20. Moreno A., Grishman R., Lpez S., Sánchez F., Sekine S. A treebank of Spanish and its application to parsing, Proceedings of the Second International Conference on Language Resources and Evaluation (LREC).— Athens, 2000.— P. 107–111.
21. Abeillé A., Clément L. A tagged reference corpus for French // LINC'99 Proceedings, EACL workshop.— Bergen, 1999.
22. Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M. T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R. The Italian syntactic-semantic treebank: architecture, annotation, tools and evaluation // Proceedings of the COLING Workshop on Linguistically Interpreted Corpora (LINC-2000).— Luxembourg, 2000, 6 August.— P. 18–27.
23. Huang C.-R., Chen K.-J., Chen F.-Y., Chen K.-J., Gao Z.-M., Chen K.-Y. Sinica treebank: design criteria, annotation guidelines, and on-line interface // Proceedings of 2nd Chinese Language Processing Workshop // ACL-2000.— Hong Kong, 2000. P. 29–37.
24. Kurohashi S., Nagao M. Building a Japanese parsed corpus // Ed. A. Abeille Treebanks: building and using syntactically annotated corpora.— Kluwer Academic Publishers, 2003.
25. Oflazer K., Say B., Hakkani-Tür D. Z., Tür G. Building a Turkish treebank // Ed. A. Abeille Treebanks: building and using syntactically annotated corpora.— Kluwer Academic Publishers, 2003.
26. Lönngren L. (Ed.) Chastotnyj slovar' sovremennogo russkogo jazyka. Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia 32.— Uppsala, 1993.
27. Падучева Е. В. О способах представления синтаксической структуры предложения // Вопр. языкознания.— 1964.— № 2.
28. Mel'čuk I. Dependency syntax: theory and practice.— Albany: NY, SUNY Press, 1988.
29. Кривнова О. Ф., Чардин И. С. Паузирование при автоматическом синтезе речи. // Материалы конференции "Теория и практика речевых исследований" (APCO-99).— М., 1999.
30. Järvinen T. Bank of English and beyond // Ed. A. Abeille Treebanks: building and using syntactically annotated corpora.— Kluwer Academic Publishers, 2003.
31. Арутюнова Н. Д. Синтаксис // Общее языкознание: Внутренняя структура языка.— М.: Наука, 1972.— С. 254–342.
32. Manning C. D., Schütze H. Foundations of statistical natural language processing.— Cambridge: MA, MIT Press, 1999.
33. Charniak E. Statistical parsing with a context-free grammar and word statistics // Proceedings of AAAI'97, 1997.— P. 598–603.
34. Charniak E. A maximum-entropy-inspired parser / Proceedings of NAACL-2000.— Seattle, 2000.
35. Collins M. J. Three generative, lexicalised models for statistical parsing // Proceedings of ACL35/EACL8.— 1997.— P. 16–23.
36. Bod R. Beyond grammar: an experience-based theory of language // CSLI Publications.— Cambridge University Press, 1998.
37. Rosenfeld R. Two decades of statistical language modeling: where do we go from here? // Proceedings of the IEEE, 88(8), 2000.
38. Carroll G., Charniak E. Two experiments on learning probabilistic dependency grammars from corpora // Workshop Notes for Statistically-Based NLP Techniques. AAAI, 1992.— P. 1–13.
39. Zeman D. A statistical approach to parsing of Czech // Prague Bulletin of Mathematical Linguistics. Vol. 69.— Karlova, Praha: Univerzita Karlova, 1998.— P. 29–37.
40. Paskin M. A. Grammatical bigrams // Ed. T. Dietterich, S. Becker, Z. Gharahmani, Advances in Neural Information Processing Systems 14.— Cambridge: MA, MIT Press, 2001.
41. Yuret D. Discovery of Linguistic Relations Using Lexical Attraction // PhD thesis.— MIT, 1998.
42. Иомдин Л. Л., Сизов В. Г., Цинман Л. Л. Использование эмпирических весов при синтаксическом анализе. // Тр. конф. "Когнитивное моделирование в лингвистике".— Дивногорск, 2001.
43. Чардин И. С. Использование аннотированного корпуса при снятии синтаксической неоднозначности в лингвистическом процессоре ЭТАП-3 // Материалы 2-ой Всероссийской конф. "Теория и практика речевых исследований" (APCO-2001).— М., 2001.
44. Black E., Abney S., Flickenger D., Gdaniec C., Grishman R., Harrison P., Hindle D., Ingria R., Jelinek F., Klavans J., Liberman M., Marcus M., Roukos S., Santorini B., Strzalkowski T. A procedure for quantitatively comparing the syntactic coverage of English grammar // Proceedings of Speech and Natural Language Workshop, DARPA, February 1991.— P. 306–311.

