Электронные базы метаданных и построение диалектологического атласа

Electronic Databases of Metadata in the Development of the Dialectological Atlas

А. В. Шеймович

A. V. Sheimovich

В статье обсуждается необходимость создания электронной базы метаданных, собранных при лингвистическом анкетировании информантов. Анкетирование проводилось на протяжении примерно 10 лет в процессе сбора диалектного материала для электронных корпусов и атласа тюркских языков Южной Сибири.

Ключевые слова: полевая лингвистика, диалектология, метаданные, базы данных

The article discusses the need to create an electronic database of metadata collected during the linguistic survey of informants. The survey was conducted for about 10 years in the process of collecting dialect material for electronic corpora and the atlas of the Turkic languages of Southern Siberia.

Keywords: field linguistics, dialectology, metadata, databases
DOI: 10.37892/2313-5816-2022-2-185-197

Метаданными[1] в полевой лингвистике обычно называют вспомогательные экстралингвистические данные, при-

[1] Метаданные (от др.-греч. *μετά* «за, после, рядом, в середине» и данные) — информация о другой информации или данные, относящиеся к дополнительной информации о содержимом или объекте. Метаданные раскрывают сведения о признаках и свойствах, характеризующих какие-либо сущности, позволяющие автоматически искать и управлять ими в больших информационных потоках.

лагающиеся к лингвистической информации, собранной в полевых условиях, и помогающие более адекватной ее интерпретации.

Существует множество публикаций, посвященных различным проблемам полевой лингвистики, включая методику сбора социолингвистической информации и других метаданных, предшествующих сбору собственно языкового материала [Архипов 2008: 6–9].

В зависимости от целей, с которыми собирается языковой материал, различные параметры метаданных могут приобретать большую или меньшую актуальность.

Ниже идет речь о систематизации и хранении метаданных, собранных экспедиционным коллективом проекта «Создание электронного диалектологического атласа тюркских языков России» (рук. А. В. Дыбо) в процессе картирования тюркских языков и диалектов Южной Сибири. Сибирь является одним из многих языковых ареалов, входящих в сферу действия этого проекта, длящегося уже более 10 лет. В процессе экспедиционной работы происходит сбор (и постоянный добор недостающего) языкового материала, что должно позволять строить и редактировать электронные лингвистические карты, отражающие значения языковых признаков, наиболее ярко характеризующих диалектные различия в области фонетики, фонологии, морфонологии, морфологии и лексики тюркских языков; подробнее об этом см. [Дыбо и др. 2020].

Экспедиции по этому проекту несколько отличаются от экспедиций типологов классического образца (за который можно принять экспедиции ОТиПл МГУ, в частности дагестанские экспедиции А. Е. Кибрика 1967–1998 гг. [Кибрик 1992; Кибрик 2007]): вместо достаточно подробного и всестороннего обследования идиомов одного или нескольких сёл примерно со второго полевого сезона работы над Атласом мы перешли к т. н. методике «коврового опроса» территорий, запланированных для нанесения на карты. В качестве базы экспедиции выбиралось село, из которого были транспортно достижимы основные на-

селенные пункты района, где жили носители интересующего нас идиома. Жителей этих пунктов опрашивали по стандартным наборам анкет на определенные диагностические признаки (см. [Дыбо и др. 2020]), по которым планировалось построить классификацию говоров этого ареала. Затем база экспедиции перемещалась в соседний район, откуда объезжался следующий «куст» деревень.

При таком способе работы по сравнению с классическим вариантом сильно возрастает количество опрошенных информантов и мест их проживания, и, соответственно, метаданных. Меняется структура и самих метаданных, значительно возрастает количество выделяемых параметров. Так, становится важным фиксировать ряд дополнительных метаданных, как, например, географические координаты населенных пунктов[2].

Ср. ниже таблицы 1 и 2.

Таблица 1 — пример таблицы метаданных, собранных во время экспедиции 2015 г. в Ширинском р-не Хакасии. Работа велась в сёлах Трошкин (преимущественно), Белый Балахчин, Черное Озеро, Шира. В этот период собранный материал еще не использовался непосредственно для нужд диалектологического Атласа; пополнялись базы электронных корпусов миноритарных языков и т. н. «анкетного проекта» — «Разработка анкет для сбора материалов к интегральному описанию миноритарных тюркских языков и диалектов России» (рук. А. В. Дыбо).

