# Основные тенденции развития многоязычной корпусной лингвистики

# (Часть 2)

***Потапова Родмонга Кондратьевна,***
*доктор филологических наук, профессор*

Corpus of Interactional Data (CID) представляет собой УРБД интерактивных аудиовизуальных материалов, собранных и затранскрибированных в Лаборатории языка и речи Университета Экс-ан-Прованса. CID является уникальным источником информации для анализа французской разговорной речи с учётом разных лингвистических уровней (фонетического, просодического, синтаксического, семантического, прагматического и мимико-жестикулярного).

CID стоит в ряду проектов, связанных с формированием крупных лингвистических БД (и, в частности, УРБД): MATE (Multilevel Annotation, Tools Engineering — Многоуровневое маркирование, технические инструменты); ATLAS (Architecture and Tools for Linguistic Analysis System — Архитектура и инструменты для лингвистических аналитических систем); NITE (Natural Interactivity Tools Engineering — инженерное обеспечение инструментами для естественного общения); Map Task — HCRC (картографирование); DAMSL (Dialog Act Markup in Several Layers — Маркирование диалоговых актов на нескольких уровнях); Verbmobil. Вместе с тем существует очень мало УРБД применительно к французскому языку. Те же, которые существуют (например, COPRAIX, VALIBEL), не всегда доступны. Кроме того, указанные ресурсы создавались для решения сравнительно небольшого круга задач, что накладывает существенные ограничения на возможности их использования для других исследований. Данная ситуация обусловила необходимость в создании такой УРБД, как CID.

Существующие УРБД не являются достаточными для качественного ведения многоуровневого исследования речи. УРБД должна включать записи речи суммарной длительностью в несколько сот часов. Однако лишь небольшая часть этого корпуса сопровождается текстовой репрезентацией. Недавно были предприняты попытки транскрибирования радиотелефонной речи (УРБД ESTER). В других научных областях, таких как лингвистика речевой коммуникации, составляются УРБД всё более сопоставимые по объёму с, например, УРБД CLAPI (Лион). Однако здесь возникает ещё одна трудность: осуществление записи речевого материала в естественной среде (деловой и бытовой дискурс) оказывается делом деликатным. Наконец, процедура собственно анализа фонетико-просодических уровней является чрезвычайно трудоёмкой, поэтому УРБД, содержащие хотя бы тридцать минут звучащего материала, уже могут считаться крупными. Всё это в совокупности объясняет востребованность и необходимость УРБД CID.

УРБД CID создана для удовлетворения информационных потребностей, связанных с разными лингвистическими уровнями: начиная с периферийного (фонетического) и до самого

высокого (дискурсивного, интерактивного), между которыми находятся просодический, синтаксический, семантический, прагматический уровни, а также мимико-жестикуляторного [Bertrand, Blache, Espesser et al. 2006: 31—35].

Поставленная общая задача диктует два, казалось бы, взаимоисключающих требования: обеспечить высокое качество записи, позволяющее проводить анализ на сегментном и супрасегментном (просодическом) уровнях, и в то же время не отбраковывать и не исключать из УРБД диалоги, заслуживающие внимания с учётом уровня коммуникации (например, организации речевых средств, реакции слушателя и т.п.), даже в тех случаях, где качество записи посредственное.

Этапы маркирования и, главным образом, транскрибирования, в силу их трудоёмкости, выполнялись коллективом исследователей. Кроме того, в дальнейшем планируется пополнение УРБД, что обусловливает необходимость в прозрачной и достаточно жёсткой формальной структуре УРБД и столь же прозрачном, формальном и жёстком протоколе записи информации (маркеров) различных уровней.

В настоящее время УРБД CID содержит приблизительно 8 часов диалога на французском языке. Каждый диалог (с участием двух говорящих) длится около 1 часа. Шестнадцать дикторов (10 женщин и 6 мужчин) являются уроженцами разных регионов, однако большинство из них уже в течение многих лет проживают на юго-востоке Франции.

CID представляет собой нечто среднее между многоцелевой УРБД аутентичного речевого материала, подобной CLAPI, и УРБД типа Map Task, которая разрабатывалась под одну конкретную задачу. Собственно, последний тип УРБД лёг в основу множества вариантов УРБД, приспособленных к специфике различных языков (итальянского, шведского, французского и т.п.), а также к решению различных задач.

Дикторы — участники диалогов, включённых в CID, опирались на инструкцию, которая использовалась в качестве инструмента «тематической поддержки», т.е. для того чтобы достаточно быстро завязать разговор и не запинаться в той или иной ситуации. Инструкция не накладывала каких-либо жёстких ограничений на ход диалога, который, таким образом, был реализован максимально естественно.

Запись проводилась в безэховой камере. Участники диалога находились на расстоянии около метра друг от друга, что при обычном разговоре считается естественным. На собеседниках были надеты микрошлемы, позволяющие регистрировать каждый голос на отдельной дорожке. Полученное таким образом оптимальное качество речевого материала позволяет использовать его в других исследованиях. Одно из очевидных преимуществ CID в том, что благодаря раздельной записи голосов места, где реплики накладываются друг на друга, остаются пригодными для всех видов акустического анализа и транскрибирования, что особенно ценно, так как явления, происходящие именно на участках перекрытия реплик диалога, остаются пока малоизученными; в то же время часто звучат утверждения (пока не имеющие подтверждения за отсутствием соответствующего материала), что они играют очень важную роль в структурной организации дискурса.

Кроме аудиозаписи в УРБД CID включена и видеозапись диалогов, что позволяет использовать УРБД для изучения полимодального дискурса.

Некоторое внимание уделялось также психологической подготовке информантов. Участники диалогов должны были быть знакомы с местом записи, чтобы свести к минимуму влияние фактора стресса на естественность речи. Все участники являются сотрудниками (научными сотрудниками или аспирантами) лаборатории, которая занималась формированием УРБД. При делении группы на пары учитывалась степень знакомства информантов и их привычки к общению друг с другом. Наличие привычки гарантирует, что у информантов имеется реальный опыт общения, что способствует большей непринуждённости. В этом случае им легче при необходимости отклониться от инструкции и в то же время продолжить следовать всем требованиям эксперимента.

Как уже было сказано выше, все информанты руководствовались инструкцией, в то же время сохраняя естественность речи и позволяя себе свободные импровизации. CID включал в себя разнообразные материалы, в числе которых множество нарративных (главным образом, там, где участники действовали по инструкции), а также доказательных, объяснительных или описательных типов дискурса. Полученные диалоги имеют характер непринуждённых бесед: речь может быть достаточно плавной, иногда с перебоями (заполненные паузы, вступления, фальстарты и т.п.). Структурная организация речи подчиняется принципам чередования плавных и неплавных процессов. Наблюдаются так называемые плавные переходы, в которых речь собеседников чередуется достаточно гладко, ритмично, то есть без слишком долгих пауз или «вторжений» в речь собеседника, а также неплавные переходы — многочисленные «перебивы-вторжения» в речь.