УДК 811.161.1'1:159.946

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

# Корпус русского языка и восприятие речи*

*Анализируются основные проблемы, возникающие при построении модели восприятия речи на базе корпуса русских текстов. Теоретически и экспериментально обосновывается необходимость в различении функционально ориентированных словарей: генеративного и перцептивного, вокабулой которого выступает словоформа. Изучается роль ударения и фонетического слова в перцептивных процессах, в том числе роль фонетического слова в перцептивном словаре. Описывается программа сегментации текста с использованием словаря.*

В настоящее время лингвистика во многом избавилась от раннегенеративистских иллюзий, в частности, от уверенности, что лингвистические механизмы как таковые могут быть познаны с привлечением весьма ограниченного набора примеров (обычно сочиненных самим лингвистом). На смену этим достаточно наивным представлениям приходит понимание необходимости строить исследование даже самого "мелкого" фрагмента языковой системы с использованием репрезентативного

*Работа выполнена при поддержке РГНФ, грант № 03-04-00226а.

множества текстов соответствующего языка. Оговоримся, что имеется в виду репрезентативность как в количественном, так и в качественном отношении — по представленности жанров, стилей и т. п. Такое множество текстов стало уже традиционным называть *корпусом*. Приступая к исследованию конкретной проблемы, лингвист может (а в реальной ситуации, как правило, должен) составлять свой собственный корпус.

В последние десятилетия усилия лингвистов многих стран направлены на создание национальных, или универсальных, интегральных корпусов. Хотя критерии репрезентативности такого корпуса пока не вполне ясны, ясна задача: корпус должен обладать количественными и качественными параметрами, необходимыми и достаточными для построения на его основе адекватных словаря и грамматики соответствующего языка.

Адекватность словаря определяется, с этой точки зрения, тем, насколько мала вероятность встретить в произвольном тексте — вне текстов корпуса — словарную единицу (слово, словоформу, фразеологизм), отсутствующую в словаре. "Произвольность" текста не следует понимать буквально: для любого корпуса, даже универсального, допустимы ограничения — например, невключение текстов диалектного характера.

Адекватность грамматики мы предпочли бы трактовать как характеристику действующей, динамической системы, обеспечивающей речевую деятельность. Иначе говоря, грамматика для нас — это механизм порождения и/или восприятия текста (речи). Адекватность такой грамматики — это ее способность порождать правильные (нормативные) тексты и только их (критерии нормативности задаются отдельно), а также анализировать с получением заданного результата (транскрипция, семантическая запись и т. п.) правильные тексты (с учетом допустимых отклонений от правильности, см. [1] и др.).

Уже использование логической связки "и/или" выше дает понять, что мы, не отрицая единства грамматического механизма на некотором уровне, признаем, тем не менее, возможность и даже необходимость выделять грамматику, отвечающую за порождение речи, и грамматику, "заведующую" восприятием речи. Более того, в этом различении, восходящем к Л. В. Щербе с его активной и пассивной грамматиками, мы идем дальше, разграничивая также *словари:* генеративный (обслуживающий порождение речи)[1] и перцептивный (обслуживающий восприятие речи). Именно последний, как компонент модели восприятия речи, будет интересовать нас в настоящей статье.

Прежде, однако, воспроизведем основные аргументы в пользу, как нам представляется, признания относительной самостоятельности перцептивного словаря [3].

Главной отличительной особенностью перцептивного словаря нам видится характер его единицы: в качестве таковой есть основания указать *словоформу*. Можно считать экспериментально доказанным, что важным ключом для идентификации слова при его восприятии (изолированно или в тексте) выступает *частотность* данного слова. Но частотность слова как лексемы — в известном смысле фикция. Реальной частотностью характеризуются именно отдельные словоформы слова, причем разные словоформы одного и того же слова могут существенно отличаться по частотности[2].

Точно так же можно считать доказанным, что еще один важный ключ, используемый для предварительной, грубой классификации слова при восприятии речи, — это его акцентный контур. Но и акцентный контур — даже более непосредственно, нежели частотность — есть признак словоформы, а не лексемы. Разные словоформы одной и той же лексемы могут обладать разными акцентными контурами, совокупность которых образует так называемую акцентную кривую, ср., например, *сад, са́ду, (в)саду́, (в)сада́х* и т. п. Акцентная кривая создается, главным образом, перемещением ударения с основы на окончание или наоборот.

Признание словоформы основной единицей перцептивного словаря, разумеется, приводит к значительному увеличению его объема. В то же время это возрастание объема значительно меньше, чем можно было бы предположить априори; связано это с тем, что отнюдь не каждая лексема обладает полным набором словоформ, отвечающим категориям, которые присущи ее классу/подклассу. Специальное статистическое изучение такого рода ограничений представило бы отдельный интерес.

Увеличивая словарь, опора на словоформу в то же время сильно упрощает процедуру идентификации единиц текста при их восприятии, во многом сводя эту процедуру к прямому сличению отрезка текста и единицы словаря — минуя процесс лемматизации, неизбежный, если мы имеем дело с традиционным словарем лексем, а не словоформ.

