А.В. Венцов, Н.А. Слепокурова (Санкт-Петербург)

ОБ ОПЫТЕ СОЗДАНИЯ ФОНЕТИЧЕСКИ АННОТИРОВАННОГО КОРПУСА РУССКОЙ СПОНТАННОЙ РЕЧИ[1]

Создание функциональной модели восприятия речи человеком предполагает в качестве первого этапа разработку алгоритмов преобразования непрерывного речевого сигнала в дискретную последовательность адекватных лексических единиц. При этом алгоритм должен решать проблему «восстановления» редуцированных словоформ, особенно часто появляющихся в спонтанной речи.

Считается, что процесс такого преобразования включает операцию сегментации непрерывного речевого потока и идентификацию выделенных сегментов с лексическими единицами языка, которая требует обращения к внутреннему перцептивному словарю носителя языка [Kassevitch et al. 2000]. Использование этого словаря может обеспечивать и распознавание редуцированных словоформ.

Таким образом, возникает необходимость располагать информацией о содержимом перцептивного словаря, с одной стороны, и иметь четкое представление обо всех особенностях акустической реализации словоформ русского языка, с другой. Решить эту проблему можно только на основе анализа звучания словоформ русского языка в корпусе спонтанной устной речи и формирования на базе такого корпуса частотного словаря реализаций, отражающего статистику возможных вариантов. Последнее потребует создания для каждого звучащего текста корпуса двух параллельных описаний: орфографического и транскрипционного, предназначенного для отражения реальной звуковой картины. Следует отметить, что на сегодняшний день в доступных отечественных и зарубежных корпусах устной речи содержатся лишь орфографические описания звучащих текстов, фрагментарно дополняемые фонетической транскрипцией, полученной путем автоматического транскрибирования соответствующих орфографических описаний.

Алфавит используемых при слуховой транскрипции символов должен быть приспособлен к задачам последующей ком-

[1] Работа выполнена при поддержке гранта РФФИ № 09-06-00244-а.

пьютерной обработки. Для этой цели плохо подходит система символов IPA, поэтому для решения наших задач был специально сформирован упрощенный алфавит, частично совпадающий с системой X-SAMPA, разработанной именно для компьютерного описания фонетических параметров речевого сигнала.

Наш корпус состоит из оцифрованных спонтанных диалогических текстов общей длительностью звучания около 2-х часов, сегментированных на паузы и отрезки непрерывного «говорения», снабженные орфографическим описанием.

Для одного из текстов этого корпуса длительностью около 15 минут к орфографическому описанию отрезков «говорения» было добавлено их транскрипционное описание с использованием принятой системы символов. Транскрибирование производилось опытными экспертами на слух с одновременным использованием динамических спектрограмм, синхронизированных на экране компьютера с анализируемым участком звукового файла. Чтобы избежать возможного влияния лексики на решения экспертов, идентификация сегментов речевого сигнала осуществлялась на интервалах, не превышавших длительности слога.

Ударность гласных определялась путем субъективного сопоставления степени выделенности слогов в последовательных парах цепочки слогов, образующих анализируемый отрезок речевого сигнала: 1-го со 2-м, 2-го с 3-м, 3-го с 4-м, и так далее. Решение об ударности принималось в том случае, если очередной слог оказывался более выделенным, чем предшествующий и последующий.

В результате был получен файл-описатель проанализированного фрагмента звучащего текста, содержащий одновременно орфографическую и транскрипционную записи каждого отрезка речи между паузами. Это позволило создать пилотный «двуязычный» (орфографически-транскрипционный) частотный словарь для 2209 словоупотреблений, содержащихся в исследованном тексте. Полученный словарь подтвердил надежность принятой системы транскрибирования для решения поставленной задачи и впервые дал возможность статистической оценки редукции словоформ в спонтанной речи.