А. В. Венцов, Ю. О. Нигматулина, О. В. Раева,
Е. И. Риехакайнен, Н. А. Слепокурова
A. V. Ventsov, Y. O. Nigmatulina, O. V. Raeva,
E. I. Riekhakaynen, N. A. Slepokurova

КОРПУС РУССКИХ СПОНТАННЫХ ТЕКСТОВ: СТРУКТУРА И ЕДИНИЦЫ

CORPUS OF SPONTANEOUS RUSSIAN TEXTS: STRUCTURE AND ITEMS

Аннотация. В докладе описывается Корпус русских спонтанных текстов, создаваемый с целью получения материала для разработки функциональной модели восприятия речи человеком. Словоформы, на границах которых в спонтанной речи происходит стяжение звуков, рассматриваются как кандидаты в нечленимые единицы в составе Корпуса и Частотного словаря, созданного на его основе.

Abstract. The Corpus of spontaneous Russian texts described in the paper is being created as a source of material for the functional model of spoken word recognition. If the last sound of a word form in the Corpus is contracted to the first sound of the next one, such word forms are considered to be an indivisible item both in the Corpus and in the Frequency Word List formed automatically on the basis of the Corpus.

Введение

Для разработки функциональной модели восприятия речи необходимо изучить особенности того сигнала, с которым слушающий сталкивается при восприятии речи в естественных условиях. В перцептивных исследованиях для получения подобных сведений все чаще используются корпусы спонтанных текстов[1]. Представляется, что в таком корпусе, наряду с орфографической

[1] См., например: *Ernestus M., Baayen H., Schreuder R.* The Recognition of Reduced Word Forms // Brain and Language. 2002. Vol. 81 (1-3). P. 162-173; *Brouwer S.* Processing Strongly Reduced Forms in Casual Speech : Theses for the Degree of Ph.D. Nijmegen, 2010 и др.

расшифровкой записей, должна быть обязательно представлена и акустико-фонетическая транскрипция, которая может быть принята в качестве результата обработки речевого сигнала начальными уровнями слуховой системы, а также позволит определить, как именно реализуются те или иные единицы в потоке речи.

К сожалению, ни в одном из доступных корпусов устной русской речи не представлена сплошная фонетическая расшифровка записей. В устном и мультимедийном подкорпусах Национального корпуса русского языка (далее – НКРЯ)[2] дана только орфографическая расшифровка. Корпусы, созданные в рамках проекта «Рассказы о сновидениях и другие корпуса звучащей речи»[3], снабжены дискурсивной транскрипцией, которая описывает лишь просодические особенности текстов и такие нюансы произнесения, как «губные смычки <…>, придыхание, ускоренное произнесение, сниженный регистр и т.д.»[4]. В корпусе «Один речевой день» фонетическая расшифровка на данный момент осуществлена лишь выборочно, кроме того, этот корпус не является общедоступным (хотя часть материалов данного корпуса представлена в НКРЯ)[5]. Как следствие, эти корпусы не могут использоваться для проведения экспериментальных исследований в области восприятия речи и построения функциональной модели механизмов распознавания человеком естественной речи без их предварительной доработки.

---

[2] URL: **http://ruscorpora.ru/**

[3] URL: **http://spokencorpora.ru/**

[4] О дискурсивной транскрипции // Рассказы о сновидениях и другие корпуса звучащей речи. URL: **http://spokencorpora.ru/showtranshelp.ru.** Дата обращения: 14.04.2013.

[5] См., например, *Богданова Н.В. и др.* Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М., 2010. С. 41–47.

Целью создания Корпуса русских спонтанных текстов, который представлен в докладе, является в первую очередь получение материала для разработки возможных алгоритмов преобразования непрерывного речевого акустического сигнала в линейную последовательность лексических единиц. Подобная формулировка цели обусловила наличие в корпусе не только орфографической, но и детальной фонетической транскрипции.