Таблица 1 очень компактна: значительная часть данных явно осталась в блокнотах у сборщиков вместе с т. н. «первичными метаданными». Первичными метаданными я называю информацию, полученную зачастую еще до встречи с информантом, например у главы поселковой администрации, у библиотекаря или директора школы

[2] Помимо того что координаты необходимы для последующей работы над картой, они помогают различать одноименные населенные пункты в разных районах и дают представление о местонахождении сёл, к настоящему времени уже исчезнувших, где информант мог родиться.

(ФИО, адрес, примерный возраст; в последние годы это иногда мог быть даже номер мобильного телефона), у родственника или у встреченного в магазине односельчанина («зайдите к бабе Наде, второй дом от угла, зеленый забор, хорошо язык помнит»; «спросите на соседней улице учительницу истории Нину Васильевну, она всё вам расскажет»). С метаданными работали постфактум, т. к. их записывали в отдельный аудиофайл на языке информанта, который требовал отдельной расшифровки, чаще всего с помощью носителя языка; в таблице 1 предусмотрено больше позиций для данных о собранном материале и о сборщиках и меньше — об информантах.

Таблица 1

Экспедиция в Ширинский р-н Хакасии, август 2015 г.

№	Язык, диалект, селение	Назв. файлов	Содержание (предмет опроса)	Информант (ФИО, год. рожд., место рожд.)	Экспедиция (организация, число, месяц, год)	Опросчик	Расшифровка (назв. файла, кто расшифровывал)
1.	Хакасский, качинский, Трошкино	ZojaEfimovna_razgovor.flac	Разговор	Аёшина (Кокова) Зоя (Соня) Ефимовна, 1927, родилась в с. Фыркал	ИЯз 2–14 августа 2015 г.	Э. В. Султрекова	Зоя Ефимовна разговор.docx, Э. В. Султрекова

Разница по времени между созданием таблиц 1 и 2 — четыре года.

Адаптируясь к нуждам проекта диалектологического атласа, структура метаданных в табл. 2 значительно усложнилась. Таблица содержит значительно большее количество куда более подробно структурированной информации[3], перекочевавшей в нее из блокнотов, текс-

[3] Вышесказанное не значит, что в 2015 и 2019 гг. лингвисты собирали с информантов разные наборы метаданных. Но акценты при опросах смещались с одних факторов на дру-

товых файлов, электронных писем. В экспедициях последних лет более последовательно стало соблюдаться правило: в конце рабочего дня каждый участник опросов должен внести в общую таблицу биографические и иные данные о своих информантах и детали опроса. Это положительно повлияло на объем и полноту собранных метаданных и плачевно сказалось на удобстве работы с таблицей в прежнем формате.

Очевидно, что большие, тяжелые массивы метаданных, представленные в виде таблиц в текстовых документах, становятся крайне неудобны для обработки, для поиска внутри них.

Внутри метаданных можно выделить следующие группы:

1. Данные, относящиеся к информанту, биографические и социолингвистические: имя, год и место рождения, места проживания, образование, род деятельности (позиции 2, 4–10, 12);
2. Данные, относящиеся к географическому ареалу, который занимает исследуемый диалект, «геоданные»: места рождения и места проживания информантов, места опроса, их географические координаты (позиции 5–10) [группы 1 и 2 частично пересекаются, но если в первой группе важна ситуация отдельного носителя языка, то во второй – ситуация всей группы носителей]; к этому же типу данных относится информация о диалектной принадлежности информанта (позиция 3).
3. Данные, относящиеся к собранной языковой информации: какие анкеты были опрошены, их названия; сколько и каких файлов записано (аудио-, видео-, фото).

гие. Например, при анализе материала для составления карты диалектного ареала данные о месте опроса имеют меньший вес по сравнению с данными о том, где сформировались языковые навыки информанта (место рождения, место обучения в начальной и средней школе, место проживания до 15–18 лет).

Таблица 2

Экспедиция в Орджникидзевский р-н Хакасии, июль-август 2019 г.

