

Научная статья
УДК 81'33
DOI: 10.18101/2305-459X-2020-4-23-29

ПРИНЦИПЫ ПРОСОДИЧЕСКОГО АННОТИРОВАНИЯ В ЗВУКОВОМ КОРПУСЕ БУРЯТСКИХ ДИАЛЕКТОВ

© Абаева Юлия Догоржаповна

кандидат филологических наук, научный сотрудник,
Институт монголоведения, буддологии и тибетологии СО РАН
Россия, 670047, г. Улан-Удэ, ул. Сахьяновой, 6
julaba@yandex.ru

Аннотация. В работе рассматривается один из ключевых моментов аннотирования звукового корпуса — просодическая транскрипция. Анализ имеющихся корпусов русского языка показал, что в звуковых корпусах, нацеленных на исследование данных устной речи, обязательно используется просодическая транскрипция той или иной степени подробности. В результате изучения системы аннотирования в корпусах русского языка выявлено, что в некоторых из них используются интонационные транскрипции, созданные еще в докорпусный период, но не потерявшие своей актуальности, например системы Е. А. Брызгуновой, С. В. Кодзасова. А также и новые разработки, например система дискурсивной транскрипции корпуса «Рассказы о сновидениях» и др. Цель настоящего исследования заключается в анализе имеющихся просодических транскрипций, выявлении основных принципов транскрибирования. В результате анализа впервые в бурятском языкознании разработана просодическая транскрипция для аннотирования звукового корпуса бурятских диалектов. Актуальность работы обусловлена тем, что звуковой диалектный корпус является необходимым средством сохранения диалектных особенностей, а также средством изучения современного состояния бурятских диалектов и тем самым вносит вклад в развитие современной бурятской диалектологии.

Ключевые слова: звуковой корпус; диалектный корпус; интонация; просодическая транскрипция; просодическая аннотация; синтагма; интонационный центр; тональный акцент; мелодика; темп; пауза.

Благодарность. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-012-00491а.

Для цитирования

Ю. Д. Абаева. Принципы просодического аннотирования в звуковом корпусе бурятских диалектов // Вестник Бурятского государственного университета. Язык. Литература. Культура. 2020. Вып. 4. С. 23–29.

Актуальность создания звукового корпуса бурятских диалектов связана, во-первых, с недостаточным количеством современных исследований в области бурятской диалектологии. Во-вторых, в связи со сложившейся языковой ситуацией, когда с каждым поколением снижается уровень владения родным языком. Между тем современная лингвистика располагает средствами, которые позволяют хранить и систематизировать большие объемы звучащих тек-

стов, которые, с одной стороны, являются отличной базой для лингвистических исследований, с другой стороны, позволяют сохранять и фиксировать звуковую материю языка.

Основой звуковых корпусных исследований, помимо качественно собранного языкового материала, является качественно проведенная его обработка: сегментация и аннотирование. От аннотации зависит дальнейшее использование корпуса, его исследовательские возможности. Поскольку описываемый звуковой корпус задумывался нами как средство исследования прежде всего суперсегментных характеристик звучащей речи, а именно интонации, большое внимание было уделено разработке принципов просодического аннотирования, созданию транскрипции для бурятской интонации.

В современном языкознании накоплен значительный опыт создания звуковых корпусов и, соответственно, принципов их аннотирования. В наших работах мы в первую очередь опираемся на опыт отечественного языкознания, поэтому анализ способов просодического аннотирования и транскрибирования проводился на имеющихся звуковых корпусах русского языка. Их анализ показал, что в ряде корпусов используются просодические транскрипции, созданные в еще «докорпусный» период. Это прежде всего система интонационных конструкций (ИК) Е. А. Брызгуновой, а также комбинаторная модель фразовой просодии С. В. Кодзасова, которые слегка модифицируются, но до сих пор не потеряли своей актуальности.