1. Описание Корпуса

1.1. Общая информация

В корпусе представлены расшифровки записей теле- и радиопередач, проводившихся в режиме спонтанной речи. Тексты продолжительностью звучания 224 минуты разбиты на межпаузальные отрезки и снабжены орфографической расшифровкой. Для фрагмента записи длительностью 90 минут (10 488 с/у) осуществлена также сплошная акустико-фонетическая расшифровка. Создание конкорданса по текстам корпуса, снабженным фонетической транскрипцией, доступно на сайте Корпуса русского литературного языка[6].

1.2. Принципы транскрибирования текстов

Транскрибирование осуществляется вручную опытными фонетистами – сотрудниками Лаборатории моделирования речевой деятельности и Кафедры общего языкознания Санкт-Петербургского государственного университета[7]. При транскрибировании эксперты последовательно рассматривают (на слух и с помощью динамических спектрограмм) короткие асемантичные фрагменты текста, что позволяет в максимальной степени опираться на акустические параметры речевого сигнала и исключить, насколько это возможно, влияние лексико-грамматической

---

[6] URL: **http://www.narusco.ru**

[7] Работа по расшифровке первых 90 минут звучания была выполнена в 2009–2011 гг. при поддержке гранта РФФИ №09-06-00244а.

информации. Для описания используется специальный набор символов, частично совпадающий с системой X-SAMPA, разработанной для компьютерного описания фонетических параметров речи[8].

1.3. Частотный словарь словоформ русской спонтанной речи

На базе той части корпуса, для которой имеется как орфографическая, так и фонетическая расшифровка, был создан «двуязычный» (орфографически-транскрипционный) частотный словарь, состоящий из 6651 строки – уникальной комбинации орфографического описания и фонетической транскрипции с указанием частоты встречаемости конкретной акустической реализации в текстах корпуса. Строки, имеющие одинаковое орфографическое описание, объединены в кластеры для удобства анализа вариантов произнесения одной и той же словоформы.

Например:

хорошая [xarúše] *1*
хорошая [xaróšы] *1*

2. Граница между словами в корпусе устных текстов[9]

Вопрос о границах слова в звучащем тексте и о целесообразности выделения т.н. фонетических слов неоднократно обсуждался в фонетической литературе и не имеет однозначного решения[10].

---

[8] Описание системы представлено на **http://www.narusco.ru/.** Далее в тексте для удобства читателей все примеры будут даны в транскрипции, предложенной Л.В. Щербой.

[9] Исследование, результаты которого представлены в данном разделе, осуществляется при поддержке Гранта Президента Российской Федерации для молодых российских ученых МК-3646-2013-6.

[10] См. обзор в: *Апушкина И.Е., Венцов А.В., Слепокурова Н.А.* О фонетическом слове // Анализ разговорной русской речи (АР^3–2008). Труды второго междисциплинарного семинара. СПб., 2008. С. 31–36.

Поскольку на материале Корпуса спонтанных текстов планировалось в первую очередь создать Частотный словарь словоформ, предполагающий получение списка всех возможных вариантов произнесения для каждого из орфографических слов, в ходе транскрибирования решено было считать словами то, что разделяется пробелами на письме.

Исключение было сделано лишь для некоторых т.н. составных слов («сочетаний, эквивалентных слову», например, *то_есть, потому_что* и др.), целесообразность выделения которых в качестве самостоятельных единиц в корпусе текстов была показана ранее на материале Корпуса русского литературного языка[11]. Для обозначения таких единиц в орфографической расшифровке используется знак «_». Как показал анализ имеющегося у нас материала, подобные сочетания демонстрируют свойства неделимого слова не только на грамматическом и семантическом уровнях, но и на фонетическом уровне: фонетическая граница между элементами подобных сочетаний как правило стирается. Например, *всё_равно [fs'o::rnó], то_есть [tes'], [tьıs']* и др.

