикографический аспекты. Научный руководитель: Е.И. Зиновьева. СПбГУ, 2021 URL: https://dspace.spbu.ru/bitstream/11701/29971/1/VKR_Voloda_U.V.docx (дата обращения: 07.07.2024)

2. Тимофеева М.К. Возможности использования сервиса RusVectōrēs для выявления семантических ассоциатов глаголов русского языка // Научный диалог. 2018. №9. URL: https://cyberleninka.ru/article/n/vozmozhnosti-ispolzovaniya-servisa-rusvect-r-s-dlya-vyyavleniya-semanticheskih-assotsiatov-glagolov-russkogo-yazyka (дата обращения: 07.07.2024) - 2018а

3. Тимофеева М.К. Типология семантических отношений, выявляемых посредством инструмента RusVectōrēs // Научный диалог. 2018. №8. URL: https://cyberleninka.ru/article/n/tipologiya-semanticheskih-otnosheniy-vyyavlyaemyh-posredstvom-instrumenta-rusvect-r-s (дата обращения: 07.07.2024) - 2018б