Транскрипция в CID, главным образом, орфографическая, опирающаяся на транскрипции GARS. Вместе с тем она а) отражает все типичные явления устной речи, такие как заполненные паузы («ээ-э», «ммм», «хмм» и т.п.), фальстарты, повторы, усечённые слова; б) эксплицитно указывает на имена собственные, географические названия, прямую речь и т.п.; в) содержит некоторые детали фонетического характера (шва, региональная особенность, специфическое произношение и т.п.), необходимые на следующих этапах фонетизации и совмещения с аудиосигналом.

Транскрибирование выполнено с использованием программного пакета PRAAT экспертами-фонетистами. На начальном этапе фонетизации эксперты осуществляли проверку работы программы, чтобы свести к минимуму число возможных ошибок. Перед транскрибированием выполнялось предварительное автоматическое членение речевого сигнала на межпаузальные единицы (ME). ME — блоки речи, ограниченные паузами молчания, как минимум, в 200 мс (длительность может варьироваться в зависимости от ряда факторов). Автоматическая процедура сегментации на ME заключалась в выделении прежде всего глухих / звонких звуков и в установлении пороговой величины для определения паузы определённой длительности. ME часто используется в качестве опорной единицы при формировании УРБД большого объёма. По своей формальной и объективной природе эта единица отличается от других просодических единиц, таких, например, как интонационные единицы, членение которых требует ручной работы экспертов, которые, кроме того, могут давать противоречивые интерпретации.

Автоматическая сегментация на ME не только облегчила транскрибирование, но и улучшила эффективность работы на этапах фонетизации и совмещения с аудиосигналом. В тех случаях, когда автоматическая сегментация оказывалась ошибочной, она корректировалась экспертами. На следующем этапе выполнялось автоматическое преобразова-

ние орфографической транскрипции в цепочку символов SAMPA. Для этого использовалась автоматическая программа «Фонетист» (см. табл. 4).

*Таблица 4*

**Фонетико-орфографические соответствия**

| Орфографическое представление | Фонетическое представление |
|---|---|
| je_suis | Sui |
| Allé | Ale |
| Heu | @ |
| c'est-à-dire | stAdiR |
| Nourrir | nuRiR@ (южное произношение) |

Использованная в ходе работ по формированию CID программа совмещения цепочки символов SAMPA с аудиосигналом была создана в LORIA Д.Фором и И. Лапри. Она опирается на метод HMM (http://www.loria.fr/equipes/parole/). В качестве исходной информации эта программа принимает список фонем и аудиосигнал. Результат на выходе — временно́е местоположение каждой фонемы относительно начала сигнала.

Следует отметить специфические трудности, которые возникали на этапах фонетизации и совмещения. Так, например, например, «je sais» — «я знаю» в беглой разговорной речи иногда произносится как «chai». В подобных случаях решение принималось следующим образом: на этапе транскрибирования и фонетизации отклонения от произносительной нормы передавались транскрипцией. Таким образом, цепочка символов на входе программы оказывалась максимально соответствующей реальному речевому сигналу.

Помимо фонетического и орфографического уровней маркирования в УРБД CID существуют и другие уровни, которые образуют мультимодальную структуру CID.

Уровень сегментного фонетического маркирования имеет (впрочем, как и уровни просодического и мимико-жестикуляторного маркирования) отличительную черту: единицы этого уровня необходимым образом соотносятся с физической реальностью речевого сигнала. Процесс соотнесения маркеров со звучащим материалом УРБД связан с выбором определённых теоретических и методологических подходов. В противоположность просодическому и мимико-жестикуляторному уровням, фонетический уровень не ставит проблем, касающихся выбора кода маркирования. В номенклатуре фонетических единиц нет множества возможных теоретических моделей. Зато предметом обсуждения может стать степень точности маркирования (вводить ли какие-то обозначения для переходных процессов артикуляторных фаз и т.п.), а также точности определения позиций пограничных маркеров.

Основная проблема фонетического маркирования состоит во временно́м определении сегментных единиц речи. Иногда бывает очень сложно найти соответствие между абстрактным и дискретным фонетическим кодом и более или менее непрерывным речевым сигналом.

Одно из явлений, обуславливающих сложность процесса фонетического маркирования, — коартикуляция. Каждая фонема характеризуется совокупностью артикуляторных признаков. Так, для гласного /u/, помимо прочего, характерен такой признак, как огубленность. На артикуляторном уровне этот признак характеризуется выдвижением губ вперед и их округлением. Данный жест обычно предшествует акустическому эффекту звука, то есть он начинается гораздо раньше, чем произносится сам гласный: в слоге /su/ округление губ начинается с момента начала произнесения /s/. Если в приведённом примере представлены два звука в контакте, то было показано, что характерные артикуляторные следы некоторых звуков могут быть идентифицированы на большем расстоянии от самого звука[1]. Следствием коартикуляции является то, что временна́я протяжённость фонетической единицы оказывается достаточно неопределённой и чаще всего превосходит длительность сегмента, который необходимо идентифицировать по речевому сигналу.

Другая сложность связана с отсутствием физических маркеров границ сегментов в потоке речи. В речевом сигнале можно обнаружить много различных «переломных» точек, однако далеко не всегда они соответствуют границам фонетических единиц. Так, одной из таких точек является начало эксплозии взрывного звука или аффрикаты, однако здесь границы сегментов нет. В то же время переход от одного сегмента к следующему может проходить без заметного разрыва, как это имеет место в стечениях гласных звуков. Очевидно, что в этом случае постановка маркера начала/конца сегмента носит произвольный характер и зависит от теоретического и методологического выбора эксперта.

Можно ли в связи с этим утверждать, что любая сегментация речевого сигнала может считаться ошибочной? Вообще говоря — да. Этот вопрос в течение долгого времени вызывал дискуссии в научном сообществе. Необходимо допустить, что разметка границ фонетических единиц — операция, опирающаяся на произвольные критерии и отвечающая необходимости анализа этих единиц. Часто поднимается вопрос о возможности маркирования лишь центра (ядра) квазистационарной части каждой звуковой единицы, что могло бы разрешить проблему границ. Однако такой выбор не позволяет оперировать таким важным параметром, как длительность фонетической единицы, и делает проблематичной синхронизацию многоуровневого маркирования.

При формировании УРБД CID был сделан выбор в пользу постановки маркеров начала и конца каждой единицы сегментного уровня. При этом было сделано допущение о том, что сигнал, заключённый между маркерами, является акустическим соответствием не звука, а лишь части звука. Было показано, что коартикуляция более детально маркируется индексами места образования, в то время как способ образования мог бы служить основанием для более детального временно́го маркирования. То, что какие-то части сегментов могут при таком подходе оказаться вне участков, ограниченных маркерами, представлялось авторам CID не столь существенным моментом, так как выбранная ими система маркирования позволяла осуществить синхронизацию маркирования на различных уровнях [Bertrand, Blache, Espesser et al. 2006: 37—38].