Возникает еще одна проблема. Выше мы упоминали о релевантности акцентного контура словоформы в качестве ключа для ее идентификации. Но акцентный контур характеризует не словоформу как таковую, а *фонетическую словоформу* (ФС), т. е. фонетическое слово, которое состоит из знаменательной словоформы плюс клитики. Деление текста на ФС может довольно существенно расходиться с сегментацией на слова (словоформы) как лексико-грамматические единицы, ср. *Ты / бы / ко / мне / раньше / с / этим / пришел* и *Тыбы / комне / раньше / сэтим / пришел* (косая черта указывает на границу между словами, условно воспроизводится орфографическая запись).

Из релевантности именно ФС как "носителя" акцентного контура по крайней мере *может* следовать, что и единицей равно текста и словаря (а их идентичность принципиальна) выступает не просто словоформа, а ФС.

Рассмотрим указанные и иные относящиеся к ним вопросы в определенной последовательности. 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

[1] Понятие генеративного словаря здесь никак не соотносится с понятием словаря в генеративной лингвистике [2].

[2] Разумеется, частотность словоформы, которая отлична от частотности лексемы, — это особенность русского и аналогичных ему языков с развитой морфологией. Данная проблема может быть периферийной или даже несущественной для аналитических и тем более изолирующих языков (вероятно, именно поэтому аналогичные вопросы не рассматриваются в многочисленных работах на материале английского языка).

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

На настоящий момент общий объем нашего корпуса — 1 031 920 словоупотреблений. На основании подкорпуса объемом в 322 тыс. словоупотреблений организован частотный словарь словоформ, включающий 63 742 единицы и словарь фонетических слов объемом 84 174 единицы. Этот подкорпус имеется также в транскрибированном виде. Автоматическое транскрибирование текстов осуществлялось с помощью версии фонологического транскриптора на базе кириллицы (автор программы А. В. Венцов).

В данной статье мы попытались отразить как методический подход, так и основные направления исследований авторского коллектива в заявленной области.

## КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ СЕГМЕНТАЦИИ И ИДЕНТИФИКАЦИИ ГРАФИЧЕСКОЙ ЗАПИСИ ТЕКСТА

Наличие корпуса и словаря словоформ позволило осуществить компьютерное моделирование сегментации графической беспробельной записи текста через идентификацию, т. е. путем сличения с единицами словаря. Мы исходим из того, что подобная процедура на материале "сплошной" графической записи может рассматриваться как некоторое приближение к работе с материалом звучащего текста, а используемые принципы компьютерного моделирования до некоторой степени соответствуют процессам восприятия речи человеком. Сделанный акцент на процедуре сегментации через идентификацию ни в коей мере не означает отказ от исследования автономного механизма сегментации (независимой от идентификации), но лишь признание относительно небольшого удельного веса автономной сегментации на слова в восприятии речи (подробнее см. об этом ниже).

Программа реализует алгоритм, восходящий к ранней версии "модели когорты" (ср. [4] и др. работы; см. также [5]). Существенно отметить, что в большинстве ранних работ, выполненных в русле "модели когорты", материалом, подлежащим распознаванию, выступали изолированные слова — соответственно проблема сегментации вообще не возникала. В отличие от этого, наш алгоритм принципиально нацелен на обработку слитной речи — на данной стадии исследования в ее графическом представлении, а именно орфографической и транскрипционной (в терминах фонем) записей.

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.

Сам процесс выбора начинается сразу же, как только в буфере появляются первые один-два символа. По ним из словаря выбираются все подходящие слова, т. е. начинающиеся на тот же символ или последовательность символов слова, которые и образуют "когорту". По мере поступления в буфер следующих символов, из когорты удаляются все слова, не согласующиеся по началу с имеющейся в буфере цепочкой, и процесс этот продолжается до тех пор, пока в когорте не останется одно единственное слово, которое и будет считаться идентификатором распознаваемого отрезка текста.

Создатели "модели когорты" предполагали, что по мере накопления информации о фонемном составе слова будет резко сокращаться объем когорты и процесс идентификации должен сходиться достаточно быстро и эффективно (особенно если принять во внимание возможность априорного контекстного ограничения словаря, из которого производится начальная выборка когорты, что обычно не учитывается). Сделанные нами самые предварительные расчеты для русского языка показали, что объем выборки действительно стремительно сокращается по мере появления во входном буфере все новых фонем, особенно если при составлении когорты принять во внимание ритмическую структуру распознаваемого слова.

Но все это относилось к идентификации изолированных слов. Мы же попытались использовать ту же идею при "работе" с непрерывной последовательностью слов, не разделенных какими бы то ни было метками сегментации, т. е. возможности того же алгоритма оценивались применительно к распознаванию слитной речи, которая характеризуется как раз отсутствием границ между словами, образующими высказывание (синтагму). Одна из вытекающих при этом сложных проблем заключается в том, что единый процесс идентификации-сегментации предполагает нахождение правой границы слова.

