Е. В. Рахилина, В. П. Фесенко
E. V. Rakhilina, V. P. Fesenko

ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ КОРПУСА ТЕКСТОВ XIX ВЕКА В ЛИНГВИСТИЧЕСКОМ ИССЛЕДОВАНИИ[1]

OPPORTUNITIES FOR CORPUS OF XIX CENTURY TEXTS USAGE IN LINGUISTIC RESEARCH

Аннотация. Доклад посвящен Корпусу текстов XIX века, созданному в Школе лингвистики НИУ ВШЭ. Корпус специальным образом размечен и содержит информацию о лингвистических единицах, которые изменили значение или форму в период с XIX до XXI век. Рассматривается возможность применения данных корпуса для микродиахронических исследований в области синтаксиса (в частности, проблемы выбора родительного/ винительного падежа объекта при переходном глаголе с отрицанием), а также небольших лексикосемантических сдвигов (ср. *не уметь, нездоров*) и др.
Ключевые слова. Корпус, XIX век, отрицание, изменение семантики, изменение сочетаемости, историческая грамматика, русский язык.

Abstract. The paper describes the Corpus of XIX-th century texts (School of Linguistics, RSU HSE). The Corpus has a sophisticated system of annotation and contains information about linguistic items that have changed their meaning and/or form since XIX-th century. The paper discusses how the Corpus can be used as an additional linguistic tool for microdiachronicre- search on syntactic problems (like genitive of negation) and minor lexical semantic shifts (cf. *ne umet'* 'to be hopeless at doing smth', *nezdorov* 'healthless').
Keywords. Corpus, XIX century, negation, semantic change, historical grammar, Russian.

Проект «Корпус текстов XIX века» стартовал в сентябре 2016 года на базе Школы лингвистики НИУ ВШЭ под руководством Е. В. Рахилиной. Объектом разметки Корпуса текстов XIX века являются морфологические, лексические и синтаксические конструкции, обнаруживающие различия между нормами русского языка XIX и XXI веков. Цель создания корпуса — составление базы данных текстов XIX века и разметка в них языковых единиц разного рода, отличающих языковые нормы XIX и XXI веков. Данные корпуса могут быть применены для поиска и описания этих единиц, включая конструкции. Ближайшим результатом работы над корпусом может стать составление списка устаревших конструкций XIX века, который в дальнейшем может лечь в основу соответствующего словаря. Информация из корпуса мо-

[1] При поддержке гранта РНФ 16-18-02071 «Пограничный русский: оценка сложности восприятия русского текста в теоретическом, экспериментальном и статистическом аспектах».

жет использоваться как для научного описания нормы русского языка в диахронии, так и как помощь в понимании текстов, написанных два века назад.

Составление базы началось с разметки текста романа М. Ю. Лермонтова «Герой нашего времени» (объем — 43 566 слов). На сегодняшний день в корпусе размечены или находятся на стадии разработки роман И. А. Гончарова «Обыкновенная история» (97 868), повесть И. С. Тургенева «Ася» (13 931) и рассказ «Мой сосед Радилов» (2 289). Помимо художественных, база корпуса будет включать тексты других стилей.

С помощью корпуса возможен поиск по типам конструкций, изменившихся с XIX века, а также по конкретным словам и конструкциям. Такой поиск позволяет находить примеры для исследования языковых конструкций, а также отмечать новые лингвистические сюжеты для изучения истории языка.

Иллюстративным материалом к докладу послужит использование Корпуса текстов XIX века для решения комплекса вопросов, связанных с отрицанием:

— проблема объектного генитива при отрицании;
— изменение функционирования сочетания *не умеет;*
— изменение семантики краткого прилагательного *нездоров.*

1. Выбор падежа объекта при отрицании

Наиболее известная проблема русистики, связанная с отрицанием, — выбор падежа объекта при переходном глаголе с отрицанием. Об отклонениях от существовавшей с XVIII века нормы замены винительного падежа на родительный при отрицании с научной точки зрения первым начать говорить А. И. Томсон в начале XX века [Томсон 1902]. После этого на протяжении XX века и до сего дня она привлекает внимание теоретиков лингвистики (см., например, [Борщев, Парти 1998; Объектный генитив 2008]). Однако о неполном соответствии грамматической нормы выбору падежа в реальной речи писал еще А. С. Пушкин в первой половине XIX века (см. [Винокур 1959]). Корпус текстов XIX века с разметкой конструкций, отличающихся от принятых в современном русском языке, позволяет по-новому подойти к исследованию этой проблемы.