В спонтанной речи к проблеме границ сегментов добавляется проблема идентификации реально произнесённых фонетических единиц. В ходе работы над УРБД CID эксперты часто транскрибировали сегменты, отсутствующие в сигнале; или, реже, звуки,

[1] Данный признак, связанный с законом антиципации (упреждения и захождения артикуляторных жестов) был описан намного ранее на материале других языков (см., например, Потапова Р.К., Линднер Г. Особенности немецкого произношения. М.: Высшая школа, 1991. 319 с.).

появляющиеся в речевом сигнале, оказывались не записанными экспертами, что не являлось следствием недостаточной квалификации экспертов. Процесс восприятия речи человеком включает реконструкцию отсутствующей или искажённой фонетической информации, и это является его фундаментальным свойством. Реконструкция реализуется вне уровня сознания человека и, таким образом, ускользает от самонаблюдения слушающего. Следовательно, эксперт, выполняющий транскрипцию, может включить в неё и те звуки, которые не были реально произнесены диктором, а только лишь должны были присутствовать в речи. Поэтому после транскрибирования речевого материала УРБД CID проводилась дополнительная коррекция.

В ряде случаев возникали проблемы сегментации спонтанной речи вследствие образования полизвукотипов, особенно на месте цепочек VCV, содержащих звонкий согласный.

Не менее сложной задачей вследствие наличия большого количества параметров, которые необходимо учитывать, является просодическое маркирование. Лишь немногие системы позволяют фиксировать достаточно полную совокупность просодических явлений. Так, TOBI и INSTINT в большей степени ориентированы на интонационные явления. Кроме того, «привязка» к французскому языку такой системы, как TOBI, осложняется её требованиями к априорным знаниям о фонологической структуре исследуемого языка. Напротив, одно из главных преимуществ INSTINT состоит в том, что эта система опирается на акустический анализ, не подразумевающий априори каких-либо знаний о фонологической системе языка. Как утверждают создатели INSTINT, эта система может быть использована в отношении любого языка [Di Cristo et al. 2004].

Что касается именно французского языка, как раз самые последние работы затрагивают вопросы связей между просодическими и дискурсивными явлениями, или же в них делаются попытки описания просодических вариаций, связанных, например, с региональными вариантами, а также предлагаются более полные системы маркирования в виде многоярусной решётки, позволяющие одновременно кодировать феномены общего характера (например, диапазон частоты основного тона) и локального характера (акцентуация, выделение и т.п.) на разных уровнях: временно́м, метрическом и интонационном. К числу таких систем относится, например, IVTS (адаптированная к системе IViE), которая позволяет фиксировать различные аспекты просодической вариативности.

Система маркирования, учитывающая совокупность просодических феноменов, оказывается необходимой в случаях, когда стоит задача установления связей между различными элементами хотя бы одного просодического уровня, не говоря уже о разных лингвистических уровнях. Выполнение этой задачи применительно к УРБД с объёмом звучащего материала в несколько часов оказывается слишком тяжёлым и дорогостоящим делом в отсутствие комплекса средств, автоматизирующих максимально возможное число этапов кодировки.

При разработке УРБД CID была выбрана система, сочетающая ручной и автоматический методы обработки, которая опирается в первом случае на слухо-

вую идентификацию экспертами частных просодических феноменов (подход, сходный с TOBI) и на подход MOMEL-INSTINT во втором случае [Hirst at al. 2000].

По этой причине при маркировании CID были использованы несколько видов маркеров. Во-первых, применялась нотация INSTINT, которая позволяет автоматически кодировать тональные целевые сегменты. Система INSTINT использует алгоритм MOMEL, позволяющий моделировать кривую $F_0$, и даёт на выходе последовательность лингвистически релевантных точечных целей. Система INSTINT имеет алфавит из восьми символов: Top, Middle и Bottom определяются в целом, по отношению к регистру каждого говорящего; Higher, Same и Lower определяются по отношению к предыдущим позициям, как и Downstepped и Upstepped, относящиеся к более слабым изменениям.

Другой набор знаков применялся при ручном маркировании. Один из них связан с просодической фразировкой высказываний, то есть с определением областей просодических единиц. Выделялись интонационные единицы (les unités intonatives) и единицы акцентуации (les unités accentuelles). Специальный маркер был введён для неоднозначных случаев или случаев, которые невозможно отнести к первой или второй категориям. Подобные случаи могут быть связаны с присутствием дискурсивных маркеров (таких как «что?» (quoi), «видишь ли» (tu vois) и т.п.).

Ряд маркеров относился к «интонационным контурам». Были приняты следующие символы: ровный / flat (fl), малый подъём / minor rising (mr); другие малые контуры / other minors (m0); нисходящие / falling (F); восходяще-нисходящие / rising-falling (RF1); восходяще-нисходящие с предпоследнего слога / rising-falling from penultimate (RF2); большой подъём, предполагающий продолжение / major continuation rising (MCR); терминальный подъём / terminal rising (TR); вопросительный подъём / question rising (QR); подъём при перечислении / enumerative rising (ER); падение при перечислении / enumerative falling (EF).

Принцип морфосинтаксического маркирования заключался в присвоении словам высказывания соответствующих категорий. Существует несколько систем, или классификаторов, позволяющих с большим или меньшим успехом автоматически выполнять эту задачу.

В частности, для французского языка известны такие системы, как WinBrill, Cordial и LPL. Последняя была применена в ходе работы над УРБД CID. Она использует стохастические данные, полученные при отработке обучающей выборки, для того чтобы определить, какие морфосинтаксические маркеры наиболее вероятны для того или иного высказывания. В настоящее время авторский коллектив CID работает над адаптацией классификатора LPL к задаче пополнения УРБД [Bertrand, Blache, Espesser, 2006].

Синтаксическое маркирование остаётся сложной задачей, с трудом поддающейся автоматизации. Несмотря на это, существует ряд автоматических анализаторов, которые могут использоваться, как минимум, в качестве основы для осуществления маркирования. В связи с этим, в зависимости от желаемого уровня описания, различают подходы двух типов. Самое простое маркирование (проставление скобок) может опираться только на методы поверхностного анализа (shallow parsing). Тем не менее, возможно также более тонкое маркирование, выполняемое с помощью более сложных анализаторов. Последние, в отличие от поверхностных анализаторов, позволяют идентифицировать не только единицы и их структуры, но также и синтаксические отношения, связывающие их одновременно с грамматическими функциями этих единиц. Для всех случаев (эта ремарка применима для всех представленных здесь уровней анализа) техника маркирования и используемые формальные приёмы независимы от теоретического

подхода, выбранного для синтаксического описания (например, HPSG, GP или грамматики соподчинённости).

Поверхностные анализаторы предоставляют информацию о границах составляющих текста. Этот тип анализа используется также в более широких областях: например, в области передачи информации, диалоговых систем, систем синтеза речи. Данный тип программы опирается на совокупность правил, определяющих левые и правые границы составляющих в зависимости от анализируемой составляющей и различных свойств читаемого слова.