№	Имя информанта	Язык/диалект/говор	Год рожд.	Место рождения и координаты			Место опроса и координаты			Дата опроса	Краткие биогр. сведения, образование, передвижения (цель — выяснить, где человек выучился говорить, с кем и на каком языке разговаривал)	Что спрашивали	Какие файлы записаны, их кол-во и как были обработаны	Имя сборщика	Примечания (на что писали, в каких усл.)
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1.	С.Н.И.[4]	кызыльский	1958	Черное Озеро	54. 688269	89. 429505	Подкамень	54. 820664	89. 449376	1.08.19	В Подкамне с 1969, нач. школа в Ошколе, 7-й кл. в Новомарьясово (интернат), 8-й класс в Устинкино, служба в армии в Москве	анкета на изоглоссы, кр. однослоги; *к-г-х*; полные однослоги (до половины спрошено подряд, потом с упором на *и/е*)	кр. однослоги – нарезка, изоглоссы нарезаны и внсены в базу	К.Е.В. ИЯз РАН	Zoom Pro

[4] В оригинальной таблице содеоржатся полные ФИО информантов, которые дали на это согласие; здесь приведены инициалы в целях экономии места и анонимности.

3.1 Данные, относящиеся к условиям работы: дата[5] и место опроса (и в целом район проведения экспедиции); на что велась запись (тип диктофона; смартфон) и в каких условиях (если они значительно влияют на качество записи) (позиции 8–11, 16);
3.2 Данные, относящиеся к последующей обработке собранной информации (расшифрована, обработана соответствующей программой, внесена в базу данных, выложена в интернет и т. п.) (позиция 14); местонахождение архива файлов-первоисточников (в таблице 2 не реализовано);
4. Данные, относящиеся к людям, собиравшим информацию: ФИО, место работы (позиция 14).

Для оптимального представления лингвистического признака на карте необходимо увязать между собой по крайней мере три первых типа метаданных.

Диалектная принадлежность информанта (позиция 3 «Язык/диалект/говор») заполняется в последнюю очередь и является результатом анализа полученных лингвистических материалов в сопоставлении с персональными данными информантов и географической информацией. Для этой работы важно зафиксировать «в связке» ряд параметров, относящихся как к географическому положению населенного пункта, где живут носители идиома, так и к «языковой биографии» опрашиваемых носителей. Для более точного определения диалектной принадлежности информанта требуется возможно более подробное структурирование биографических и географических сведений, т. е. (в перспективе) разбиение содержимого столбца 12 на несколько позиций (смена мест жительства по годам, образование по годам и по населенным пунктам, смена мест

[5] На результаты работы влияет не только точная дата опроса (позиция 11), но и в целом временной промежуток полевых работ (в табл. 2 он вынесен в заголовок), т. к. важно знать, не пришелся ли он на сезон сенокоса, например, или сбора кедрового ореха в тайге.

работы по годам и по населенным пунктам). Часто оказывается, что, хотя человек довольно долго живет в населенном пункте, где проводится опрос, говорит он все же на диалекте того места, где родился, — если прожил там достаточно долго, общался на родном языке с представителями старших поколений, ходил в начальную школу и пр. Так часто бывает с женщинами, вышедшими замуж в другой район и сохраняющими при этом диалектные черты своего родного села. В случае, когда за годы, проведенные на новом месте, родной диалект не был вытеснен местным, обычно приходится относить диалект такого информанта не к месту сбора данных, а к месту, где проходило его детство и становление языковых навыков (часто совпадающее с местом рождения). Поэтому зафиксировать это место часто бывает важнее, чем географическую локацию опроса/место проживания в настоящий момент. Хотя случаи сильного влияния диалекта места проживания на родной диалект у информантов тоже встречаются.

Метаданные, касающиеся собранных и обработанных материалов, становятся важны, когда из большого массива звуковых и текстовых файлов необходимо быстро извлечь несколько таких, по которым планируется строить карту на конкретную изоглоссу. Здесь бы сильно помогла сортировка по одному или нескольким параметрам метаданных: название анкеты (в перспективе — разбить столбец 13 на позиции по числу опрашиваемых в текущем сезоне анкет, дав им короткие условные обозначения), название/координаты населенного пункта, название идиома/диалекта. Это облегчило бы получение названий файлов, содержащих, к примеру, все стословники для шорского диалекта хакасского языка для поселка Анчуль.

Структурированную таким образом информацию уже невозможно хранить в виде таблицы в текстовом документе. Становится очевидной потребность в интерактивной электронной базе метаданных, создание которой, к сожалению, пока только планируется. Объединенная с картографической базой, база метаданных может стать одним

из компонентов системы связанных баз (электронных словарей, грамматик и корпусов текстов языков и диалектов), на основе обращений к которой и должна строиться работа Электронного диалектологического атласа тюркских языков.