В нашей модели анализируемый текст считывается из файла слово за словом и записывается в строку без пробелов и знаков препинания. Начальная часть строки длиной в 7–9 открытых слогов представляет собой буфер, с содержимым которого работает в дальнейшем программа. Объем буфера выбран на основании имеющихся данных об объеме оперативной (кратковременной) памяти человека (7±2 слога). На этом этапе алгоритм работы программы, скорее всего, не соответствует предполагаемому алгоритму работы системы распознавания речи человеком и выбран таковым только из удобства программной реализации процесса.

По первому символу строки-буфера начинается процесс образования текущей когорты. Для орфографической записи при этом применяются следующие правила: (1) если первая буква не является допустимым однобуквенным словом, не содержащим ударного гласного (союзом, предлогом), то происходит только определение объема когорты, сама же когорта как набор слов не создается (это чисто программистский ход, экономящий время); если первая буква является допустимым однобуквенным словом, то из соответствующей словарной статьи в промежуточный буфер записывается слово-кандидат, а из остальных словарных статей выбираются данные об их объеме для сбора статистики; (2) заполнение когорты производится по двум первым буквам буфера-строки (или только по первой, когда это ударный гласный, поскольку по чисто техническим причинам ударные гласные представлены в текстах и в словарных статьях двухсимвольными сочетаниями: собственно гласный и знак

ударения "+"; равным образом согласные тоже могут иметь двухсимвольные соответствия с учетом "ъ" или "ь")[3], (3) буфер слов-кандидатов заполняется до тех пор, пока N первых символов в исходном буфере совпадают хотя бы с одним словом в когорте и прекращается, когда добавление еще одного элемента создает комбинацию, не представленную в словаре; вслед за этим начинается анализ слов-кандидатов. Правила работы с транскрипционной записью полностью аналогичны приведенным выше.

В данный момент при выборе окончательного варианта из всех слов-кандидатов принято самое простое правило: окончательным считается слово, последним занесенное в список — при условии, что сохраняется возможность идентификации через словарь "оставшейся" цепочки. Это вполне соответствует правилу отбора, сформулированному в теории когорты: выбирается только слово, полностью и без остатка совпадающее с входной последовательностью символов.

На материале как беспробельной орфографической, так и транскрипционной записи рассмотренных текстов точность работы компьютерной сегментации посредством идентификации составила более 98%. Столь высокую результативность описанных правил мы можем рассматривать как косвенное (в силу специфичности исходного материала), но убедительное подтверждение "работоспособности" алгоритма, базирующегося на основных положениях модели когорты.

## ПЕРЦЕПТИВНЫЙ СЛОВАРЬ

Одна из задач нашей работы заключается в проверке выдвинутой гипотезы о существовании особого перцептивного словаря. В качестве одного из средств верификации гипотезы был использован свободный ассоциативный эксперимент, где в роли стимулов используются как словарные, так и несловарные формы слов.

Предварительный ассоциативный эксперимент в его устно-письменном варианте был ранее проведен студенткой А. Морозовой (рук. Е. В. Глазанова) на материале, включающем все финитные формы глаголов. В протоколах зафиксировано в среднем более 15% реакций, явно, непосредственно обусловленных грамматической формой глагола-стимула. В большинстве случаев это относится к парадигматическим реакциям, например, *берешь — отдаешь*. Частичную обусловленность реакций формой глагола-стимула можно видеть в парах более сложных типов, например, *берешь — отдавай* или даже *брал — не отдаст*, и, наконец, в синтагматических реакциях с согласованием глагола-стимула и имени-реакции, ср. пары *брал — папа, брало — оно, берешь — ты* и т. д. С учетом всех вариантов, где представлена частичная обусловленность грамматики реакции грамматикой стимула, можно утверждать, что такая связь характеризует до 99% пар "стимул-реакция" в описываемом эксперименте.

Возможно, особенности методики устно-письменного эксперимента (переключение модальности, наличие нескольких реакций на один стимул) лишь отчасти позволяют использовать ее в решении поставленной задачи. В настоящее время проводится серия устно-устных ассоциативных экспериментов, в которых список стимулов включает различные формы существительных и глаголов. Данный эксперимент проводится с участием как взрослых испытуемых, так и детей шести лет, языковые механизмы которых находятся в стадии развития. Имеющиеся на настоящий момент предварительные результаты не противоречат высказанной гипотезе.

Основываясь на этих предварительных результатах, естественно предположить, что испытуемые непосредственно переходят от словоформы как стимула к словоформе как реакции. Поскольку выбору реакции с необходимостью предшествует основанная на обращении к словарю идентификация стимула, приходится признать, что вход в словарь в данном случае — это обнаружение соответствующей *словоформы*. В противном случае мы должны были бы полагать, что сначала осуществляется процесс лемматизации, а затем — возвращение к уже "использованной" словоформе для установления информации о ее характеристиках, которые служат основанием для выбора словоформы-реакции. Иначе говоря, ассоциативные эксперименты подтверждают гипотезу о словоформе как основной единице перцептивного словаря.