С недавнего времени существуют несколько средств запроса к синтаксически маркированным УРБД как результат усилий, направленных на формирование синтаксически маркированных французских баз данных. Средства для формирования запросов в синтаксически маркированных УРБД позволяют учитывать соотношения по принципу вычленения основной информации.

Данное маркирование разработано в настоящий момент для уровня лексики и отношений между лексическими единицами. Речь идёт об упорядочивании элементов, релевантных для построения смысла дискурса, и одновременно о маркировке лексических единиц и связывающих их отношений. Составление подобного массива — первый шаг в формализации семантических и дискурсивных связей.

Семантическое маркирование, целью которого является упорядочение лексических и межлексических единиц, должно включать, как минимум, три информационных уровня:

— на первом уровне маркирования отмечаются семантические функции. Под этим подразумевается точная маркировка вклада лексической единицы (например, состояние, процесс и переход) в упорядочение событийной структуры фразы, которая позволяет отобразить особый уровень информации в лексической семантике. Если точнее, то она предназначена для классификации объектов мира. Структура имеет четыре функции, уточняющие семантические признаки единицы информации: конститутивные, формальные, целевые и агентивные. В соответствии с этим уточняется отношение объекта к его составляющим, характерные признаки объекта, функции объекта и активных участников, связанных с объектом.

На втором уровне отмечается информация онтологического типа. Эта совокупность семантических черт отбирается на основе предварительных лингвистических исследований, доказавших свою эффективность во многих языках.

Наконец, на третьем уровне отмечаются отношения между лексическими единицами. Речь идёт, таким образом, одновременно о фиксации отношений иерархического порядка (например, анафорических), но также (и в основном) об указании на то, как был получен смысл каждой единицы высказывания, что может выполняться с учётом взаимодействий той или иной единицы с другими полисемическими единицами высказывания. Поскольку семантическая единица формируется только в контексте, необходимо отметить вклад каждого уровня в формирование смысла.

Для того чтобы вести речь о прагматическом уровне, необходимо уточнить содержание этого термина. В литературе он ассоциируется с различными теоретическими направлениями, методологиями, с множеством самых различных объектов исследования. В данном случае этот термин охватывает три перспективы, определяющие уровни разного характера: например, языковые действия, феномены разговорного порядка, связанные с конструированием речевых оборотов, и другие уровни, также релевантные в нарративном аспекте.

Ниже приведён пример маркирования. Первый элемент, имеющий временной образец, служит индексом для других элементов, которым самим присваивается индекс по их позиции в структурном элементе, в который они входят.

```
<el индекс=«32» начало= «5.8588» конец= «6.0908»>
<имя-атрибут= «SpellSp1»>графемы</атрибут></el…
<el индекс=«26» начало= «32» конец= «32»>
<имя-атрибут= «Имя Нарицательное»>Нарицательное</ атрибут >
<имя-атрибут=Согласие>4</атрибут></el>…
<el индекс=«15» начало= «31» конец= «32»>
<имя-атрибут= «Именное предложение «>Стандарт</атрибут>
<имя-атрибут=Согласие>4</атрибут></el>…
```

Промаркированы также и другие типичные элементы устной речи. Первая линия маркирования относится к дискурсивным маркерам типа «что», «вот», «видишь ли», «знаешь ли», «наконец». Две другие линии маркирования касаются слуховых феноменов. Среди слуховых сигналов различаются простые и сложные сигналы: некоторые повторы, повторные формулировки, метавопросы, завершения и т.п. Одна линия маркирования посвящена формальным категориям («ммм», «ну да», «а, нет», «а, ну да», «согласен» и т.п.), другая линия посвящена функциональным категориям. Подразумевается, что деление на сложные категории уже включает функциональный аспект (минимальное выслушивание, взятие на заметку, оценка, суждение).

Последняя линия маркирования относится к типу единиц, используемых для учёта речевых оборотов: единицы построения оборотов речи / turn-constructional units (TCUs). TCUs — это единицы, считающиеся потенциально «завершёнными» с синтаксической, просодической и прагматической точки зрения. Конец TCU представляет собой место потенциального завершения, отсылающее к так называемому переходному месту / transition-relevance place (TRP).

TCUs являются, таким образом, наименьшими лингвистически незавершёнными и завершёнными единицами, релевантными на уровне коммуникации. Конечная TCU представляет собой оборот, состоящий из одной единицы, в то время как неконечная TCU — один из семантически или прагматически незавершённых компонентов сложного оборота, определённого, например, в терминах дискурсивной деятельности (каузальные конструкции, повествовательная последовательность и т.п.). CID маркирован с учётом конечных и неконечных TCU.

Если, в отличие от фонетического маркирования, совмещение дискурсивной деятельности с речевым сигналом и не является абсолютной необходимостью, то прагматическое маркирование подобных языковых феноменов ставит, тем не менее, ряд проблем. Первая проблема заключается в понимании самого термина «маркирование». Вторая проблема — в принятии во внимание фактора временной развёртки речевого высказывания, а следовательно, и границ наблюдаемого явления.

Размытость дискурсивных и нарративных границ и их контекстуальный фон (принятие в расчёт ситуации общения, скрытые обстоятельства, реакции говорящих и т.д.) делают вопрос маркирования проблематичным. Становится необходимым работать по возможности с наиболее полным источником информации, включая видеоряд. Полимодальный анализ на данном уровне имеет, следовательно, первостепенное значение, по крайней мере, по двум причинам: первая касается интерпретации исследуемых явлений, которая может быть лишь улучшена благодаря учёту всех прочих уровней, участвующих в формировании смысла, вторая связана с совершенствованием и обогащением процессов маркирования.

Невербальное маркирование УРБД находится в процессе разработки. В рамках исследования было использовано программное обеспечение ANVIL, учитывающее ручную жестикуляцию и выражение лица, движения головы, направление взгляда. Этот код был дополнен и адаптирован в соответствии с потребностями исследования (в частности, были введены обозначения для движений корпуса, а также уточнены эталоны для дейктических жестов) [Bertrand, Blache, Espesser 2006: 50—51].

Жесты, включённые в номенклатуру маркеров, перечислены в таблице 5.

*Таблица 5*

**Перечень жестов, маркированных в системе CID**

| Голова / Лицо | Руки | Корпус |
|---|---|---|
| Выражение лица | Симметрия / асимметрия жеста | Движения корпуса |
| Движение бровей | Траектория руки | |
| Открытие глаз | Конфигурация руки | |
| Направление взгляда | Семиотический тип жеста | |
| Открытие рта | Фазы жеста | |
| Конфигурация губ | Вершина | |
| Движения головы | Точка контакта | |
| Эмоции | Высота реализации жеста | |
| | Позиция жеста в пространстве жестикуляции говорящего | |

Жест включает разные фазы, такие как фаза подготовки (рука, например, покидает положение покоя и вступает в реализацию жеста), фаза собственно реализации жеста, затем фаза исхода (когда рука для того, чтобы возобновить тот же тип жеста, возвращается в позицию покоя). Перед конечной фазой жест может быть задержан. К этим фазам добавлен параметр, отмечающий экстремум жеста (точку, когда жест достигает своего максимального развёртывания по отношению к положению покоя). Наконец, точка контакта используется для адаптивных жестов, фиксирующих контакт между рукой и какой-либо частью корпуса либо говорящего, либо партнёра по коммуникации.