Поиск в Корпусе текстов XIX века по помете *genneg* дает конструкции с родительным падежом при отрицании, в которых современный

носитель языка выбрал бы винительный падеж. Например: *Она покраснела и не хотела назвать дня, вспомнив свою милую выходку* («Княжна Мери»). В современных исследованиях объект *день* рассматривается как референтный и требует винительного падежа. В докладе подробно, с привлечением материала НКРЯ, обсуждается процесс изменения именных ограничений на эту отрицательную конструкцию, связанных с референциальным статусом существительного.

2. Сочетаемость выражения *не уметь*

Помимо классической проблемы падежа при отрицании, интересную динамику показывает конструкция *не уметь*. Поиск в Корпусе текстов XIX века дает следующие контексты, отличающиеся от современных: *Ведь этакой народ! — сказал он, —* и *хлеба по-русски назвать не умеет, а выучил: «офицер, дай на водку!»* («Бэла»); *Так иногда отличный анатомик не умеет вылечить от лихорадки; Он изучал все живые струны сердца человеческого, как изучают жилы трупа, но никогда не умел он воспользоваться своим знанием; — Вы также переменились, — отвечала она, бросив на него быстрый взгляд, в котором он не умел разобрать тайной насмешки* («Княжна Мери»). Можно заметить, что все зависящие от *не умеет* глаголы — совершенного вида. Дальнейший анализ примеров из НКРЯ показывает, что в XIX веке доля примеров с сочетанием *не уметь + уСВ* значительно выше, чем в XXI веке: в среднем 43 % из всех сочетаний *не уметь + V*. В XXI веке этот средний показатель равен 14 %. Притом эти 14 % примеров принадлежат авторам, родившимся не позднее 1969 года, воспитанным на литературе более раннего периода.

Любопытно отметить, что в XIX веке достаточно частотно использование в качестве зависимого от *не уметь* слова существительного *грамота* в дательном падеже, что сегодня соответствует сочетанию *не обучен*: *он не умеет грамоте;* есть также маргинальный контекст с предложным управлением: *не умеет в грамоте*. Кроме того, в XIX веке встречаются контексты с зависимым от *не уметь* придаточным изъяснительным предложением, равным сегодняшнему *не знать*: *не умел с чего начать; не умею, как вас назвать; что сказать, он не умеет* и пр.

Примечательно также, что в XIX веке было возможно опущение (с точки зрения современной нормы) глагола *говорить* при *не уметь* в конструкции со значением владения иностранным *языком:* не умеет по-немецки, по-барски не умею, не умел по-русски ни слова.

Обсуждается, как в дальнейшем общее значение слова *уметь* распределилось между глаголами *мочь, знать* и *быть обученным.*

3. Семантика краткого прилагательного *нездоров*

Интересным объектом микроисторического исследования может быть краткое прилагательное *нездоров.* В Корпусе XIX века встречаем следующий пример: *Печорин был долго нездоров, исхудал, бедняжка; только никогда с этих пор мы не говорили о Бэле: я видел, что это ему будет неприятно, так зачем же?* («Бэла»). В этом примере *был долго нездоров* «в переводе» на современный русский язык означает *долго болел.*

Анализ примеров НКРЯ показывает, что в XIX веке *нездоров* было полным синонимом слова *болен.* Так можно было сказать и о человеке, у которого разболелась голова или начался насморк, но и о человеке, страдающем оспой, лежащем на смертном одре или раненом. Допустимым было сочетание *нездоров* с наречиями *очень, сильно.* Любопытно, что слова *нездоров* и *болен* не различались по своей семантико-синтаксической сочетаемости с существительным: *болен простудой* и *нездоров флюсом, нездоров сильным кашлем.*

В XXI веке *нездоров* встречается в двух основных контекстах. Прежде всего, это небольшое недомогание: *Но я нынче нездоров, Мне что-то тяжело, пойду засну.* [Музыкальные паузы режиссера Васильева (2003) // «Театральная жизнь», 2003.05.26]; *Элмар был нездоров и хотел отдохнуть…* [Армен Медведев. Территория кино (1999–2001)]; *Он слегка нездоров и расстроен…* [Александр Проханов. Господин Гексоген (2001)].