Как отмечалось во вступительном разделе статьи, есть основания полагать, что единицей перцептивного словаря выступает не просто словоформа, а словоформа *фонетическая*. 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, 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

[3] В одной из модификаций программы вводилось важное изменение по сравнению с "классической" моделью когорты: образование текущей когорты осуществлялось не по первому символу, а по предударной части и ударному гласному анализируемой цепочки. При использовании данного варианта повышалось быстродействие программы.

серьезной "помехой" для оперирования фонетическими словами как особыми единицами ввиду их очевидной неоднозначности. Однако наши исследования показывают, что важность данной проблемы не следует преувеличивать. Во-первых, экспериментально было не раз 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. Во-вторых, значимость подобных пар не следует переоценивать 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 сформированного 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 девять 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 85 000 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, по-видимому, следует, что фонетические слова в словаре представлены скорее косвенно — как словоформы, омонимичные сочетаниям словоформ и их клитик. Омонимичность разрешается обращением 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.

## ФОНЕТИЧЕСКОЕ СЛОВО И РЕДУКЦИЯ

В этом разделе мы представим дополнительные экспериментальные данные, относящиеся к роли ФС в процессах восприятия речи.

ФС для русского языка неразрывно связано с ударением. С точки зрения восприятия речи — это, как многократно упоминалось, означает, что, опознавая ударные слоги в тексте, носитель языка членит текст на фонетические слова. Членение может осуществляться с точностью до числа ФС и с точностью до фиксирования межсловных единиц, где под словами, опять-таки, должны пониматься слова фонетические.

Установление межсловных границ было бы возможным, если бы границы акцентного контура были *перцептивно опознаваемыми*. Теория пограничных сигналов Н. С. Трубецкого по существу предполагает такой вариант: по крайней мере со времен А. А. Потебни известно, что русское слово характеризуется разными степенями редукции гласного (слога), которые определяются позицией относительно ударного слога в пределах слова, и, соответственно, зная тип редукции — умея его определять в тексте, — мы получаем информацию о "местоположении" начала/конца слова в речевой цепи.

Однако в действительности носителю языка едва ли доступны подобные операции. Даже если считать, что традиционные представления о "дуге редукции" в пределах слова верны, из этого еще не следует, что соответствующая информация принадлежит к перцептивно полезным признакам, используемым в процессе восприятия речи.

Об этом говорят и эмпирические данные наблюдений над восприятием реальной речи. Так, лишь семантическая неинтерпретируемость мешает воспринимать строку известной песни *сказал кочегар кочегару* как *сказалка чигарка чигару* или *сказалка чигар качигару*. Такие перераспределения границ были бы очевидным образом невозможны, если бы информация о типе редукции реально использовалась. Вполне естественно, что подобные ошибки в изобилии дает ситуация восприятия речи на фоне шума, когда затруднен доступ к информации о сегментной структуре слова и, следовательно, о семантических характеристиках высказывания. Примерами могут служить замены наподобие *зеленый крокодил* → *наверно приходил, черешни поспели* → *лежи в постели, живу воспоминаниями* → *желает понимания* и т. д.

Иначе говоря, информация о редукции, скорее всего, не используется для определения границ фонетического слова.

Те же эксперименты по восприятию речи в условиях маскировки дают, однако, и замены принципиально иного типа, которые ставят под сомнение незыблемость самого по себе положения о том, что число ударений везде совпадает с числом ФС, например, *ловля птиц* → *коллектив* [6]. Из внеэкспериментальных свидетельств, которые также колеблют принятое положение о взаимооднозначном соответствии между ударениями и ФС, можно указать на каламбуры наподобие знаменитых минаевских: *муж, побледнев как штукатурка, воскликнул — это штука турка!* или *даже к финским скалам бурым обращаюсь с каламбуром; писать стихи — моя стихия, и легко пишу стихи я*. Если бы *штука турка* и *штукатурка* уверенно различались как, соответственно, два (фонетических) слова *vs.* одно (фонетическое) слово за счет наличия двух *vs.* одного ударения, то эффект каламбура, очевидно, не возникал бы.