В настоящее время на основе этих стандартов создан файл (см. листинг 1). В приведённом фрагменте указано, что линия маркирования Eyebrows (движения бровями) входит в группу маркирования движений лица (group name=Face) и что для движений бровями можно встретить такие значения, как frowning (нахмуривание бровей) или raising (поднятие бровей). Введено также значение other (другое) для редких случаев, когда говорящий может, например, поднять одну бровь.

*Листинг 1*

| | |
|---|---|
| <?xml version=«1.0»кодировка=«ISO-8859-1»?><br><annotation-spec><br>— <head><br>— <valuetype-def><br>— <valueset name= «EyebrowsType «><br><value-el color=«#9df4a»>Frowning</value-el><br><value-el color=«#f1f07a»>Raising</value-el><br><value-el color=«#f5ce16»>Other</value-el><br></valueset><br></valuetype-def><br></head> | Данная первая строка описывает файл как файл xml<br>Наименование маркирования<br>В первой части файла даны все возможные в линии значения маркирования, здесь — *frowning, raising, other* |
| — <body><br>— <group name=«Face»><br>— <track-spec name=«Eyebrows»type=«primary»><br><attribute name> =«Eyebrows»<br>valuetype=«Eyebrows type» display=«true»/><br></track-spec><br></group><br></body><br></annotation-spec> | Во второй части файла описана линия *eyebrows* и указано к тому же, что линия эта первична (она не зависит от других линий), но входит в состав группы *face*, группы маркирования движений лица. Синтаксис XML требует, чтобы все скобки были закрыты. |

Создание файла позволило, в частности, продумать иначе структуру маркирования жестов: так, вместо того, чтобы создавать множество специфических маркеров, таких как:

*Таблица 6*

**Примеры маркеров**

| | | |
|---|---|---|
| Gaze > | Sideways > | left / right |
| Head > | Side turn > | left / right |
| | Single tilt > | left / right |
| Trunk > | Sideways > | left / right |
| Hands > | Single hand > | left / right |

— достаточно ввести один раз значение left / right, применимое к движениям любых частей тела. Если движение не предполагает параметра left / right, как например, eyebrow raising / frowning, то к ярлыку left / right по умолчанию применяется значение none. Таким же образом функционирует маркер concrete / abstract, который применим только к дейктическим жестам и направлению взгляда. Параметр же contact point относится только к приспособительным жестам, но, однако, присутствует и отмечен как none для всех других жестов.

Как только маркеры вводятся в ANVIL, программное обеспечение порождает файл XML, указывающий для каждой совокупности маркеров ранг, время начала и время конца.

Значения по умолчанию не учитываются в заключительном файле маркеров и, следовательно, его не загромождают ненужной для того или иного типа жеста информацией.

Помимо маркирования языка жестов, ANVIL позволяет группировать описанные выше совокупности маркеров других уровней. Данное программное средство позволяет не только импортировать совокупности маркеров, например, из PRAAT, но также их менять при условии предварительной спецификации ярлыков, используемых в спецификационном файле. Это привело к необходимости создания иерархии маркеров в рамках каждого уровня: например, последовательности маркеров, представленные в линейном виде в формате PRAAT, были построены в соответствии с иерархией, отвечающей требованиям структуры формата XML, входного и выходного формата ANVIL. Что касается других уровней, маркирование которых не выполнялось в PRAAT, то они выполнены в формате XML, при этом маркеры также были специфицированы в файле spec.

Следует отметить, что формат ANVIL предусматривает возможность создания стольких файлов spec, сколько требуется, используя тот или иной уровень маркирования, а также тот или иной набор тэгов. Возможна также последующая непосредственная модификация этих файлов, поскольку они создаются в формате XML.

Интерес к применению комплексного подхода в маркировании УРБД (в частности, многомодальных УРБД) основан на перспективе самых широких областей применения этих данных. В частности, предусматривается возможность запросов высокого уровня с привлечением различных областей маркирования, позволяющих более систематично исследовать типы взаимодействия, которые могут иметь место между ними. Необходима временна́я синхронизация, если маркированные объекты чётко идентифицированы и отсегментированы. Вопрос синхронизации — это главная проблема мультимодальных УРБД: как установить связь между объектами, принадлежащими к разным областям, без единой справочной базы? В ходе работы над УРБД CID было выбрано решение, опирающееся на позиционирование каждого объекта с помощью специальной системы. Данный принцип заключается в указании, если это возможно, для каждого объекта нескольких ориентиров. Разумеется, большинство объектов, напрямую связанных с речевым сигналом, имеют соотношение с временно́й позицией. Равным образом, как указано выше, некоторые объекты из других областей могут быть также «привязаны» к временно́му сигналу, что касается, например, слов и морфосинтаксических конструкций. В этом случае может быть предложен также другой тип привязки: «позиция» объекта в цепи (например, ранг слова в последовательности или фразе). Такой тип более традиционен для обработки письменного материала, но в данном случае он применим и к устно-речевому дискурсу. Наконец, для включения в комплекс семантических данных предложено использование индексации элементов речи. В целом, ориентировка каждого элемента может проводиться через комплексную «привязку», поясняющую все три составляющие (временну́ю реализацию сигнала, положение в речевой цепи, индекс контекста) или их часть.

После того как этот процесс закончен, становятся возможными запросы в УРБД с использованием различных уровней маркирования. В этом случае речь

идёт о запросах одновременно по нескольким областям при синхронизации поисков через комплексную систему «привязок». Следующие примеры запроса иллюстрируют функционирование системы:

**Q0 Найти паузы внутри синтаксической единицы**

Требуемые области — синтаксис и просодия. Поиск заключается в идентификации объектов особого просодического типа, границы которых включены в границы объекта, принадлежащего к другой области — синтаксису. В этом случае, временные и позиционные «привязки» позволят идентифицировать объекты.

**Q1 Найти дейктические жесты, ассоциируемые с местоимением**

Данный запрос относится одновременно к области жестов и морфосинтаксической области. Он заключается в идентификации в одной области (жестов) всех объектов особого типа (дейктических) и выявлении перекрытия этой области объектами типа «местоимение» в определённых пределах (интервалы которых включены в пределы интервалов объектов, относящихся к жестам).

**Q2 Найти синтаксические субъекты**

В данном случае отобранные объекты должны иметь синтаксическую характеристику (быть фрагментами), семантическое свойство (быть субъектами) и принадлежать к особому просодическому контуру (например, с восходящей мелодикой). Здесь полезны «привязки» временные и позиционные.

**Q3 Найти каналы обработки связи**

Цель данного запроса — идентификация особых речевых маркеров, реализуемых говорящим и слушающим.

