# КОРПУС РУССКИХ ТЕКСТОВ И МОДЕЛЬ ВОСПРИЯТИЯ РЕЧИ

А. В. Венцов, В. Б. Касевич, Е. В. Ягунова

В современной лингвистике укрепляется тенденция основывать любое исследование языковой и речевой действительности с использованием репрезентативного множества текстов. Такое множество текстов, организованное определенным образом, стало традиционным называть корпусом.

В последние десятилетия усилия лингвистов многих стран направлены на создание национальных, универсальных корпусов. При отсутствии точных критериев репрезентативности такого корпуса ясно общая задача: корпус должен обладать такими количественными и качественными параметрами, которые окажутся необходимыми и достаточными для построения на его основе адекватных словаря и грамматики языка.

Ни в коей мере не отрицая единства грамматического механизма на некотором уровне, мы полагаем необходимым выделять грамматику, отвечающую за порождение речи, и грамматику, отвечающую за восприятие речи, как две относительно самостоятельные грамматики. Более того, в этом различении, восходящем к идеям Л.В.Щербы об активной и пассивной грамматике, мы последовательно разграничиваем не только грамматики, но и словари: словарь, обслуживающий порождение речи, и словарь, обслуживающий восприятие речи (в дальнейшем – словарь для говорящего и словарь для слушающего). Именно последний, рассматриваемый как компонент модели восприятия речи, будет обсуждаться в настоящем докладе.

Прежде всего, повторим наши основные аргументы в пользу признания относительной самостоятельности словаря для слушающего (Венцов, Касевич 1998). Яркой отличительной особенностью этого словаря нам представляется характер его единицы: в качестве таковой мы полагаем словоформу (с исключением порождающих ее грамматических (морфологических) правил).

Можно считать экспериментально доказанным, что при восприятии (идентификации) слова (как изолированного, так и в тексте) человек опирается на частотность данного слова. Однако реальной частотностью характеризуются именно словоформы, причем разные формы одного и того же слова могут существенно различаться по частотности.

Экспериментально доказано и то, что перцептивно значим такой признак слова, как его акцентный контур. Акцентный контур еще более непосредственным и очевидным образом, чем частотность, есть признак словоформы, но никак не лексемы.

Выдвижение словоформы на роль основной единицы словаря для слушающего, разумеется, приводит к значительному увеличению его объема. В то же время опора на словоформу сильно упрощает процедуру идентификации единиц текста при их восприятии, во многом сводя эту процедуру к непосредственному сличению отрезка текста и единицы словаря, т.е. минуя дополнительную процедуру лемматизации, неизбежную при работе с традиционным словарем лексем.

Возникает дополнительная проблема. Как не раз было демонстрировано, в том числе авторами проекта, акцентный контур слова – один из основных признаков для первичной классификации слов при их восприятии. Приходится учитывать, однако, что акцентный контур текстовой единицы характеризует *фонетическое* слово, а не слово с семантико-грамматической точки зрения. Возникает проблема: либо перцептивный словарь представляет собой инвентарь (систему) не просто словоформ, но *фонетических* слов-словоформ, либо существуют алгоритмы перехода от фонетического слова к семантико-грамматическому (последние могут различаться довольно существенно, ср. одно и то же высказывание в терминах фонетических и семантико-грамматических слов: *тыбы лучше ейже ирассказал* и *ты бы лучше ей же и рассказал*).

В докладе излагаются некоторые результаты работы над проектом, основная цель которого – разработать подходы к созданию модели восприятия речи на базе корпуса русских текстов. Мы попытались отразить как методологический подход, так и основные направления исследований авторского коллектива в заявленной области.

Для начала зафиксируем исходные позиции, которые заключаются, по-видимому, в следующем.