Наконец, можно добавить, что неочевидно просодическое (акцентное) противопоставление пар наподобие *на диване* (одно ФС) и *дядя Ваня* (два ФС). По поводу последнего из упомянутых типов один из авторов настоящей статьи пишет: "Очевидно, не формулируемое явным образом рассуждение, которое ведет к традиционному разграничению сочетаний типа *дядя Ваня* и *на диване*, должно выглядеть следующим образом: при полном сохранении просодических характеристик (при сохранении акцентного контура) вместо *дядя Ваня* можно ожидать, например, сочетание *тетя Таня*. Но в этом сочетании в слове *тетя* имеем фонему /о/, а /о/ не может быть безударным (если отвлечься от малочисленных исключений). Следовательно, само по себе наличие /о/ ... свидетельствует о двуударности — а тем самым о наличии двух ФС в сочетании *тетя Таня* и, по аналогии, в *дядя Ваня* (в отличие от *на диване*), что и требовалось доказать" [7, с. 107].

Принятие приведенного рассуждения предполагает учет теснейшей взаимосвязи просодических (акцентная структура слова) и сегментных характеристик слова (редукции или даже чередования фонем). В связи с этим можно вспомнить, что в литературе существуют концепции, согласно которым выделяются не только ударные/безударные слоги, но и сильные/слабые, или тяжелые/легкие — нередуцированные и редуцированные соответственно. Если в пределах одного языка нет попарного совпадения членов указанных противопоставлений, т. е. безударный слог не всегда редуцированный, а ударный — не всегда сильный, то возникает возможность вычленения ФС по двум относительно независимым критериям: наличию/отсутствию ударения и наличию/отсутствию (и типу) редукции. В сущности, к такому подходу близка не получившая дальнейшего развития позиция Э. Пальгрэма, который предлагал различать нексусные и курсусные единицы [8].

Теоретически реальным выглядит предположение, когда ФС будет определяться одновременно по набору признаков, как акцентных, так и "редукционных". Тогда мы получим некоторое множество структурных типов ФС, в котором, например, *на диване* будет характеризоваться в терминах признаков [+ одноударн.], [+ полноредуц.], а сочетание *дядя Ваня* попадет в другой подкласс с признаками [+ одноударн.], [– полноредуц.].

Ясно, однако, что такого рода теоретические гипотезы должны проверяться экспериментально, ибо, как и в случае с акцентным контуром, априори неизвестно, какие именно признаки реально используются носителями языка в речевой деятельности (при восприятии речи).

С целью проверки соответствующих гипотез был проведен ряд экспериментов. Исследовалась возможность своего рода перцептивной нейтрализации противопоставления одного ФС двум. Изучались следующие типы такого неочевидного противопоставления.

1. Пары, включающие слово и словосочетание из двух знаменательных слов, совпадающее с первым членом пары по фонемному составу и месту ударения во втором слове, ср. *барбариса — бар Бориса*. Всего исследовалось 18 таких пар. На базе списка пар было составлено 64 фразы: каждое слово и словосочетание было помещено как в нейтральный (допускающий обе интерпретации), так и в однозначно диктующий выбор контекст (в дальнейшем "однозначный"); слова и словосочетания находились, как правило, в конечной позиции во фразе; возможность просодического выделения слов в словосочетаниях минимизировалась.

2. Пары наподобие *на диване — дядя Ваня, на отложка — наша ложка*, в которых первый элемент традиционно трактуется как одно, а второй — как два ФС (всего 10 пар). На базе этого списка было составлено 66 фраз: каждое слово и словосочетание было помещено как в нейтральный так и в однозначный контекст; слова и словосочетания находились как в начальной, так и конечной позиции во фразе; возможность просодического выделения слов в словосочетаниях также минимизировалась.

3. Сочетания глаголов (разной ритмической структуры) с постпозитивным личным местоимением или частицей, напр. *читали мы — читали бы — читали бы мы* ... и т. д. На базе этого списка было составлено 180 фраз наподобие *Читали мы эту книгу — Читали бы эту книгу — Читали бы мы эту книгу* ...

Выше описанные фразы в случайном порядке были включены в состав большой таблицы, содержащей разнообразные фразы, которая была прочитана в естественном темпе диктором-женщиной, опытным лингвистом-педагогом. На настоящий момент записана, но еще не обработана, аналогичная таблица, прочитанная в естественном темпе диктором-мужчиной.

Методика работы предполагает сочетание инструментального и перцептивного анализа. Инструментальный анализ включает в себя анализ акцентного контура стимула (рассматриваемого как в составе фразы, так и изолированно) по следующим параметрам: длительность, интенсивность и диапазон изменений частот основного тона (ЧОТ) для слогов рассматриваемых слов и сочетаний.

Перцептивный анализ состоит из четырех серий экспериментов, в которых испытуемым было предложено прослушать (1) изолированно предъявляемые стимулы, выделенные из фраз, и (2) фразы с нейтральным контекстом и выбрать один из двух вариантов, предложенных в анкете. Две серии содержали интактный материал (без зашумления) и две — в условиях маскировки белым шумом при соотношении сигнал/шум 0 дБ. В качестве испытуемых выступали студенты-филологи, для каждой серии использовалось более 20 испытуемых.