Возможность осуществления подобных запросов позволяет, таким образом, весьма полно использовать ресурсы мультимодальной УРБД. Такой тип исследования в УРБД большого объёма на сегодня не имеет аналогов, которые содержали бы достаточно полное многоуровневое маркирование. Подход, применённый в УРБД CID, представляет собой эффективное решение, позволяющее синхронно отображать маркеры различных уровней. Таким образом, речь идёт о важном вкладе в анализ мультимодальных систем.

Рост числа лингвистических работ, посвящённых просодии, речи и грамматике, с одной стороны, а также развитие исследовательских подходов к УРБД, с другой стороны, побуждает к размышлениям о символическом и дискретном отображении непрерывных просодических элементов. Вопрос кодировки, символического и дискретного отображения супрасегментных единиц не является нейтральным, особенно если им задаваться с целью установления параллели с отображением сегментного уровня, предложенным IPA (International Phonetic Association и International Phonetic Alphabet — Международная фонетическая ассоциация и международный фонетический алфавит). Данная система кодировки опирается на несколько гипотез, касающихся континуума речи и его анализа, а именно: а) одни аспекты речи лингвистически релевантны, в то время как другие — нет. Это затрагивает вопрос о возможности выполнения широких (или фонематических) транскрипций и узких (или фонетических) транскрипций; б) континуум речи может быть частично отображён как последовательность сегментов.

Важно отметить, что единственное общеизвестное допущение, которое делает IPA, — это то, что континуум речи может быть отображён в форме дискретных сегментов. Однако использование чёткого символа для отображения звука или сегмента не обязательно превращает последний в фонему.

Один из источников ошибок связан с трудностями задачи транскрибирования. Действительно, транскриптор может по-своему кодировать услышанные звуки, делать выбор, объясняемый либо знанием фонологической системы транскрибируемого языка, либо незнанием этого языка. В некоторых случаях он может выбрать транскрипцию сегментов не в том виде, в каком они были действительно произнесены, а как образцы фонем. Следовательно, для разработки или оценки любой системы транскрипции просодических единиц необходимо ввести систему ограничений и провести сопоставление четырёх систем транскрибирования просодических единиц: IPA и специфические символы кодировки супрасегментных единиц, INSTINT, ToBI и IVTS — систему транскрипции, разработанную на основе IViE. При этом ставится двойная цель:
— представить каждую систему транскрипции и
— оценить способы, с помощью которых эти различные системы могут транскрибировать некоторые типы особых просодических явлений, в частности, акцентуацию, мелодику и тональность [Delais — Roussarie, Post, Portes 2006: 63—65].

Как и ToBI, система IVTS использует несколько категорий маркирования для кодировки различных просодических и лингвистических данных. В IVTS транскрипция организована на шести уровнях, из которых четыре используются для записи просодических феноменов. Транскрипция, таким образом, принимает следующую форму:

*Таблица 7*

**Уровни транскрибирования в системе IVTS**

| |
|---|
| { Категория «Комментарии» (или *Comments tier*) |
| { Категория «Фонология» (или *Phonological tier*) |
| { Категория «Общее фонетическое восприятие» |
| Декодирование просодической информации |
| (или *Global auditory phonetic tier*) |
| { Категория «Локальное фонетическое восприятие» |
| (или *Local auditory phonetic tier*) |
| { Категория «Ритм» (или *Rythmic tier*) |
| { Категория «Слово» (или *Orthographic tier*) |

Каждый из уровней служит для кодировки отдельных данных. Категория «Слово» используется для совмещения произнесённых слов с участками сигналов, которые им соответствуют. В других категориях обозначения совмещаются с определёнными точками сигнала, например:
а) участком слога, воспринимаемого как выделенный;
б) границами интонационных областей и т.п.

В категории «Комментарии» точки, взятые для совмещения с сигналом, соответствуют зонам, к которым относятся комментарии. Различные виды принятых совмещений зависят от их релевантности для обрабатываемого языка и от характера работы привязки мелодических контуров.

В категории «Ритм» обозначение P указывает на то, что отмеченный слог выделен сильнее, чем соседние слоги. На акустическом уровне это

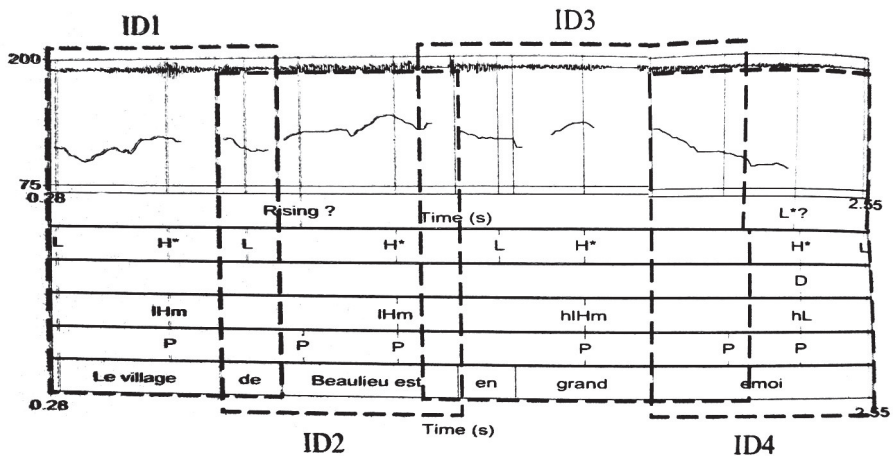


*Рис. 4. Сегментация с учётом реализации акцентуации (ID) фразы «le village de Beaulieu est en grand émoi» (деревня Больё в большом волнении)*

может характеризоваться увеличением длительности, мелодическим контуром и т.п. Отметим, что P указывает на воспринимаемое качество, но не обязательно на абстрактное структурное свойство слова или группы слов с лексическим ударением. Кроме того, обозначения P присваиваются слогам, на которых реализуются особые мелодические движения, — их предстоит совместить с обозначениями категорий «Локальное фонетическое восприятие» и «Фонология». При этом последняя операция не является обязательной.

Категория «Локальное фонетическое восприятие» используется для записи формы мелодических движений, выполненных на выделенных слогах, а также на примыкающих к ним слогах (см. рис. 4).

Здесь акцент делается на мелодической конфигурации и характере совмещения отдельных участков. Мелодические изменения более общего плана, такие как регистр или нисходящая мелодика, на данном уровне не кодируются. Транскрипция мелодических изменений проводится на перцептивно-слуховой базе, а не на основе акустического анализа частоты основного тона. Она выполняется при внимательном прослушивании части сигнала, соответствующего области реализации акцентуации (ID). Протяжённость этой области варьирует в зависимости от языков. Во французском языке любая ID включает а) выделенный слог, обозначенный P; б) все предшествующие ему слоги до предыдущего выделенного слога или до границы основной интонационной области; в) непосредственно следующий за ним слог. Согласно этому определению высказывание «le village de Beaulieu est en grand émoi» (деревня Больё в большом волнении) делится на четыре ID, по одной на каждый выделенный слог, отмеченный мелодическим движением.

