ОПЫТ ДИСКУРСИВНОЙ РАЗМЕТКИ КОРПУСА БЕСПИСЬМЕННОГО ЯЗЫКА КУЛЛУИ

Ю.В. Мазурова
Институт языкознания РАН
mazurova.julia@iling-ran.ru

Андрей Александрович Кибрик не был непосредственным руководителем моих научных исследований, однако я с начала своей научной деятельности работаю под его руководством в Институте языкознания — сначала в группе «Языки мира», потом в секторе ареальной лингвистики. Андрей Александрович никогда не учил меня чему-то напрямую (за исключением курса «Языки мира» на третьем, кажется, курсе), но из его книг, статей, выступлений, совместной работы над разными проектами и особенно из неформальных обсуждений всего и вся я усвоила важные для себя идеи, которые потом (иногда неожиданным образом) воплотились в моих собственных исследованиях. Я научилась у Андрея Александровича очень многому как в научном смысле, так и в человеческом, за что я ему безмерно благодарна.

В этой небольшой статье речь пойдёт о звуковом корпусе устных текстов на малом индоарийском языке куллуи, распространённом в штате Химачал-Прадеш, Индия. Корпус создан на основе полевых материалов, собранных сотрудниками Института языкознания РАН и Института востоковедения РАН Ю.В. Мазуровой, Е.А. Ренковской и А.С. Крыловой в ходе экспедиций в округ Куллу с 2014 по 2018 г. (Мазурова 2018) Подробная информация об исследовании, сам устный корпус и записанные тексты доступны на сайте проекта[1].

Значение работы «Рассказы о сновидениях: Корпусное исследование устного русского дискурса» под редакцией А.А. Кибрика и В.И. Подлесской выходит далеко за рамки частной задачи, с которой она начиналась. Как отмечают и сами авторы, «разработанные здесь принципы достаточно легко переносятся на материал других языков» (Кибрик, Подлесская 2009: 550) Фундаментальные выводы, к которым приходят авторы, оказываются чрезвычайно актуальными и для такой достаточно далеко отстоящей сферы как документация малых бесписьменных языков (Кибрик, Майсак 2021: 30–32).

[1] http://pahari-languages.ru. Дата обращения 20.03.2023

DOI 10.37892/978-5-6049527-2-6-7

Теоретические и методологические рамки, взятые за основу в этом исследовании (Кибрик, Подлесская 2009: 31), оказались очень полезны для решения задач документации языка. Это четыре основных тезиса: 1) признание приоритета и центральности устной формы языка; 2) ориентация на корпусный метод сбора и анализа данных; 3) необходимость дискурсивной транскрипции как инструмента репрезентации устных корпусных данных; 4) когнитивный теоретический подход. В этой статье мы на конкретных примерах разберём прикладное значение этих тезисов применительно к устному корпусу языка куллуи.

Что касается первого тезиса, то для куллуи устная форма является не приоритетной, а вообще единственной формой существования языка. Те немногие примеры записи куллуи и других малых языков химачали письмом деванагари в местных студенческих журналах, которые мы находили во время экспедиций в Химачал-Прадеш, а также переводы Библии на некоторые языки этой группы являются проявлениями очень периферийного языкового активизма, но никак не влияют на бытование языка в целом. Специальное письмо танкри (Ренковская, Крылова 2021) для языков химачали вышло из употребления ещё в середине XX в., но и тогда оно использовалось весьма ограниченно.

Из этого следует и очевидность второго тезиса: полноценно описывать грамматику куллуи на современном уровне можно только на основе собранного и размеченного корпуса устных текстов и анкет, как отмечено в (Renkovskaya et al. 2020: 155). Наш корпус состоит из двух частей — естественных текстов и элицитированных грамматических анкет, необходимых для изучения отдельных грамматических явлений. В корпусе имеется техническая возможность при поиске выбрать один из этих подкорпусов. В этой статье речь будет идти только о подкорпусе естественных текстов — это записанные в ходе экспедиций рассказы носителей языка о жизни, о своей деревне, традициях, праздниках и местных преданиях.

Главный рассматриваемый в этой статье тезис — необходимость дискурсивной разметки в корпусе. На самом деле это не так очевидно, и мы далеко не сразу пришли к пониманию такой необходимости.

Поскольку мы первые начали изучать куллуи на современном научном уровне, перед нами стояли базовые задачи — изучение фонетики и фонологии, разработка транскрипции, сбор словаря, описание основных морфологических и синтаксических явлений. Однако невозможно изучить более низкие уровни языка без обращения к более высоким. Невозможно построить фонологию без обращения к значению морфем. Точно так же нельзя разобраться в грамматических явлениях, не обращаясь в тех или иных случаях к суперсегментной и дискурсивной структуре текста в целом. В конце концов стала очевидна необходимость какой-то систематической разметки этих явлений в корпусе.