В настоящей статье мы опишем лишь часть из полученных данных, выделив следующие результаты экспериментов:

1) различение рассматриваемых пар для носителей языка представляет немалую сложность, для отдельных стимулов число ошибок доходит до 78%;

2) наличие фразового (нейтрального) контекста упрощает правильный выбор (для интактного материала);

3) в целом ошибки в выборе варианта для стимулов с предположительно двумя ФС случаются несколько чаще, чем с предположительно одним ФС, для фраз это различие больше, чем для изолированного предъявления;

4) изолированно предъявляемые стимулы, выделенные из нейтрального контекста, порождают несколько меньше ошибок, чем стимулы, извлеченные из "однозначного" контекста, однако это различие незначительно;

5) на сложность выбора варианта в паре оказывают существенное влияние тип связи между словами (грамматической, лексической, что мы сейчас не обсуждаем) и фонетические параметры (длина ФС в слогах, место ударения в ФС, расстояние между ударными слогами в словосочетании и некоторые другие).

Данные инструментального анализа стимулов мы пока оставим в стороне. Ясно, однако, что уже перцептивные данные свидетельствуют: в русском языке нет полной однозначности в противопоставлении ФС и словосочетания за счет одно-/двуударности. Имеет место своего рода нейтрализации противопоставления слова словосочетанию в русском языке. Возможно, наряду с общеизвестными положениями о редукции сегментных единиц следует ввести представление о *просодической редукции* — в частности, *редукции ударения*.

В описанных экспериментах не учитывались параметры, связанные с редукцией гласных и слогов в целом. Хотя выше мы подвергли сомнению перцептивную релевантность редукции безударных слогов, есть основания отдельно рассматривать редукционные характеристики *заударной части слова как целого*. Можно предположить, что само по себе наличие такого сильно редуцированного участка (плохо поддающегося членению на фонемы) длиной, обычно, более чем в один слог, с некоторой степенью вероятности соотносится с сигналом о границе между словами [9].

## РОЛЬ АНЛАУТА, ИНЛАУТА И АУСЛАУТА СЛОВ В СЕГМЕНТАЦИИ ТЕКСТА

Как не раз говорилось, сегментация через идентификацию возможна лишь для знакомых слов. Как осуществляет сегментацию носитель языка, сталкиваясь с новыми, незнакомыми словами? В литературе высказывались предположения о том, что опорой такой сегментации могут служить фонотактические закономерности. В этой публикации мы представляем результаты предварительной статистической обработки текстов, имеющей целью выявить потенциальные сигналы межсловных границ.

Мы не ставим задачу моделирования процедуры сегментации с учетом закономерностей фонотактики и статистических характеристик появления сочетаний согласных в разных позициях в слове. Это задача дальнейших исследований.

На материале нашего корпуса текстов были получены данные о частоте встречаемости каждого из возможных сочетаний согласных[4] 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 *й'n, лй', хn, т'н, т'н', мй', т'п*[5]. Таким образом, 63% сочетаний согласных может служить сигналом наличия границы. Целый ряд сочетаний, встречающихся только в середине слова (но не в начале, не в конце и не на стыке слов), может рассматриваться как своего рода отрицательный пограничный сигнал. Такие сочетания составляют 8% от всех возможных (300 сочетаний), а в качестве примера таковых можем привести часто встречающиеся *пщ', ств', шк', нск', ч'к', цтв'*[6].

Для более систематической оценки позиционных сочетаний с учетом различий в их частоте встречаемости в начальной, конечной или серединной позиции, все сочетания были поделены на два класса частотности[7].

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 сигнала о начале слова можно рассматривать 41 малочастотное сочетание, встречающееся только в начале (но 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.

Набор сочетаний, частотных для начала слова, но малочастотных 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'*[8]. Думается, что эти сочетания могут выступать в роли вероятностного указателя на левую границу слова.

Было выявлено, что 40% слов имеют ударение на последнем слоге слова, еще почти 40% — после первого заударного слога и около 18% — после второго заударного. Таким образом, почти в 97% случаев правая граница слова в тексте будет не дальше второго заударного слога. Эта область, вероятно, и задает первичные ориентиры для поиска

[4] Под сочетанием согласных здесь и далее понимаем как собственно сочетание нескольких согласных, так и нуль согласного или одиночный согласный.

[5] С абсолютной частотой встречаемости — *й'n* (413), *лй'* (254), *хn* (224), *т'н* (185), *т'н'* (179), *мй'* (158), *т'п* (156). Суммарная частота встречаемости для сочетаний согласных как в начальной, так и в конечной позиции — 160 683.

[6] С абсолютной частотой встречаемости — *пщ'* (331), *ств'* (246), *шк'* (226), *нск'* (214), *ч'к'* (191), *цтв'* (176).

[7] Для определения класса частотности абсолютные частоты встречаемости наносились на логарифмическую шкалу: минимальное значение соответствовало 1, максимальное — максимальной частоте встречаемости. Полученная таким образом шкала делилась на 2 равные части, а сочетания согласных — по нахождению в той или иной части — на 2 класса: частотные и малочастотные (в дальнейшем — "сочетания частотные и малочастотные").