Вначале система транскрипции IPA была разработана для кодирования сегментной информации. При этом был предложен набор символов для кодировки некоторых просодических феноменов как метрического, так и тонального характера. Для кодировки метрических феноменов IPA предлагает два различных символа, «ˈ» и «ˌ»: первый для отображения слогов, получающих первичное ударение, второй для слогов, получающих вторичное ударение. Другая серия используется для отображения просодических при-

знаков. Приняты два уровня структуризации: основание (или малая группа), представленная символом « | », и интонационная (большая) группа, представленная символом « || ». Для феноменов тонального характера в IPA имеются две серии символов: одна для статических тонов, другая для модулированных тонов. Эти символы созданы для транскрибирования лексических тонов в таких языках, как китайский; однако они не позволяют выполнять кодировку интонационных фразовых явлений. В этом случае могут быть использованы некоторые символы широкого значения: символ нисходящей шкалы $ и восходящей шкалы #, символы нисходящей мелодики ( и восходящей мелодики &.

Для маркирования ударных слогов IPA предлагает фиксировать различие между первичным и вторичным ударениями, то есть привлекать фонологический уровень. Такой подход предполагает, что ритмическое функционирование языка известно, а значит, можно определить, является ли физическая выделенность первичным или вторичным ударением. Следовательно, на сегментном уровне возможности, связанной с различением между широкой и узкой транскрипцией, здесь, по-видимому, не существует.

Вторая трудность соотносится с применением символов сегментации на просодические группы. IPA предлагает установить различие между двумя уровнями просодической структуризации: малой группой и интонационной группой. При этом ничего не говорится о критериях определения этих групп [Delais Roussarie et al. 2006].

Изучение различных примеров в Handbook of the International Phonetic Association (1999) показывает, что дело обстоит отнюдь не так гладко. В каталанском языке выбор между двумя уровнями кажется подчинённым использованию или неиспользованию терминального контура: если последовательность завершается продолжением, то граница между малыми группами используется, в противном случае используется граница между большими группами. В транскрипции французского языка границы между большими группами используются как после контура продолжения, так и после финального контура. Следовательно, выбор между тем или другим символом оказывается часто делом тонким и зависит от транскриптора. Отметим, впрочем, что эта проблема сегментации существует постоянно в большинстве систем транскрипции, за исключением, возможно, ToBI и break indices, образующих более прочный и надёжный инвентарь.

Институт фонетики Экс-ан-Прованса (Франция) предлагает теоретическую модель и инструменты маркирования интонационных систем. Оригинальность подхода заключается в том, что предложено средство автоматического маркирования на двух уровнях: уровне «фонетической» репрезентации, порождённой алгоритмом MOMEL (Modeling MELody), и уровне «поверхностно-фонологической» репрезентации, автоматически выполняемой с помощью алфавита INSTINT (International Transcription System for INTonation).

Алгоритм MOMEL и «фонетическая» репрезентация частоты основного тона преобразуют прерывистую кривую — следствие огрублённого распозна-

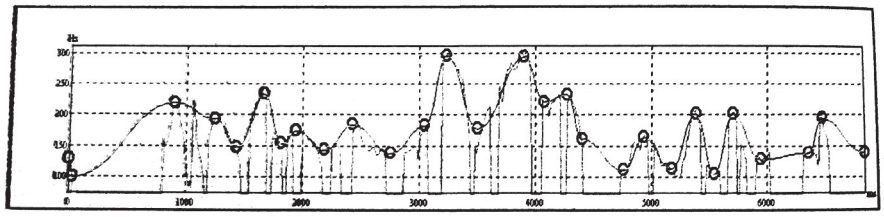

*Рис. 5. «Целевые точки» и сплошная кривая, порождённая алгоритмом MOMEL на основе выделения частоты основного тона*

вания частоты основного тона — в сплошную кривую, которая является интонационно релевантной. Роль алгоритма заключается в отделении «макропросодической» составляющей от «микропросодической» составляющей, которая не принимается во внимание как лингвистически нерелевантная. На выходе MOMEL производит совокупность точек, охарактеризованных с помощью пары «временна́я локализация / $F_0$». Эти точки затем объединяются. Зоны отклонения представляются вершинами и впадинами.

Алфавит INSTINT включает набор из восьми абстрактных тональных символов. Три из этих символов кодируют «абсолютные» тоны, разграничивающие общую протяжённость регистра говорящего на отрезке «интонационной единицы» (соответствующей максимальной единице просодической фразы для данной модели): речь идёт о символах T для Top, M для Mid и B для Bottom. Пять остальных символов кодируют «относительные» тоны, значение которых зависит от значения предыдущего тона. Относительные тоны делятся на две подкатегории: неитеративные тоны (H — для высоких, S — для монотонных и L — для низких) и итеративные тоны (U — для повышения и D — для понижения). Помимо орфографических символов второй набор диакритики используется преимущественно в рамках транскрипции текстов.

Возможно получение автоматической кодировки речевых данных INSTINT с помощью алгоритма, учитывающего два дополнительных параметра: ключ и регистр. И тот, и другой зависят одновременно от говорящего и высказывания. Абсолютные тона T и B определены как границы тонального регистра говорящего, симметрично распределённого вокруг ключа, характеризующего значение тона M.

* * *

В заключение следует подчеркнуть, что языковые ресурсы являются важнейшим компонентом процесса создания и эксплуатации различного рода информационных систем, реализующих лингвистические функции, направленные на обработку естественного языка в его различных проявлениях (применительно к печатным и рукописным текстам, а также к звучащей речи).

В области корпусной лингвистики современные компьютерные технологии ускоряют и упрощают процедуры лингвистической обработки больших массивов текстов в их письменном и устном вариантах. По сути, лингвистический корпус — это своего рода информационно-справочная система, основанная на текстовых ресурсах на некотором языке в электронной форме при наличии особой дополнительной информации о свойствах лингвистического материала.

Таким образом, корпусная лингвистика — это бурно развивающаяся отрасль «лингвистической индустрии», предназначенная как для проведения научных исследований, так и для решения целого ряда прикладных задач.