Если для транскрипции и морфологической разметки в отечественных корпусных проектах по документации уже сложился некий общепринятый узус правил глоссирования, поиска и т. п., то для разметки других языковых уровней такого пока не наблюдается[2]. При создании корпуса устных текстов возникают определённые практические вопросы: разбиение на клаузы, внутреннее членение предложения на группы, обозначение интонации, отношения говорящего к высказыванию. Все эти вопросы тесно связаны с сегментацией, а также с просодической и суперсегментной организацией текста.

В отечественных корпусных исследованиях пока не сложилось общепринятых решений по поводу суперсегментной разметки в корпусах: в каждом проекте выбирается то, что необходимо для конкретных исследовательских задач. Не претендуя на исчерпывающее исследование этого вопроса, перечислю основные возможности, представленные в наиболее известных и разработанных устных корпусах, представленных на сайтах «Международная лаборатория языковой конвергенции НИУ ВШЭ»[3], «Малые языки России»[4], «Устный корпус осетинского языка»[5]. Для языков с письменной традицией или диалектов таких языков в устных текстах ожидаемо используется пунктуация (осетинский, бесермянский, хакасский, башкирский, луговой марийский, муиринский даргинский, кадарский даргинский, эвенкийский). Для некоторых языков используются только точки, а сегментация на более мелкие единицы идёт просто по строчкам (кетский, ительменский). В корпусе лугового марийского есть элемент, обозначающий речевой сбой (=), но он используется окказионально, наряду с запятой и многоточием в том же смысле.

В корпусе куллуи мы приняли решение использовать не пунктуацию, а адаптированные элементы дискурсивной разметки, подробно разработанные в (Кибрик, Подлесская 2009). Куллуи — язык бесписьменный, поэтому никакой пунктуационной нормы там, разумеется, нет, а привносить в куллуи пунктуацию из родного нам русского языка — возможное, но не самое удачное решение, так как в русском языке знаки препинания, особенно запятая, слишком многозначны. Из знаков пунктуации мы используем только точку, вопросительный и восклицательный знаки.

---

[2] В статье речь идёт только об устных корпусах малых языков: разумеется, есть корпуса, которые создаются именно для изучения дискурсивных и других явлений, и в них используется специально разработанная для этих целей разметка.

[3] https://ilcl.hse.ru/corpora, http://lingconlab.ru/resources_ru.html. Дата обращения 20.03.2023.

[4] https://minlang.iling-ran.ru/corpora. Дата обращения 20.03.2023.

[5] https://www.ossetic-studies.org/ru/texts. Дата обращения 20.03.2023.

Для текущих исследовательских задач оказались важны следующие типы суперсегментной разметки: 1) паузы, 2) интонация, 3) речевые сбои, 4) переключение на другой язык, 5) смех, 6) неясные фрагменты.

Дискурсивная разметка в корпусе куллуи

|||	Пауза (длина паузы соответствует количеству палочек — от одной до трёх)
/	Восходящая интонация
\	Нисходящая интонация
?	Вопросительная интонация
!	Восклицательная интонация
=	Обрыв слова или фразы
[]	Переключение кода
⟨laugh⟩	Смех
⟨…⟩	Неразборчивый фрагмент

В примере (1) представлены следующие виды дискурсивной разметки: пауза, восходящая интонация и обрыв слова:

(1) *dzebe ase skul-a-b dza-a tʰi na* |
когда 1Pl.Dir школа-Obl-Acc/Dat идти-Ger\Ipfv Cop.Pst Disc
bəɣ-a / mədza la = ɐtʃʰa lag-a tʰi dzɛɳɖ-a .
очень-M удовольствие *** хорошо казаться-Ger\Ipfv Cop.Pst такой-M
‘Когда мы ходили в школу, нам очень нравилось’.

Дискурсивный показатель *na* ‘не так ли’ омонимичен показателю отрицания, однако пауза после *na* снимает синтаксическую неоднозначность, показывая, что интонационно *na* примыкает к первой синтагме. Знак = показывает, что элемент «*la*» является не отдельным словом, а исправлением конструкции: вместо «*mədza laga*» (такого выражения не существует) говорящий выбрал простое и стандартное выражение «*ɐtʃʰa laga*». Восходящая же интонация после *bəɣa* вероятнее всего как раз отражает сомнения информанта в выборе дальнейшей конструкции.

В данном случае разметка позволяет нам правильно отглоссировать и перевести пример — без обращения к суперсегментному уровню этого нельзя сделать. Разумеется, при обработке текстов мы сами переслушиваем аудиозапись и принимаем во внимание паузы и интонацию, однако эксплицитная разметка позволяет потом пользоваться и перепроверять те или иные решения, а не полагаться на то, что исследователь услышал и понял в конкретный момент времени, опираясь на свою интуицию.