[8] Со значением вероятности появления в начальной и конечной позиции (через тире): *фп* (91% — 9%), *гд'* (83% — 17%), *ср* (83% — 17%), *кр'* (54% — 2%), *св'* (28% — 1%).

границы слова. Дальнейшее уточнение может дать фонотактика, о которой и шла речь выше.

Для сопоставительного исследования роли анлаута и ауслаута в идентификации слов были проведены три серии экспериментов на материале орфографической записи текста без знаков препинания и пробелов между словами[9]. В них испытуемым предложено восстановить недостающую информацию, а именно: буквы, число и место которых указано на бланке, и расставить метки словоделения.

Материал для трех экспериментальных серий представлял собой следующие модификации:

1) из каждого фонетического слова (ФС) текста удалялся его анлаут — первый открытый слог типа Г, СГ, ССГ, СССГ,

2) из каждого ФС текста удалялся его ауслаут — последний гласный слова с последующими согласными типа Г, ГС, ГСС,

3) из каждого ФС текста удалялся его анлаут и ауслаут.

В таблице приведена информация об удаляемых сегментах в этих экспериментальных сериях.

**Распределение буквенной информации в удаляемых анлаутах и ауслаутах**

| | анлаут серия 1 | ауслаут серия 2 | анлаут и ауслаут серия 3 |
|---|---|---|---|
| удалено букв | 26% | 18% | 41% |
| распределение удаленных 1-, 2-, 3-, 4- и 5-буквенных сочетаний | | | |
| 1-буквенные (один гласный) | **26%** | **59%** | 0% |
| 2-буквенные | **50%** | **36%** | 21% |
| 3-буквенные | 24% | 3% | 38% |
| 4-буквенные | 0% | 2% | 27% |
| 5-буквенные | 0% | 0% | 14% |

С каждой экспериментальной серией работало не менее 11 человек — студенты-филологи.

Правильное опознание ФС составляет 80%, 94% и 27% для серии 1, 2 и 3 соответственно[10]. Правильное опознание лексико-грамматического слова (без учета клитик и формы слова) существенно возрастает (с 27% до 48%) лишь для серии 3.

В соответствии с моделью когорты, предполагалось существенное различие "перцептивной силы" анлаута и ауслаута[11], что в целом и подтвердилось (14% разницы между сериями 1 и 2).

В заключение, отметим: есть все основания надеяться, что именно сочетание методов корпусной лингвистики, с одной стороны, и экспериментального подхода — с другой, позволят существенно продвинуться в моделировании речевой деятельности.

## СПИСОК ЛИТЕРАТУРЫ


1. Касевич В. Б. Элементы общей лингвистики.— М.: Наука, 1977.

2. Pustejovsky J. The generative lexicon. Camb., Mass., 1995.

3. Венцов А. В., Касевич В. Б. Словарь для модели восприятия речи // Вестн. СПб. ун-та.— 1998.— Сер. 2. Вып. 3.— С. 32–39.

4. 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.

5. Венцов А. В., Касевич В. Б. Проблемы восприятия речи.— СПб.: Изд-во С.-Петербург. ун-та, 1994.

6. Касевич В. Б., Шабельникова Е. М., Рыбин В. В. Ударение и тон в языке и речевой деятельности.— Л., 1990.

7. Касевич В. Б. Еще о понятии фонетического слова // Проблемы фонетики IV.— М.: Наука, 2001.

8. Pulgram E. Syllable, word, nexus, cursus. The Hague; Paris, 1970.

9. Овчаренко Е. Б. Реализация и восприятие заударных морфных комплексов и функциональная нагрузка морфем: Дис. на соиск. уч. степени канд. филол. наук.— Л., 1988.

10. Богомазов Г. М. Роль ритмической структуры слова при восприятии письменного и звучащего текста // 100 лет экспериментальной фонетике.— СПб.— 2001.— С. 27–29.

11. Noteboom S., Vlugt M. J., van der. A search for a word-beginning superiority effect // JASA.— 1988.— Vol. 84.— № 6.— P. 2018–2032.


[9] Эксперимент проводился при участии Е. В. Грудевой.

[10] Ср. с близкими данными, полученными Г. М. Богомазовым [10] на материале сходных опытов.

[11] Ряд опытов по определению соотносительного вклада анлаута и ауслаута слова в его идентификацию проводил С. Нотебоом. В резюме к одной из статей, посвященной данной проблеме, говорится: "...процесс активации слов при восприятии устной речи в равной степени чувствителен к стимульной информации, которую несет как начало слова, так и его конец. Особая роль начала слова все же остается, поскольку анлаут дает возможность обеспечить временную связь между характеристиками стимула и [когортой] слов-кандидатов" [11, с. 2018].