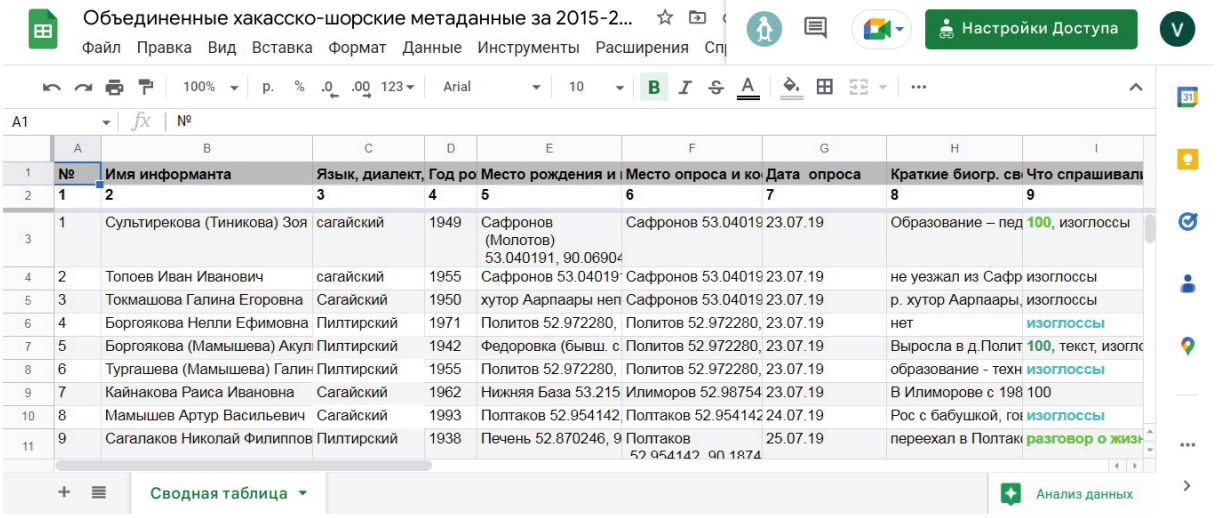
Сырой образец подобной базы, в которой в отдельные позиции вынесены далеко не все описанные выше параметры метаданных, представлен ниже (см. таблица 3).

Планируется построить электронную базу метаданных[6], собранных в 2006–2022 гг. (и далее) в экспедициях под руководством А. В. Дыбо по языкам и диалектам Южной Сибири (Хакасия, Шория, Алтай, Тува), и последовательно расширять ее на другие тюркоязычные регионы.

[6] Предварительно планируется делать это средствами СУБД Starling [см. Крылов, Тер-Аванесова].

Таблица 3

Объединенные метаданные по экспедициям в Хакасию — Шорию за 2015–2022 гг.



	A	B	C	D	E	F	G	H	I
1	№	Имя информанта	Язык, диалект,	Год ро	Место рождения и	Место опроса и ко	Дата опроса	Краткие биогр. св	Что спрашивали
2	1	2	3	4	5	6	7	8	9
3	1	Сультирекова (Тиникова) Зоя	сагайский	1949	Сафронов (Молотов) 53.040191, 90.06904	Сафронов 53.04019	23.07.19	Образование – пед	100, изоглоссы
4	2	Топоев Иван Иванович	сагайский	1955	Сафронов 53.04019	Сафронов 53.04019	23.07.19	не уезжал из Сафр	изоглоссы
5	3	Токмашова Галина Егоровна	Сагайский	1950	хутор Аарпаары неп	Сафронов 53.04019	23.07.19	р. хутор Аарпаары,	изоглоссы
6	4	Боргоякова Нелли Ефимовна	Пилтирский	1971	Политов 52.972280,	Политов 52.972280,	23.07.19	нет	изоглоссы
7	5	Боргоякова (Мамышева) Акул	Пилтирский	1942	Федоровка (бывш. с	Политов 52.972280,	23.07.19	Выросла в д.Полит	100, текст, изогло
8	6	Тургашева (Мамышева) Галин	Пилтирский	1955	Политов 52.972280,	Политов 52.972280,	23.07.19	образование - техн	изоглоссы
9	7	Кайнакова Раиса Ивановна	Сагайский	1962	Нижняя База 53.215	Илиморов 52.98754	23.07.19	В Илиморове с 198	100
10	8	Мамышев Артур Васильевич	Сагайский	1993	Полтаков 52.954142	Полтаков 52.954142	24.07.19	Рос с бабушкой, го	изоглоссы
11	9	Сагалаков Николай Филиппов	Пилтирский	1938	Печень 52.870246, 9	Полтаков 52.954142, 90.1874	25.07.19	переехал в Полтак	разговор о жизн

Литература

Архипов А. В. Документирование малых языков: научные и технические аспекты // *Языковое разнообразие в киберпространстве: российский и зарубежный опыт*. Москва, 2008, 76–83.