Довольно частым случаем, где встречаются речевые сбои, является выбор падежа местоимения. В куллуи есть два основных случая, когда говорящему нужно выбрать либо прямой, либо косвенный падеж. Во-первых, это дифференцированное оформление прямого объекта в зависимости от его референциального статуса. Второй случай — это выбор между номинативной и эргативной конструкцией. В примере (2) речевой сбой происходит, когда говорящему нужно выбрать падеж одного и того же местоимения ‘он’ сначала в номинативной (с непереходным глаголом ‘уходить’), а потом в эргативной (с переходным глаголом ‘класть’) конструкции. Носительница языка несколько раз повторяет местоимения в разных падежах, меняя свой изначальный выбор конструкции:

(2) *tebe teie bol-u ki ʁtsʰa haũ a-u*
тогда 3Sg.M.Dist.Erg говорить-Pfv.Sg.M что хорошо 1Sg приходить-Pfv.Sg.M
tə teie = sɔ =
и 3Sg.M.Dist.Erg 3Sg.Dist.Dir
sɔ *nɔʈʰ-a tɔkʰe-nə*
3Sg.Dist.Dir идти-Pfv.Sg.M там-ABL
tə ***teie*** *newle age ɖah-u*
и 3Sg.M.Dist.Erg мангуст-OBL APUD класть-PFV.SG.M
apɳa beʈ-a her-ɳ-e-r-i tẽje.
свой-M сын-Dir.M смотреть-Inf-Obl-Gen-F для
‘Он сказал: «Хорошо, я пойду», он ушел, он положил сына рядом с мангустом, чтобы тот присматривал’.

Записанный без дискурсивных помет текст может оказаться двумя разными клаузами или содержать повтор, ошибку и самоисправление. Бывает, что носитель начинает фразу, имея в виду одно, а потом меняет своё речевое намерение. Если для грамматики учитывать только сегментный уровень, то там встретится много неясных примеров, противоречащих друг другу. При этом надо помнить, что на данный момент собранный корпус — это всё, что у нас есть, нет никакого «нормированного» языка или «квалифицированного» носителя, к которому мы могли бы обратиться за подтверждением или уточнением правильности примера. Более того, языками-посредниками (английским и хинди) носители куллуи зачастую владеют только на разговорном уровне, а концепция точного перевода с одного языка на другой им совсем не так очевидна, как хотелось бы нам, исследователям. Поэтому полностью полагаться на перевод, данный во время расшифровки, нельзя: он может быть неточным, и его нужно перепроверять во время обработки. Таким образом, анализ грамматики неизбежно влечёт анализ дискурсивной структуры текста.

Приходится делать попытки заглянуть в голову носителю и понять, что он имел в виду, когда произносил фразу. Это и есть практически ориентированный когнитивный анализ дискурса.

Исследование дискурсивных явлений в корпусе куллуи на данный момент — это вспомогательная задача для описания грамматики этого языка. Глубокий анализ и описание дискурса в куллуи — дело будущего. Однако, как показывает наш опыт, некоторые предварительные дискурсивные исследования и систематическая разметка корпуса совершенно необходимы для адекватного описания более базовых уровней — морфологии и синтаксиса.

Литература

Кибрик А.А., Майсак Т.А. 2021. Правила дискурсивной транскрипции для описательных и документационных исследований. *Rhema. Рема*, № 2, 23–45.

Кибрик А.А., Подлесская В.И. (ред.) 2009. *Рассказы о сновидениях: корпусное исследование устного русского дискурса*. М.: Языки славянских культур.

Мазурова Ю.В. 2018. Малые индоарийские языки Северной Индии: язык куллуи. *Вестник РФФИ. Гуманитарные и общественные науки*, № 4, 82–91.

Ренковская Е.А., Крылова А.С. 2021. Манускрипты на танкри из коллекции Кхубрама Кхушдиля (Химачал Прадеш, Индия): первые результаты работы с текстами. В кн.: *Российские исследования Гималаев и Тибета 2021 — природа и культура* (материалы конференции, Санкт-Петербург, 23-24 ноября 2021 года). СПб: Европейский дом, 40–41.

Renkovskaya E., Krylova A., Mazurova J. 2020. Documentation of Himachali Pahari languages as a step towards language maintenance. In: B.K. Joshi, P. Pokharel Madhav, P. Joshi Maheshwar (eds.). *Language Endangerment and Language Revitalization in Himalaya* (Proceedings of the International Seminar on Endangered Languages of Himalaya, Almora 2018). Dehradun: Doon Library and Research Centre, 149–156.