По классификации Е. А. Брызгуновой выделяется семь типов ИК на основе оппозиций высказываний с одинаковым синтаксическим строением и лексическим составом или с разным синтаксическим строением, но одинаковым звуковым составом словоформ [9, с. 99]. Кроме того, Е. А. Брызгуновой была разработана и просодическая транскрипция на основе ИК для «наглядно-слухового представления звучащей речи» [2, с. 5], которая использовалась при обучении русскому произношению иностранных студентов. Основными принципами транскрипции были членение высказывания на синтагмы, обозначение интонационного центра и указание на тип мелодического движения (ИК) на интонационном центре. При детализации транскрипции обозначались участки изменения темпа, увеличение длительности звуков, паузы внутри ИК и между фразами, случаи слабо выраженной ИК и др. [2, с. 6].

В настоящее время система ИК Е. А. Брызгуновой в дополненном и расширенном виде применяется в корпусе русской спонтанной речи CoRuSS (Corpus of Russian Spontaneous Speech), создаваемом на кафедре фонетики и методики преподавания иностранных языков СПбГУ [5]. Просодическая транскрипция, применяемая в этом корпусе, была модифицирована Н. Б. Вольской и содержит 13 типов интонаций, каждый из которых имеет свои подтипы [4]. На орфографической записи отмечаются границы между синтагмами, дополнительная просодическая выделенность, тип мелодического движения (цифра по классификации Н. Б. Вольской), основное и побочное ударение в словах. Кроме того, отмечаются явления, характерные для спонтанной речи: паузы-хезитации, фальстарты и самокоррекции.

В основе системы интонационной транскрипции С. В. Кодзасова лежит идея о невозможности свести все разнообразие интонаций к некоторому ограниченному количеству интонационных моделей [7, с. 124]. Задачей разработанного им комбинаторного метода является выявление элементарных просодических компонентов, которые, сочетаясь друг с другом, дают некоторое количество просодем. В этой системе отдельно отмечаются акцентные и синтагменные (интегральные) просодии. Помимо этого, обозначаются выделительные акценты, удлинения или сокращения звуков, межсинтагменные паузы. Выглядит это следующим образом: на орфографической записи делаются пометы в виде квадратных скобок и букв для обозначения интегральных просодий, и цифр после гласного, которые указывают на вокалические центры слов, несущих тональный акцент.

Далее даются характеристики для каждой из просодий. Для интегральной просодии указываются: регистр тона, темп, громкость, фонация, случаи редукции. Для тональных акцентов указываются направление (комбинации восходящего, нисходящего и ровного движения), регистр (если он выше или ниже среднего), интервал (если он больше или меньше нейтрального, равного примерно 6–7 полутонам), локализация (на гласном, согласном, на цепочке слогов и др.), а также гортанная смычка, скрипучий голос, придыхание, двуморовость гласного сегмента.

Эта система просодической разметки была использована в одном из первых корпусов «Фонетической базе данных ИРЯ РАН» (1992–1995 гг.) [7, с. 124]. В настоящее время применяется в современных корпусах: Базе данных «Интонация русского диалога» (ИРД) [1], в Русском интонационном корпусе (RINCO) [10].

Еще одна транскрипция русской интонации ТОРИ (Transcription of Russian intonation) была создана С. Оде [6]. В работе используются термины авто-сегментной школы, разработанные для английского языка и широко используемые в других, преимущественно европейских языках: H — высокий, M — средний, L — низкий тональные уровни и их сочетания. В этой транскрипции описываются перцептивно релевантные движения высоты тона, достигающие перцептивно релевантных тональных уровней [6, с. 245]. Помимо движения тона на тональном акценте, указываются тональные уровни в начале и в конце синтагмы, а также изменения высоты тона, не придающие слову выделенности.

В последние годы активно создаются звуковые корпуса спонтанной устной речи. Хотя и не все из них нацелены конкретно на исследование суперсегментных характеристик, однако в них обязательно присутствуют элементы просодической разметки.

К таким корпусам можно отнести «Рассказы о сновидениях и другие корпуса звучащей речи»¹

¹ URL: <http://spokencorpora.ru> (дата обращения: 10.09.2020).