Дыбо А. В., Мальцева В. С., Николаев С. Л., Шеймович А. В. Диалектологическая анкета для пилотного опроса «Признаки-изоглоссы для хакасско-шорско-чулымского ареала (группы тюркских z-языков)» // *Родной язык*, 2020, 1(12): 86–119.

Кибрик А. Е. Методика полевой работы с информантом // *Очерки по общим и прикладным вопросам языкознания (Универсальное, типовое и специфичное в языке).* Москва, 1992, 262–287.

Кибрик А. Е. Диалог лингвиста с носителем: в поисках полевого метода и формата лингвистического описания // *«На меже меж Голосом и Эхом».* Сборник статей в честь Татьяны Владимировны Цивьян. Сост. Л. О. Зайонц. Москва, 2007.

Крылов С. А., Тер-Аванесова А. В. *Электронные базы данных по русским народным говорам*.
URL: http://niryaz.inion.ru/linguistics/125

Kazakevich O. Fieldwork in the Situation of Language Shift // *Strategies for Knowledge Elicitation: The Experience of the Russian School of Field Linguistics*. Springer Nature Switzerland AG. 2021, 103–118.

Crowley T. *Field Linguistics. A Beginner's Guide*. Edited and prepared for publication by Nick Thieberger. Oxford University Press, 2007.

Chelliah Sh. L., De Reuse W. J. *Handbook of descriprive linguistic fieldwork*. Springer, 2011.

References

Arkhipov A. V. Dokumentirovanie malykh yazykov: nauchnye i tekhnicheskie aspekty [Documentation of minority languages: scientific and technical aspects] // *Yazykovoe raznoobrazie v kiberprostranstve: rossiyskiy i zarubezhnyy opyt*. Moskva, 2008, 76–83. (In Russ.)

Chelliah Sh. L., De Reuse W. J. *Handbook of descriprive linguistic fieldwork*. Springer, 2011.

Crowley T. *Field Linguistics. A Beginner's Guide*. Edited and prepared for publication by Nick Thieberger. Oxford University Press, 2007.

Dybo A. V., Mal'tseva V. S., Nikolaev S. L., Sheymovich A. V. Dialektologicheskaya anketa dlya pilotnogo oprosa «Priznaki izoglossy dlya khakassko-shorsko-chulymskogo areala (gruppy tyurkskikh z-yazykov)» [Dialectological questionnaire for the pilot survey "Isogloss signs for the Khakass-Shor-Chulym area (a group of Turkic z-languages)"] // *Rodnoy yazyk*, 2020, 1(12): 86–119. (In Russ.)

Kazakevich O. Fieldwork in the Situation of Language Shift // *Strategies for Knowledge Elicitation: The Experience of the Russian School of Field Linguistics*. Springer Nature Switzerland AG. 2021, 103–118.

Kibrik A. E. Dialog lingvista s nositelem: v poiskakh polevogo metoda i formata lingvisticheskogo opisaniya [Dialogue of a linguist with a native speaker: in search of a field method and format of a linguistic description] // *«Na mezhe mezh Golosom i Ekhom»*. Sbornik statey v chest' Tat'yany Vladimirovny Tsiv'yan. Sost. L. O. Zayonts. Moskva, 2007. (In Russ.)

Kibrik A. E. Metodika polevoy raboty s informantom [Methods of field work with an informant] // *Ocherki po obshchim i prikladnym voprosam yazykoznaniya (Universal'noe, tipovoe i spetsifichnoe v yazyke)*. Moskva, 1992, 262–287. (In Russ.)

Krylov S. A., Ter-Avanesova A. V. *Elektronnye bazy dannykh po russkim narodnym govoram* [Electronic databases of Russian folk dialects]. URL: http://niryaz.inion.ru/linguistics/125 (In Russ.)

Шеймович Александра Валерьевна
Институт языкознания РАН
Москва, Россия
Sheimovich Aleksandra Valerievna
Institute of Linguistics RAS
Moscow, Russia
asheimovich@yandex.ru