Однако в ходе фонетического транскрибирования записей было замечено, что «стирание» границ между словами наблюдается не только между компонентами составных слов, но может затрагивать и любые другие словоформы, если на их стыке оказываются два гласных или два согласных звука. Например, *человек**а**_**о**бщаться [čəle+k**а**pš':a+cə]*. Данное явление, а именно слияние двух смежных звуков, в результате которого возникает один звук на месте двух прежних, в фонетике принято называть стяжением[12].

---

[11] *Венцов А.В., Грудева Е.В., Касевич В.Б., Ягунова Е.В.* Идиомы в Национальном корпусе русского литературного языка // Международная конференция «Корпусная лингвистика – 2004» Тезисы докладов (12–14 октября 2004 г., С.-Петербург). СПб, 2004. С. 17–18.

[12] *Розенталь Д.Э.* Справочник по русскому языку. Словарь лингвистических терминов. М., 2008. С. 418.

Чтобы оценить его масштаб, было проведено сравнение всех случаев соседства гласных и соседства согласных (одинаковых или различающихся по признаку глухости-звонкости и/или мягкости-твердости) на стыке словоформ в орфографической записи с транскрипцией данных фрагментов (составные слова при этом не учитывались). Оказалось, что «стягиваться» могут как гласные, так и согласные. Для получения более представительного материала была дополнительно рассмотрена запись, которая на данный момент представлена в Корпусе лишь в орфографической расшифровке (25 минут звучания). Таким образом, общий объем текста, использованного для сплошного поиска стяжений, составил 115 минут. При анализе учитывались спектральные характеристики звука/звуков на стыке словоформ: наличие стяжения констатировалось в том случае, когда на спектрограмме не было видно существенных формантных изменений, указывающих на наличие двух звуков.

Количество зафиксированных стяжений и их процент от общего числа сочетаний гласных и сочетаний согласных на стыке словоформ представлены в Таблице 1.

Таблица 1. Стяжения гласных и согласных в Корпусе русских спонтанных текстов

Стяжения	**Количество**	**Процент от общего числа проанализированных сочетаний V+V или C+C на стыке словоформ, %**
гласных	310	**55,5**
согласных	106	**74,1**

Можно констатировать, таким образом, что стяжение звуков на стыках словоформ является неотъемлемой составляющей русской спонтанной речи. Указанный масштаб этого явления засвидетельствован в литературе впервые.

Данное явление представляет большой интерес с точки зрения описания механизмов распознавания естественного речевого сигнала, поскольку размывает границы между словами и, соот-

ветственно, затрудняет сегментацию. Следовательно, сочетания со стяжениями на стыке должны быть предметом самостоятельного анализа при моделировании процессов восприятия речи человеком.

Поскольку проведение границы между словоформами, подвергшимися стяжению, в фонетической расшифровке не представляется возможным, такие единицы были также соединены в орфографической записи с помощью знака «_» и вошли в Частотный словарь словоформ спонтанной речи как самостоятельные единицы.

Таким образом, в Корпусе русских спонтанных текстов и в Частотном словаре словоформ, созданном на его основе, на лексическом уровне представлены три вида единиц: отдельные словоформы (*я, четыре* и т.д.), составные слова (*то_есть, всё_равно* и др.) и словоформы, подвергшиеся стяжению (*что_он [što̝n]/[šo̝n], дайте_им [dáet'im]* и др.).

3. Перспективы

В настоящий момент проводится детальное изучение обнаруженных в Корпусе сочетаний со стяжениями на стыке с целью выяснения стратегий, которыми может пользоваться слушающий для перцептивной сегментации подобных единиц.

Что касается дальнейшей разработки Корпуса русских спонтанных текстов в целом, то ведется работа по верификации созданной транскрипции, а также дальнейшая фонетическая расшифровка записей и пополнение общего объема корпуса, в том числе за счет привлечения записей подготовленной русской речи (дикторской речи, прочитанных текстов). Последнее позволит проверить, характерны ли явления, отмеченные в спонтанной речи, для устной речи в целом.