В его основе лежит концепция описания и дискурсивного аннотирования данных живой устной речи. Для этой цели была разработана система транскрипции, называемая дискурсивной, которая используется для графической записи звукового сигнала.

В корпусе представлено три уровня транскрипции. При минимальном уровне транскрибирования отмечаются два основных феномена, без которых, по мнению авторов, дискурсивная транскрипция не может являться таковой. Это сегментация речевого потока на минимальные отрезки, называемые в работе элементарными дискурсивными единицами (ЭДЕ), и указание на роль этих единиц в дискурсе (завершенность, незавершенность и другие типы иллюкативных значений). В более подробных транскрипциях обозначается направление движения тона на главном слове ЭДЕ, а также паузы с указанием на их большую или меньшую длительность. В полной транскрипции помимо главного акцента указывается направление движения тона на словах, произносимых с акцентным выделением, так называемых вторичных акцентах. Кроме того, обозначаются сниженный тональный регистр, редукция, эмфатическое выделение, ускоренный темп и др. [8].

Звуковой корпус «Один речевой день», разрабатываемый в СПбГУ, имеет целью исследование живой устной речи и повседневной коммуникации. Сбор материала осуществлялся путем непрерывной записи речи, производимой диктором в течение суток. Акцент делается именно на естественность речи, произведенной в привычной обстановке и с обычными коммуникантами. Аннотация записанного материала начинается с сегментации орфографической записи речи, причем «членение реплик на фразы и синтагмы осуществляется с учетом интонационно-синтаксических характеристик отрезков звуковой цепи» [12, с. 68]. В аннотировании учитываются следующие просодические характеристики: обязательно указывается завершенность (конец фразы повествовательного, вопросительного или восклицательного характера) или незавершенность (неконечная синтагма, обрыв фразы) высказывания; обозначаются паузы и паузы-хезитации с учетом их длительности и заполненности звуками; участки растягивания слов.

Эти же принципы взяты за основу при аннотировании диалектных корпусов народов России, например калмыцкого [11] и башкирского [3] языков.

При разработке системы аннотирования в звуковом корпусе бурятского языка мы опирались на описанные выше разработки, слегка изменяя и комбинируя их. За основу была взята система аннотации, используемая в корпусе «Рассказы о сновидениях...». Однако были включены некоторые дополнительные характеристики. Аннотирование звукового корпуса бурятских диалектов, как и большинства современных корпусов, осуществляется в программе ELAN¹,

¹ ELAN (версия 5.9) [компьютерная программа]. (2020). Неймеген: Институт психолингвистики имени Макса Планка. URL: <https://archive.mpi.nl/tla/elan> (дата обращения: 10.09.2020).

позволяющей создавать многоуровневые аннотации. Для анализа суперсегментных характеристик используется программа PRAAT¹.

Первоначально проводится орфографическая расшифровка записанного материала. Далее, уже на орфографической записи указываются элементы просодического аннотирования. При создании просодической транскрипции для бурятского языка в первую очередь нужно было определиться с уровнями аннотирования и теми характеристиками, которые должны быть обязательно отражены в этой транскрипции. Пока за основу взяты следующие уровни:

Phrase. Как показал анализ имеющихся транскрипций, в первую очередь речевой поток должен быть сегментирован на минимальные отрезки, которые в некоторых работах называются привычным термином «синтагма», в других либо интегральная просодия (комбинаторный метод С. В. Кодзасова), элементарная дискурсивная единица (корпус «Рассказы о сновидениях...»), реплика (корпус «Один речевой день»). Этот уровень выглядит как орфографическая запись речи, поделенная на синтагмы. В конце каждой синтагмы ставится пунктуационный знак, указывающий на иллокутивную характеристику данного отрезка: точка, запятая, вопросительный или восклицательный знак, многоточие. А также обозначаются паузы: абсолютные (...), паузы-хезитации (э-э, м-м).