Второй контекст, в котором сегодня используется *нездоров,* — психиатрический: *психически нездоров, нездоров душевно.*

Другими словами, у краткого прилагательного *нездоров* ушло значение тяжелой физической болезни, раны и под.

Как видно из приведенных примеров, разметка Корпуса текстов XIX века может быть очень полезной для выявления и дополнительного исследования «проблемных точек» в языке XIX века — тех единиц и конструкций, семантика или функционирование которых в той или иной степени отличаются от современных. Дальнейший, более глубокий анализ этих отличий может быть проведен с помощью обращения к более широкому материалу Национального корпуса русского языка.

Литература

1. *Борщев В. Б., Парти Б.* (1998), Бытийные предложения и отрицание в русском языке: семантика и коммуникативная структура // Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям (под ред. А. С. Нариньяни). Казань, Хетер, с. 173–182.
2. *Винокур Г. О.* (1959) Пушкин и русский язык (Сб. «А. С. Пушкин. 1837–1937», М., 1937) // Избранные работы по русскому языку. М.: Учпедгиз, с. 189–206.
3. Объектный генитив при отрицании в русском языке (2008) [Ред. кол.: А. Б. Летучий, Е. В. Рахилина, Т. И. Резникова; Сост. Е. В. Рахилина]. М.: Пробел — 2000. 176 с. (Исследования по теории грамматики; Вып. 5).
4. *Томсон А. И.* (1902), Винительный падеж прямого дополнения в отрицательных предложениях в русском языке. Отд. отт. из «Русского Филологического Вестника». Варшава: Типография Варшавского Учебного Округа. 43 с.

References

1. *Borshchev V. B., Partee B.* (1998), Bytijnye predlozhenija i otricanie v russkom jazyke: semantika i kommunikativnaja struktura. [Existential sentences and negation in Russian: semantics and communicative structure]. In: Trudy Mezhdunarodnogo seminara Dialog'98 po komp'juternoj lingvistike i ee prilozhenijam (pod red. A. S. Narin'jani). [Incoll. Proceedings of the International seminar on computer linguistics and its applications Dialogue'98 (ed. A. S. Narin'jani)]. Kazan', Heter, pp. 173–182.
2. *Vinokur G. O.* (1959) Pushkin i russkij jazyk (Sb. «A. S. Pushkin. 1837–1937», M., 1937) [Pushkin and Russian language (Coll. "A. S. Pushkin. 1837–1937", Moscow, 1937)]. In: Izbrannye raboty po russkomu jazyku [Selected works on Russian language]. Moscow: Uchpedgiz, pp. 189–206.
3. *Ob ektnyj genitive pri otricanii v russkom jazyke* (2008) [Object-Genitive under Negation in Russian (2008)]. Red. kol.: A. B. Letuchij, E. V. Rahilina, T. I. Reznikova; Sost. E. V. Rahilina [Ed. staff: A. B. Letuchij, E. V. Rahilina, T. I. Reznikova; Comp. E. V. Rahilina]. Moscow: Probel — 2000, 176 p. (Issledovanija po teorii grammatiki; Vyp. 5). [Research on theory of grammar; edition 5].
4. *Tomson A. I.* (1902), Vinitel'nyj padezh prjamogo dopolnenija v otricatel'nyh predlozhenijah v russkom jazyke [Accusative case of direct object in Russian]. Otd. ott. iz «Russkogo Filologicheskogo Vestnika» [Separate print from "Russian Filological Herald"]. Varshava: Tipografija Varshavskogo Uchebnogo Okruga. 43 p.

