DOI: 10.37892/2218-1393-2023-18-1-75-87

Описание языковых явлений: факты и артефакты (на материале веб-версии базы данных «Языки мира» ИЯз РАН)

А. К. Зотова, О. И. Романова (Институт языкознания РАН)

Аннотация

В статье рассматриваются возможности представления данных о специфических языковых явлениях и анализируются причины возможных неточностей при переносе данных о языках из описаний в энциклопедическом издании «Языки мира» в формат веб-версии базы данных «Языки мира». Материал может представлять интерес для специалистов в области лингвистической типологии, формализации данных и разработки лингвистических баз данных.

Ключевые слова

Языки мира, формализованное представление данных, лингвотипологические базы данных.

База данных «Языки мира» создается на основе томов лингвистической энциклопедии «Языки мира», издаваемых Институтом языкознания РАН начиная с 1993 г. В рамках этого издательского проекта опубликованы 23 книги. Работа над созданием базы данных «Языки мира» началась с середины 1980-х гг. Об истории работы над базой данных см. [Коломацкий 2023], там же см. библиографию. За годы своего существования база данных претерпела несколько этапов усовершенствования, последние из которых связаны с адаптацией к возможностям Интернета (веб-версия).

База данных «Языки мира» является лингвотипологической, и одной из основных задач работы по ее совершенствованию и расширению является обеспечение системности и сопоставимости с международными лингвистическими стандартами. Самостоятельной и существенной задачей становится перенос данных о языках из описаний в энциклопе-

дии в базу данных в виде формализованных описаний (рефератов) с минимальными потерями и искажениями лингвистической информации.

С этой целью была оптимизирована процедура подготовки рефератов в формате *Microsoft Excel*, которая позволяет оперативно обновлять элементы анкеты — списки признаков и значений признаков, а также фиксировать лакуны в описаниях языков и классифицировать причины их появления.

Многократное поэтапное редактирование рефератов заставило более пристально присмотреться к еще одному явлению — появлению случаев *переописания* языков, то есть возникновению артефактов и механизмам их образования. В данном случае под артефактом понимается «недостоверный результат научного исследования, возникающий из-за дефектов метода исследования или ошибок экспериментатора» [БРЭ 2005, 283]. Речь идет об искажении и привнесении данных, которые появляются в результате их переноса из энциклопедии в заданную структуру формализованного реферата.

Данные о языке в том виде, в каком они представлены в энциклопедии, не всегда могут быть сведены к значениям признаков, сформулированным в списке допустимых значений, о типах значений в обновленном варианте реферата см. [Зотова и др. 2022]. Для таких случаев в списке значений предусмотрена возможность сформулировать значение признака *ad hoc*, которое фиксируется в соответствующем столбце типа значений *Custom*. Предполагается, что такие значения постепенно, по мере накопления заполненных рефератов, будут интегрированы в список допустимых значений в типе *Listed*.

В рефератах многих языков есть уникальные значения *Custom*, которые по формату значительно отличаются от допустимых вариантов. Так, в описаниях ряда иранских, дардских и других слабоизученных и фрагментарно описанных языков фонологические позиции представлены простым перечнем фонем или звукотипов без каких-либо характеристик или уточнений. Учитывая сложную соотносимость символов латинского или кириллического алфавитов с акустическим обликом фонем, их адекватная классификация по признакам реферата чревата ошибками. Чтобы избежать риска появления артефактов,

было принято решение создать типизированные варианты (прототипы) представления данных, например, фонологических — о вокализме в позициях A-1 — A-3 и консонантизме в позициях A-10 — A-21. Принцип состоит в том, чтобы в начале каждого нового блока реферата (фонология, морфонология, морфология и т.д.) описать нужный признак с помощью значения *Custom* (при необходимости в сочетании с уточнением в поле «Комментарий»):

— A-1: Количество степеней подъема *Custom*: Список из *N* гласных фонем по количеству степеней подъема терминологически не интерпретируется;

— A-3: Ряды гласных *Custom*: Список из *N* гласных фонем по признаку ряда терминологически не интерпретируется.

В случаях, когда характеристика имеет количественное выражение, например, «A-4: Фонологические ступени долготы — A-4-2: Две», прототипическим становится выбор из списка допустимых значений *Listed* с уточнением в поле «Комментарий»: «Фонологический статус противопоставления гласных по долготе неясен» (бóльшая часть примеров в статье основана на материале дардских языков, в данном случае приведен пример из языка тирахи) [Языки мира 1999a]. При отсутствии количественных данных используется тип значения *Custom*: «Есть долгие и краткие гласные» + «Комментарий»: «Фонологический статус долгих гласных неясен» (калаша, пашаи).

Список допустимых значений в идеальном случае должен содержать значение «Явление *X* отсутствует». В случае, когда такого значения нет, возникает необходимость ввести его с помощью типа значения *Custom*, иногда дополняемого комментарием. Например, в позиции «H-7 Маркирование зависимого имени в генитивной конструкции» в случае отсутствия в языке генитива необходимо с помощью *Custom* ввести значение «Генитив отсутствует», т.к. в списке допустимых значений присутствуют только значения «маркированное» и «немаркированное», что имплицитно свидетельствует о наличии генитива как такового (кашмири, шина, кховар).