В программе ELAN в целях быстрого поиска нужных характеристик в дальнейших исследованиях необходимые характеристики высказывания выносятся в отдельный уровень.

Punct. Так, знак препинания обозначается отдельным уровнем. Используются термины, употребляемые в «Рассказах о сновидениях»: COMMA — запятая, PERIOD — точка, DOTS — многоточие, COLON — двоеточие и др.

Pause. Обозначаются паузы А — абсолютные, э-э, м-м — заполненные паузы-хезитации.

Word. Помимо деления на синтагмы на этом уровне приводится текст в орфографической записи с делением на слова.

Accent. Движение тона на словах с выраженным акцентом. Используются условные знаки восходящий тон /, нисходящий тон \, ровный —, их комбинации.

Main accent. Указывается главный акцент в синтагме, являющийся интонационным центром.

Level. Уровень тона указывается в терминах автосегментной школы: Н — высокий, М — средний, L — низкий. Уровень тона обозначается в начале и конце высказывания, на интонационном центре и на участках второстепенных акцентов, на которых происходит заметное изменение высоты тона.

Tempo. Участки ускоренного (fast) или замедленного (len) произнесения.

Loudness. Участки повышенной (loud) или пониженной громкости (silent).

¹ Praat: Doing Phonetics by computer. URL: <http://www.praat.org> (дата обращения: 10.09.2020).

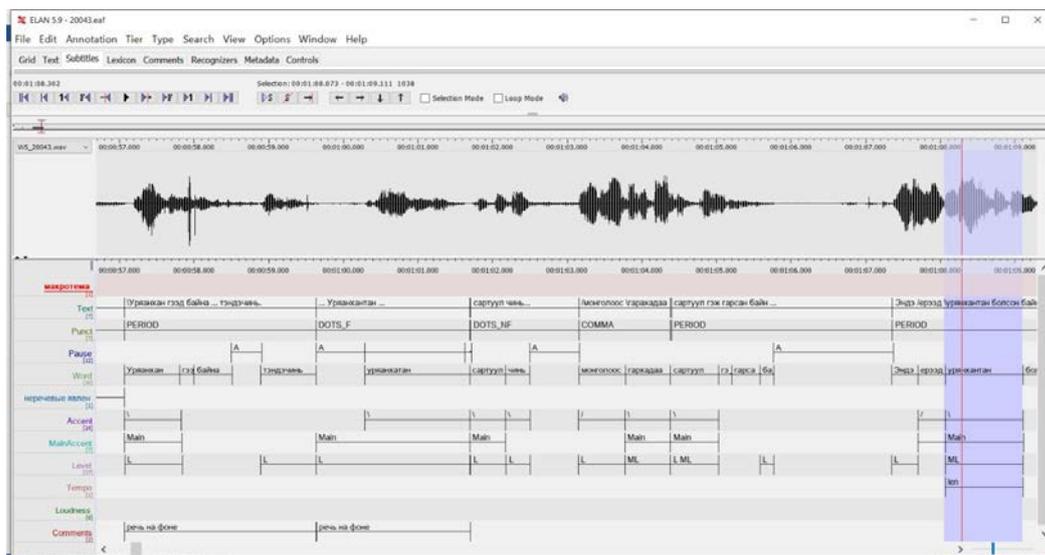


Рис. 1. Фрагмент аннотирования текста в программа ELAN

Таким образом, просодическая транскрипция является обязательным элементом аннотирования звуковых корпусов. Анализ звуковых корпусов русского языка, нацеленных на исследование как интонации, так и других аспектов живой разговорной речи, позволил выявить основные принципы просодического аннотирования и транскрибирования. В результате в бурятском языкознании впервые были разработаны принципы просодической транскрипции звучащего текста для звукового диалектного корпуса бурятского языка. Такая транскрипция позволит вести аннотирование диалектного звукового материала с целью анализа интонационных особенностей бурятских диалектов.