Моделирование процессов восприятия речи (во всяком случае, на материале русского языка) включает в себя такие подготовительные этапы, как

- формирование представительного корпуса текстов (на начальном этапе – в орфографической записи) с акцентуацией словоформ и разметкой согласно специально разработанной системе аннотирования;
- создание (на базе корпуса текстов) словаря для моделирования восприятия речи; единицей словаря выступает словоформа с индексом частотности.

К настоящему времени собраны и введены в компьютер тексты 181 автора; общий объем корпуса – 1.031.920 словоупотреблений. Из них в текстах 140 авторов проставлены ударения и частеречные пометы; объем этих текстов – 322 тыс. словоупотреблений. На основании данного материала организован частотный словарь словоформ объемом 63742 единицы.

Создана новая версия автоматического фонологического транскриптора на базе кириллицы (автор программы А.В.Венцов). С помощью транскриптора осуществлено автоматическое транскрибирование текста объемом 322 тыс. словоупотреблений, в результате получена транскрибированная версия текста объемом 261972 фонетических слова. На базе данного текста создан частотный словарь фонетических слов объемом в 84174 единицы.

Наличие корпуса и словаря словоформ позволило осуществить компьютерное моделирование сегментации графической (орфографической и транскрипционной) беспробельной записи текста через идентификацию, т.е. через сличение с единицами словаря словоформ. Мы полагаем, что подобная процедура на материале «сплошной» графической записи может рассматриваться как некоторое приближение к работе с материалом звучащего текста, а реализуемые правила до некоторой степени могут соответствовать процессам восприятия речи человеком. Акцент на процедуре сегментации через идентификацию является следствием признания наибольшей доли именно этой процедуры в восприятии речи. В то же время ни в коей мере не исключается необходимость исследования и автономного механизма сегментации (независимой от идентификации).

Программа реализует алгоритм, восходящий к ранней версии «модели когорты» (Marslen-Wilson 1990 и др. работы, см. также: Венцов, Касевич 1994). В основу алгоритма положено упрощенное предположение о том, что в буфер памяти слушающего сведения о символах, составляющих экспонент слова, поступают последовательно во времени и, соответственно, происходит накопление информации, обеспечивающей выбор подходящего слова из словаря.

На материале беспробельной как орфографической, так и транскрипционной записи рассмотренных текстов точность работы компьютерной программы моделирования процедуры сегментации через идентификацию составила более 98%. Высокую эффективность описанных правил можно рассматривать как косвенное (в силу специфичности исходного материала), но убедительное подтверждение «работоспособности» алгоритма, основывающегося на основных положениях модели когорты.

Отдельно необходимо рассмотреть проблему, связанную с допущением фонетического слова на роль единицы словаря. Очевидное возражение против признания фонетического слова основной единицей словаря для слушающего состоит в чрезмерном увеличении объема словаря. Каждое слово (словоформа) может употребляться с разными проклитиками и энклитиками, отсюда, в пределе, разрастание словаря во столько раз, сколько клитик и их сочетаний существует в языке (если не принимать во внимание, разумеется, частеречные и иные ограничения). Учитывая, однако, преимущественно эмпирический характер проблемы, авторы проекта, опираясь на реальный корпус русского языка, созданный в процессе работы над проектом, получили точные количественные данные по соотношению фонетических слов текста, единиц словаря, состоящего из фонетических слов, и словаря словоформ. Как оказалось, словарь фонетических слов, хотя и превышает, разумеется, по объему словарь словоформ, но далеко не достигает при этом теоретического предела, о котором сказано выше: реальное возрастание объема – всего 30%.

Очевидное возражение против признания фонетического слова основной единицей перцептивного словаря состоит в чрезмерном увеличении объема словаря; ясно, что каждое слово (словоформа) может употребляться с разными проклитиками и энклитиками, отсюда, в пределе, разрастание словаря во столько раз, сколько клитик и их сочетаний существует в языке (если не принимать во внимание, разумеется, частеречные и иные ограничения). Учитывая, однако, преимущественно эмпирический характер проблемы, авторы проекта, опираясь на реальный корпус русского языка, созданный в процессе работы над проектом, получили точные количественные данные по соотношению фонетических слов текста, единиц словаря, состоящего из фонетических слов, и словаря словоформ. Как оказалось, словарь фонетических слов, хотя и превышает, разумеется, по объему словарь словоформ, но

далеко не достигает при этом теоретического предела, о котором сказано выше: реальное возрастание объема – всего 30 %.