Рахилина Екатерина Владимировна
Rakhilina Ekaterina
E-mail: rakhilina@gmail.com

Фесенко Вера Павловна
Fesenko Vera
E-mail: verun4ik_18@mail.ru

Национальный исследовательский университет
«Высшая школа экономики» (Россия)
National Research University Higher School of Economy (Russia)

Е. И. Риехакайнен
E. I. Riekhakaynen

КОРПУС ТРАСКРИБИРОВАННЫХ РУССКИХ УСТНЫХ ТЕКСТОВ: ТЕКУЩИЕ ВОЗМОЖНОСТИ И ПЕРСПЕКТИВЫ

CORPUS OF TRANSCRIBED RUSSIAN SPEECH: CURRENT OPTIONS AND CHALLENGES

Аннотация. В статье на примере исследования, посвященного оценке сохранности фонетического облика словоформ в начале межпаузальных интервалов в русской спонтанной речи, демонстрируются возможности Корпуса транскрибированных русских устных текстов и намечаются перспективы его дальнейшего развития, наиболее актуальные из которых (помимо расширения Корпуса) – это создание новых уровней аннотирования (в частности, уровня идеальной транскрипции) и добавление новых поисковых возможностей.
Ключевые слова. Корпусы устной речи, русская устная речь, редукция, восприятие речи.

Abstract. The article describes a phonetic study of word forms at the beginning of interpausal intervals in spontaneous Russian that allowed to demonstrate how the Corpus of Transcribed Russian Speech can currently be used and developed. For the moment, new levels of annotation (such as ideal transcription) and new search options are the most required innovations.
Keywords. Spoken corpora, Russian speech, reduction, spoken word recognition.

1. Корпус транскрибированных русских устных текстов

Статья продолжает серию публикаций, посвященных принципам создания и использования Корпуса транскрибированных русских устных текстов (далее — Корпус), разрабатываемого сотрудниками Санкт-Петербургского государственного университета. В предыдущих статьях ([Венцов и др. 2013; Венцов и др. 2015] и др.) мы использовали различные варианты названия нашего корпуса — Корпус русской устной речи, Речевой корпус, Корпус русских спонтанных текстов. Однако, как представляется, именно формулировка «Корпус транскрибированных русских устных текстов» наилучшим образом отражает как его содержание (в нем в настоящее время представлены расшифровки не только спонтанной речи, но и, например, радиосводок Ю. Б. Левитана), так и особенности аннотирования (Корпус до сих пор остается единственным известным нам общедоступным корпусом русской речи, снабженным полной фонетической транскрипцией)[1].

Основная цель создания Корпуса — его последующее применение для моделирования восприятия естественной звучащей речи. Именно

[1] Более подробное сопоставление Корпуса с другими корпусами русской устной речи представлено в [Венцов и др. 2013: 224].

этим обусловлены те принципы аннотирования (включая обязательное наличие сплошной фонетической расшифровки), которые приняты в Корпусе (о необходимости столь подробного аннотирования см., например, в [Tucker et al. 2016]). Далее в статье будут упомянуты только те принципы аннотирования, которые являются первостепенными для исследования, о котором пойдет речь в Разделе 2.

2. Исследование сохранности фонетического облика словоформ в начале межпаузальных интервалов

2.1. Гипотеза

Экспериментальные свидетельства в пользу того, что контекст является ведущим фактором при распознавании редуцированных словоформ в естественной русской речи (см. [Риехакайнен 2016]), а также тот факт, что в момент восприятия конкретного фрагмента устной речи слушающему доступен только левый контекст, позволяют предположить, что в начальных фрагментах речевой цепи редукция должна встречаться реже, поскольку отсутствует непосредственный левый контекст, за счет которого могли бы быть восстановлены редуцированные элементы.

2.2. Методика исследования

Сформулированную выше гипотезу решено было проверить на материале одной из записей, входящих в Корпус, а именно на материале расшифровок радиопередачи «Утренний гость» (при создании конкорданса при поисковом запросе в онлайн-версии Корпуса фрагменты этого текста обозначаются guest01.wav).