Литература

1. База данных «Интонация русского диалога»: вопросительные реплики / С. В. Кодзасов [и др.] // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции «Диалог 2005». М.: Изд-во РГГУ, 2005.
2. Брызгунова Е. А. Эмоционально-стилистические различия звучащей речи / Е. А. Брызгунова. М: Изд-во Моск. ун-та, 1984. 116 с.
3. Бускунбаева Л. А., Сиразитдинов З. А. Принципы транскрибирования аудиоматериалов в диалектном корпусе башкирского языка // Финно-угорский мир в полиэтничном пространстве России: культурное наследие и новые вызовы. Сб. ст. по материалам VI Всеросс. науч. конф. финно-угроведов. 2019. С. 545–548.
4. Вольская Н. Б., Скрелин П. А. Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Анализ разговорной русской речи: третий междисциплинарный семинар (АРЗ-2009) / сост. А. Л. Ронжин. СПб., 2009. С. 28–40.
5. Вольская Н. Б., Качковская Т. В. Принципы просодической разметки в новом корпусе русской спонтанной речи CoRuSS // Фонетика сегодня. Материалы докладов и сообщений VIII Междунар. науч. конф. М.; СПб., 2016. С. 29–31.

6. Оде С. Заметки о понятии тональный акцент на примере русского языка // Проблемы фонетики / под ред. Р. Ф. Касаткиной. М.: Наука, 2007. С. 237–249.
7. Просодический строй русской речи: монография / М. Л. Каленчук [и др.]. Институт русского языка РАН. М., 1996. 256 с.
8. Рассказы о сновидениях: Корпусное исследование устного русского дискурса / под ред. А. А. Кибрика и В. И. Подлесской. М.: Языки славянских культур, 2009. 736 с.
9. Русская грамматика / под ред. Н. Ю. Шведовой. М.: Наука, 1982. Т. 1. 784 с.
10. Русский интонационный корпус: предварительный отчет / А. В. Архипов [и др.] // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11(18): в 2 т. Т. 1. Основная программа конференции. М.: Изд-во РГГУ. Т. 1. С. 18–27.
11. Устные тексты на калмыцком языке: запись и расшифровка / В. В. Куканова [и др.] // Вестник Калмыцкого университета. 2013. № 3(19). С. 56–64.
12. Шерстинова Т. Ю., Степанова С. Б., Рыко А. И. Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Материалы XXXVIII Междунар. филол. конф. Секция: «Формальные методы анализа русской речи». Март. СПб.: СПбГУ, 2009. С. 66–75.

PRINCIPLES OF PROSODIC ANNOTATION IN THE SOUND CORPORA OF THE BURYAT DIALECTS

Iuliia D. Abaeva

candidate of science in Philology, scientific fellow
Institute for Mongolian, Buddhist and Tibetan Studies
of the Siberian Division of the Russian Academy of Science
6, Sakhyanovoy Str., Ulan-Ude, 670047 Russia
julaba@yandex.ru

Abstract. Prosodic transcription is one of the most important points in annotating of the sound corpuses. The analysis of existing corpora of Russian speech has shown that in sound corpuses aimed on the study of oral speech data, prosodic transcription is necessarily used. In some corpora of Russian speech intonation transcriptions that were created in the pre-corpus period are still explored, for example, the systems of E. A. Bryzgunova, S. V. Kodzasov still have not lost their relevance. Also new methods were created, for example, the system of discursive transcription of the corpus "Stories of Dreams" and others. The purpose of this article is to analyze the existing prosodic transcriptions, to reveal the basic principles of transcription. As a result, for the first time in Buryat linguistics a prosodic transcription to annotate the sound corpus of Buryat dialects has been developed. Spoken dialectal corpus is a necessary means of preserving dialectal features, also it is a mean of studying of the current state of Buryat dialects, and so it develops the Buryat dialectology.

Keywords: spoken corpora, dialectal corpora, intonation, prosodic transcription, prosodic annotation, phrase, intonation center, tonal accent, melodic, tempo, pause.

Статья поступила в редакцию 26.10.2020; одобрена после рецензирования 18.11.2020; принята к публикации 24.12.2020.