Говоря о фонетических словах, следует учитывать существенную с точки зрения восприятия речи неоднородность этого класса единиц. Есть фонетические слова, совпадающие со словами (словоформами), которые «в любом случае» входят в перцептивный словарь, и есть фонетические слова, не совпадающие со словами – единицами словаря. Примером первых может служить фонетическое слово *народ* (*на род* и *народ*, точнее, *на рот* и *нарот*), примером вторых – *книму* (*к нему*). По-видимому, существование именно первого типа фонетических слов считается особенно серьезной «помехой» для оперирования фонетическими словами как особыми единицами ввиду их очевидной неоднозначности. Однако наши исследования показывают, что важность данной проблемы не следует преувеличивать. С одной стороны, экспериментально было показано, что носители языка не различают, вне лексического и грамматического контекста, единицы типа *народ*/*на род*. Модель восприятия речи, претендующая на адекватное воспроизведение структуры соответствующих механизмов человека и их функционирования, не может быть «лучше» своего естественного прототипа: то, что не различает человек, не должна различать и имитирующая его поведение модель. С другой стороны, значимость подобных пар невелика еще и потому, что их представленность в тексте и словаре, построенном на базе фонетических слов, весьма невелика. В нашем словаре фонетических слов, составленном на основе организованного авторами проекта корпуса русского языка, фонетические слова класса *народ* (*нарот*) составили всего 0.5% от общего числа фонетических слов. Одновременно можно отметить, что в ряде случаев различению членов пар типа *народ* / *на род* способствует несовпадающая частотность; так, в наших текстах число вхождений местоименной словоформы с предлогом *по этому* составляет 9 единиц, а слова *поэтому* - 81. Но никакой системы здесь, как и можно было ожидать, не наблюдается.

Итак, с одной стороны, организацию перцептивного словаря как словаря фонетических слов едва ли следует рассматривать как заведомо нереалистичную постановку проблемы. Его объем (на нашем материале около 85 тыс. единиц), конечно же, никоим образом не перегружает человеческую память. «Выгодность» такого словаря заключается, несомненно, в том, что процесс идентификации единиц текста здесь во многом сводится к процедуре их прямого сличения с единицами словаря, «наложения» первых на вторые (разумеется, с учетом всех процедур построения когорты и ее дальнейшей фильтрации). С другой стороны, из изложенного выше, по-видимому, следует, что фонетические слова в словаре представлены скорее косвенно – как словоформы, омонимичные сочетаниям словоформ и их клитик. Омонимичность разрешается путем обращения к высшим языковым уровням, к контексту. Там, где омонимичность не представлена, применяется стандартный алгоритм обращения к словарю, где, в числе прочих единиц, присутствуют и клитики, так что возможность/невозможность членения фонетического слова выступает как частный случай выбора между словами-кандидатами. Является при этом членимая последовательность фонетическим словом, отличным от слова семантико-грамматического, или нет, оказывается, вообще говоря, несущественным; фонетическое слово, определяемое акцентным контуром, выступает как промежуточный продукт, с которым работает алгоритм сегментации/идентификации. Отдельной проблемой, требующей решение, остается существование фонетических слов типа *Из лесу:* либо мы допускаем в словаре безударные знаменательные словоформы, либо все же вводим в словарь фонетические слова.