Во всех текстах Корпуса на данный момент осуществлена разметка на межпаузальные интервалы, при этом различаются паузы вдоха (inh), паузы вздоха (sigh), придыхание (aspir), гортанная смычка (gst(e)) и собственно паузы (paus(e) или p). Поиск по словам (в орфографии и в транскрипции) и по типам пауз доступен по адресу: http://narusco.ru/search/trn-search.php.

Из Корпуса были отобраны и проанализированы все фонетические слова, которые были употреблены непосредственно после любой из пауз. В распоряжении исследователей[2] при этом была расшифровка

[2] Отбор и первичную обработку результатов осуществлял студент СПбГУ Александр Сергеевич Смирнов под руководством автора статьи.

в виде связного текста, но поставленную задачу можно решить и с помощью онлайн-поиска, задавая последовательно все виды пауз и отбирая контексты из записей, обозначенных guest01.wav.

В выборку вошли 579 контекстов. Дальнейшее исследование заключалось в анализе сохранности фонетического облика начальных фонетических слов в каждом из случаев.

2.3. Результаты

В целом полученные результаты свидетельствуют в пользу подтверждения сформулированной гипотезы. Только в 169 (29,2 %) из 579 контекстов, попавших в выборку, было зафиксировано выпадение или изменение звуков в фонетическом слове в начале межпаузального интервала, т. е. сразу после паузы. При этом в 142 из них редукции подвергались не более двух звуков. Подавляющее большинство из этих примеров — это словоформы с редукцией на конце слов или в заударной флексии. Такие реализации могут быть однозначно восстановлены до полных или благодаря сохранности большей части словоформы (например, *которую*[3]*, украденную* и др.), или благодаря ближайшему правому контексту, в котором содержится информация, указывающая на морфологические признаки словоформы (например, *вчерашний ребёнок и* др.), т. е. левый контекст не играет в распознавании подобных реализаций ключевой роли.

Уже на данном этапе описания результатов возникают методологические вопросы: что считать идеальными произнесением и как оценить полученный процент редуцированных реализаций в начале межпаузальных интервалов? Первый вопрос в рамках данного исследования был решен самым простым способом — конкретные реализации сопоставлялись с максимально полным произнесением, однако очевидно, что в ряде случаев даже прескриптивная орфоэпическая норма будет отличаться от такого варианта произнесения (см. подробнее об этом в [Риехакайнен 2016: 72]). Чтобы каждому конкретному исследователю в дальнейшем не приходилось сталкиваться с этой проблемой, предполагается в будущем дополнить аннотирование Корпуса уровнем идеальной транскрипции, которая будет создана автоматически на основе орфографической расшифровки текста. Что касается второго вопроса, то для того, чтобы определить, что сло-

[3] Жирным шрифтом здесь и далее в примерах выделены звуки, подвергшиеся полной количественной редукции.

воформы в начале межпаузальных интервалов действительно менее склонны к редукции, чем в других позициях, необходимо знать статистику количества редуцированных словоформ по всему Корпусу. На настоящее время эти подсчеты можно провести вручную, однако создание уровня идеальной транскрипции существенно упросит решение этого вопроса.

Подробный анализ тех 27 случаев, в которых выпадению или качественному изменению подвергались более двух элементов, представлен в [Смирнов 2017], в рамках же данной статьи будет упомянута только одна группа из них, а именно те случаи, когда для надежного распознавания редуцированной словоформы требовалось обращение к левому контексту, который предшествовал паузе (например, *работ**ают***, *Влади**мирович***). Такие примеры в очередной раз демонстрируют, что паузы в спонтанной речи могут возникать не только на границе клауз (синтагм), но и внутри них. В результате поиска по всем типам пауз в выборку попали оба типа случаев. Сформулированная же выше гипотеза должна быть подтверждена прежде всего на примерах первого типа (т. е. на границах клауз), поскольку в случае т.н. «расчлененных» синтагм единство элементов, их образующих, может достигаться вопреки паузам — за счет грамматических средств. Благодаря проведенным ранее исследованиям и наличию Базы «расчлененных» дискурсивных единиц (http://narusco.ru/EDU_BASE) мы имели данные о границах клауз в анализируемом тексте. В дальнейшем подобная информация может быть также интегрирована в основную разметку Корпуса.