## Литература

1. Автоматизированное рабочее место эксперта-фоноскописта. Электронная энциклопедия, версия V1.0: http://www.estra.ru
2. *Андрющенко В.М.* Концепция и архитектура Машинного фонда русского языка. М., 1989.
3. *Белолипецкий С.И., Буря А.Г.* Специализированные СУБД для поддержки речевых баз данных // Сетевой электронный научный журнал «Системотехника». №2. 2004. М.: МГИЭМ, 2004.
4. *Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С.* Технология формирования речевых баз данных // В сб. «Организационное управление и искусственный интеллект». М.: Эдиториал УРСС, 2003.
5. *Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В.* База речевых фрагментов русского языка ISABASE // В сб.: «Интеллектуальные технологии ввода и обработки информации». М., Эдиториал УРСС, 1998.
6. *Богуславский И.М., Григорьев Н.В.* и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара ДИАЛОГ-2000. М., 2000. Т. 2. С. 41—47.
7. Корпусная лингвистика в России. / Сост. *Е.В. Рахилина* и *С.А. Шаров* // Спец. выпуск журнала НТИ. М., 2003. Сер.2. № 6, 10.
8. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование). М.: МГУ. [http://www.dialog-21.ru/archive_article.asp].
9. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование) // Сборник трудов Международного семинара Диалог'2001 по компьютерной лингвистике и её приложениям (в двух томах). Т. 2. Прикладные проблемы. М., 2001.
10. *Кривнова О.Ф.* Области применения речевых корпусов и опыт их разработки // Сборник трудов XVIII сессии РАО. М.: ГЕОС, 2006.
11. *Леонтьева Н.Н.* Автоматическое понимание текстов: Системы, модели, ресурсы. М.: Академия/Academia, 2006.
12. *Потапова Р.К., Линднер Г.* Особенности немецкого произношения. М.: Высшая школа, 1991. 319 с.
13. *Потапова Р.К.* Лингвистическое обеспечение Электронной Энциклопедии, предназначенной для экспертов-фоноскопистов (русский язык). М.: ЭСТРА, CDROM, 1998–1999.
14. *Потапова Р.К.* Новые информационные технологии и лингвистика. 4-е изд., сущ. доп. — М.: Эдиториал УРСС, 2005. — 368 с.
15. *Потапова Р.К.* Речь: коммуникация, информация, кибернетика. М.: Радио и Связь, 1997. 528 с.
16. *Потапова Р.К..* Тайна современного кентавра. Радио и связь, М., 1992.
17. *Рыков В.В.* Корпус текстов — новый тип словесного единства // Труды Международного семинара ДИАЛОГ-2003. Протвино, 2003.
18. *Сичинава Д.В.* К задаче создания корпусов русского языка в Интернете // НТИ. М., 2002. Сер.2. № 12.

19. *Скрелин П.А., Щербаков П.П.* Требования к современной фонетической базе данных для фундаментальных и прикладных исследований // Технологии информационного общества — Интернет и современное общество: труды VI Всероссийской объединенной конференции. Санкт-Петербург, 3—6 ноября 2003 г. СПб.: Изд-во Филологического ф-та СПбГУ, 2003. С. 62—63.
20. *Шаров С.А.* Параметры описания текстов корпуса, а также Корпусная лингвистика в России // НТИ. М., 2003. Сер.2. № 5—6.
21. *Arlazarov V.L., Bogdanov D.S. Krivnova O.F., Podrabinovitch A.Ya.* Creation of Russian Speech Databases: Design, Processing, Development Tools. // International Conference SPECOM'2004. Proceedings. S-Pb. Russia, 2004. Pp: 650—656.
22. *Barlow M.* Corpora for Theory and Practice. //1JCL. Amsterdam, 1996. № 1.
23. *Bel B., Blache P.* Le Centre de Ressource pour la Description de l'Oral. // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.13—18.
24. *Bertrand R., Blache P., Espesser R., Ferre G., Meunier C., Priego-Valverde B., Rauzy S.* Le CID — Corpus of Interactional Data: Protocoles, Conventions, Annotations. // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.31—60.
25. *Bohmova A.* Automatic Procedures in Tectogrammatical Tagging. //The Prague Bulletin of Mathematical Linguistics. Prague, 2001. №76. P.23—34.
26. *Collier A., Pace y M., Renouf A.* Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora. // Proceedings of the Sixth Workshop on Very Large Corpora. Montreal, 1998.
27. *Delais — Roussarie E., Post B., Portes C.* Annotation prosodique et typologia. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence. Vol.25, 2006. P. 61—95.
28. *Greenstette G., Segond F.* Multilingual Natural Language Processing // IJCL. 1997. V.2. № 1.
29. *Hajicovd E., Pajas P., Vesela K.* Corpus Annotation on the Tectogrammatical Layer: Summarizing of the First Stages of Evaluations // The Prague Bulletin of Mathematical Linguistics. Prague, 2002. №77. P. 5—18.
30. International Journal of Corpus Linguistics (IJCL). / Ed. W.Teubert. — Amsterdam, 1996—2001.
31. *Kibkalo A.A., Lotkov M.M.* Choice of Phonetic Alphabet for Russian LVCSR System // Proceedings of the International Workshop «Speech and Computer» SPECOM' 2003. (Moscow, 27—29 October, 2003) Moscow: MSLU, 2003. P. 102—105.
32. *Kucova L., Hajic ova E.* Prague Dependency Treebank: Enrichment of the Underlying Syntactic Annotation by Coreferential Mark-Up // The Prague Bulletin of Mathematical Linguistics. Prague, 2004. №81. P. 23—34.
33. *Lee Y.-J., Choi D.-L., Um Y., Lee K.-H., Kim Y.-I., Kim B.-W.* Speech Resources at SITEC in Korea // Proceedings of the 10th International Conference SPEECH and COMPUTER (SPECOM' 2005) (Patras, Greece, 17—19 October, 2005) Patras, Moscow: MSLU, 2005. P. 579—582.
34. *Loseva E., Potapova R.* Speech variability of vibrants: phonetic database for English and German // Proceedings of the 10th International Conference Speech and Computer SPECOM' 2005, Patras, Moscow: MSLU, 2005.
35. *Marcus M.P., Santorini B., Marcinkiewicz M.A.* Building a Large Annotated Corpus of English: The Penn Treebank // Computational Linguistics. 1993. Vol.19. №2. P. 313— 30.
36. *Potapova R.K, Potapov V.V.* Database of forensic phonetics knowledges (as applied to electronic encyclopaedia for Russian experts) // Proceedings of the International Conference of IAFP, York, UK, 1999. P. 6—7.
37. *Shaikevich A.* The Computer Fund of Russian Language // IJCL.-Amsterdam, 1997. V.2. №1. P. 163—167.
38. *Teubert W.* Corpus Linguistics and Lexicography // IJCL. Philadelphia, 2001.
39. http://www.mdi.ru/aspnews/body/03.12.2001_39303.html
40. http://cfrl.ru
41. http://conf.infosoc.ru/03-r2f14.html

42. http://www.auditech.ru
43. http://www.auditech.ru
44. http://www.mdi.ru/aspnews/body/03.12.2001_39303.html

*Родмонга Кондратьевна Потапова*
*Академик Международной академии информатизации, доктор филол. наук, профессор. Заслуженный работник Высшей школы РФ.*
*Зав. отделением прикладной лингвистики, зав. кафедрой прикладной и экспериментальной лингвистики, директор Центра фундаментального и прикладного речеведения Московского государственного лингвистического университета. Специалист в области романо-германского языкознания, общей и прикладной фонетики, теоретической, прикладной, экспериментальной и математической лингвистики. Автор свыше 450 научных и научно-методических публикаций.*