Еще один вопрос, отчасти уже поднятый в настоящем в докладе, связан с необходимостью исследования автономного механизма сегментации (независимой от идентификации). Этот механизм, хотя его общая доля в восприятии речи сравнительно мала, несомненно «затребован» в ряде речевых ситуаций, в частности, он обслуживает восприятие текста с новыми, незнакомыми словами. В литературе высказываются предположения о том, что опорой такой сегментации могут служить фонотактические закономерности. На материале рассматриваемого корпуса была проведена предварительная статистическая обработка текстов, имеющая своей целью выявить потенциальные сигналы межсловных границ. Были получены данные о частоте встречаемости каждого из возможных сочетаний согласных в позиции начала, середины, конца ФС, а также в позиции стыка ФС (в дальнейшем «слово»). Вся таблица насчитывает 3733 сочетания, из них 2339 встречается только на стыке слов, напр., самые частотные йп, лй, хп, т’н, т’н’, мй, т’п . Наши данные свидетельствуют, что 63% сочетаний согласных может служить потенциальным сигналом наличия границы. Ряд сочетаний, встречающихся только в середине слова, по-видимому, может рассматриваться как своего рода отрицательный пограничный сигнал. Такие сочетания составляют 8% от всех возможных (300 сочетаний), напр., часто встречающиеся пщ, ств‘, шк’, нск’, чк’, цтв’.

Для более тонкой, систематической оценки позиционных сочетаний с учетом различий в их частоте встречаемости в начальной, конечной или серединной позиции, все сочетания согласных были поделены на два класса частотности.

Сочетаний, частотных для конечной позиции, оказалось мало, практически все они являются частотными и для начальной позиции. Единственное исключение составляет [γ] – звонкий щелевой заднеязычный, позиционный вариант, появляющийся лишь в конце слова в результате озвончения перед последующим звонким шумным. Наличие [γ] естественным образом отмечает правую границу слова. В то же время, как минимум, нет уверенности в том, что данный субфонемный признак может оказаться перцептивно релевантным.

В качестве возможного сигнала о начале слова можно рассматривать 41 малочастотное сочетание, встречающееся только в начале (не в конце, не в середине и не на стыке), напр. фкл', фкв, взл, взл', фкр', вздр'. Если снять ограничение на встречаемость на стыке, то число сочетаний увеличивается до 60, т.е. в любом случае они указывают на наличие границы, а в большинстве своем – и на точное ее место.

Список сочетаний, частотных для начала слова, но низкочастотных для середины слова (или стыка) и не встречающихся в конце слова, существует, хотя он невелик по объему (11 сочетаний), напр., фп, гд', ср, кр', св'. Думается, что эти сочетания могут выступать в роли вероятностного указателя на левую границу слова.

Вероятно, опорой может служить и обращение к акцентному контуру слова. 40% слов имеют ударение на последнем слоге слова, еще почти 40% – после первого заударного слога и около 18% – после второго заударного. Следовательно, почти в 97% случаев правая граница слова в тексте будет не дальше второго заударного слога. Эта область, вероятно, и задает первичные ориентиры для поиска границы слова. Дальнейшее уточнение может дать фонотактика, о которой и шла речь выше.

В качестве как общетеоретического, так и методологического заключения, необходимо отметить: есть все основания полагать, что именно сочетание методов корпусной лингвистики, с одной стороны, и экспериментального подхода – с другой, позволят существенно продвинуться в моделировании речевой деятельности.

## Литература

1. Венцов А.В., Касевич В.Б. Проблемы восприятия речи. СПб.: Изд-во С.-Петербург. ун-та, 1994.
2. Венцов А.В., Касевич В.Б. Словарь для модели восприятия речи // Вестн. С.-Петербург. ун-та. 1998. Сер. 2. Вып. 3., С. 32-39.
3. Marslen-Wilson W.D. Activation, competition, and frequency in lexical access // Cognitive Models of Speech Processing: Psycholinguistics and Computational Perspectives / Ed. by G.T.M.Altmann. Cambridge, Mass.; London, 1990.