Поскольку во многих из проанализированных случаев редукция затронула окончания (в первую очередь прилагательных и глаголов), следующим шагом в исследовании распознавания восприятия речи должно стать моделирование того, каким образом слушающий восстанавливает редуцированную морфологическую информацию. Однако для исследований подобного рода необходимо иметь возможность поиска по различным морфологическим параметрам или хотя бы по частям речи, для чего необходимо аннотирование Корпуса на морфологическом уровне.

Таким образом, проведенное исследование показало, каким образом можно использовать уже имеющиеся возможности Корпуса, а также позволило наметить наиболее актуальные пути развития Корпуса, которые, как мы надеемся, будут воплощены в жизнь в ближайшее время.

Литература

1. *Венцов А. В., Нигматулина Ю. О., Раева О. В., Риехакайнен Е. И., Слепокурова Н. А.* (2013), Корпус русских спонтанных текстов: структура и единицы // Корпусная лингвистика — 2013: Труды международной научной конференции. СПб., с. 223–230.
2. *Венцов А. В., Нигматулина Ю. О., Раева О. В., Риехакайнен Е. И., Слепокурова Н. А.* (2015), От корпуса устной речи к базе «расчлененных» дискурсивных единиц // Корпусная лингвистика — 2015: Труды международной научной конференции. СПб., с. 154–161.
3. *Риехакайнен Е. И.* (2016), Восприятие русской устной речи: контекст + частотность. СПб.
4. *Смирнов А. С.* (2017), Без контекста: редукция в начале межпаузальных интервалов // Проблемы порождения и восприятия речи. Материалы XIV выездной школы-семинара. Череповец, в печати.
5. *Tucker B., Ernestus M.* (2016), Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon // The Mental Lexicon, 11 (3), pp. 375–400.

References

1. *Ventsov A. V., Nigmatulina Yu.O., Raeva O. V., Riekhakaynen E. I., Slepokurova N. A.* (2013), Korpus russkih spontannyh tekstov: struktura i edinitsy [Corpus of Russian spontaneous texts: structure and items]. In: Korpusnaja lingvistika — 2013: Trudy mezhdunarodnoj nauchnoj konferentsii. [Corpus linguistics — 2013: Proceedings of the international scientific conference]. St Petersburg, pp. 223–230.
2. *Ventsov A. V., Nigmatulina Yu.O., Raeva O. V., Riekhakaynen E. I., Slepokurova N. A.* (2015), Ot korpusa ustnoj rechi k baze "raschlenennyh" diskursivnyh edinits [From a speech corpus to a database of "broken" discourse units]. In: Korpusnaja lingvistika — 2015: Trudy mezhdunarodnoj nauchnoj konferentsii. [Corpus linguistics — 2015: Proceedings of the international scientific conference]. St Petersburg, pp. 154–161.
3. *Riekhakaynen E. I.* (2016), Vosprijatie russkoj ustnoj rechi: kontekst + chastotnost' [Recognition of Russian speech: context + frequency]. St Petersburg.
4. *Smirnov A. S.* (2017), Bez konteksta: reduktsija v nachale mezhpauzalnyh intervalov [Without context: reduction in the beginning of interpausal intervals]. In: Problemy porozhdenija i vosprijatija rechi. Materialy XIV vyezdnoj shkoly-seminara [Problems of speech production and recognitions. Proceedings of the 14th workshop]. Cherepovets, to appear.
5. *Tucker B., Ernestus M.* (2016), Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. In: The Mental Lexicon, 11 (3), pp. 375–400.

Риехакайнен Елена Игоревна
Санкт-Петербургский государственный университет (Россия)
Riekhakaynen Elena
Saint Petersburg State University (Russia)
E-mail: e.riehakajnen@spbu.ru