Еще один тип неоднозначных контекстов представлен ситуациями отсутствия в языке описываемого грамматического явления, что делает невозможным выбор ни одного из допустимых значений. Например, позиция реферата «К-10: Грамматические категории, выражаемые наречием» предполагает выбор из значений: «К-10-1: Отсутствуют» и «К-10-2: Степень сравнения». В языках, где наречие отсутствует как часть речи, выбор первого из этих значений чревато появлением артефакта и искажением реальной ситуации. Решением становится введение значения «Наречие отсутствует» через тип значения *Custom* (сави).

Позиция К-10 является единственной в реферате, где может быть описано наречие. В случае, когда в статье оговаривается, что данные о наречии неполны, выбор типа значения *Not stated* ведет к появлению привнесенной информации об отсутствии наречия в принципе. В этом случае решением является выбор типа значения *Explicit gap* в сочетании с комментарием: «О наречиях имеются лишь отрывочные данные» (пхалура) либо типа значения *Custom*: «Наречие отсутствует» с комментарием: «Функции наречия выполняют прилагательные» (майян).

Существуют ситуации, при которых выбранное допустимое значение не описывает явление полностью и без комментария приобретает статус артефакта. Например, в языке кашмири в позиции «К-13: Атрибутивное согласование прилагательных» допустимое значение «Всегда присутствует» в целом адекватно характеризует реальное положение вещей, потому что изменяемые прилагательные в атрибутивной синтагме согласуются с определяемым по роду, числу и падежу. Однако в этом языке есть группа неизменяемых прилагательных, и существование такой группы нужно оговорить в комментарии: «Есть изменяемые и неизменяемые прилагательные».

Причина неопределенности при выборе типа значения важна для анализа и адекватного представления данных в реферате. Добавлять значения в список *Custom* и создавать комментарий необходимо в случаях, когда текст описания не содержит однозначного указания на наличие или отсутствие явления или допускает несколько вариантов ин-

терпретации данных. Например, в позиции «A-6: Противопоставление гласных по назализации» для дардских языков нет возможности выбрать ни тип значения *Not stated*, ни значение из списка *Listed* «A-6-1: Противопоставление гласных по назализации отсутствует» либо какое-то другое допустимое значение, поскольку для этого недостаточно информации (дамели, калаша). В подобном случае через *Custom* вводится значение: «Есть назализованные и неназализованные гласные» со следующим текстом в комментарии: «Фонологический статус назализации гласных неясен».

Достаточно частой и более сложной для принятия решения является ситуация, когда выбор затруднен отсутствием не только эксплицитного описания явления или примеров в статье энциклопедии, но и иллюстраций в виде парадигм (несмотря на то, что парадигмы предусмотрены типовыми схемами статей и рекомендациями для авторов энциклопедии). К таким случаям относится заполнение позиции «K-6: Количество типов спряжения глаголов», для которой возможен выбор типа значения *Not stated* или «K-6-1: Единый тип», на том основании, что не указано противного (кховар, пхалура).

К особому типу контекстов относятся ситуации, когда описание не содержит однозначного указания на наличие или отсутствие явления. Такие контексты допускают возможность двоякой интерпретации данных в формате реферата: либо признать отсутствие в языке соответствующего явления, либо отметить отсутствие данных о нем в статье. Например, в описании языка шумашти нет прямого указания на наличие/отсутствие противопоставления гласных по признаку лабиализации. Во избежание возникновения артефакта в позиции «A-5: Противопоставление гласных по лабиализации» более корректным представляется выбор типа значения *Not stated*, а не допустимого значения «A-5-1 Противопоставление гласных по лабиализации отсутствует» из списка *Listed*. В противном случае отсутствие данных может восприниматься как отсутствие в языке явления.

С задачей разграничивать при наполнении базы данных отсутствие сведений в описании и отсутствие явления в языке сталкиваются также и разработчики морфосинтаксической базы данных *Grambank* [Skirgård, Haynie, Blasi et al. 2023] — международно-

го проекта, существующего под эгидой *Max Planck Institute* (*Leipzig*)[1]. В реферате базы данных «Языки мира» среди типов значений есть тип значения *Not stated*, аналогичный значению *Not mentioned* в анкете *Grambank.* Если вместо выбора этого типа значения выбрать значение «Отсутствует» из списка *Listed*, то ему в анкете *Grambank* соответствует значение *No category*. В формулировке участников проекта *Grambank* это проблема обозначается как *absence of evidence vs. evidence of absence* [Lesage, Haynie, Skirgård, Weber, Witzlack-Makarevich 2022, 2885].

Особой масштабной проблемой является расхождение между принятой в реферате и традиционной для конкретной описательной традиции классификацией языковых явлений. Речь идет прежде всего о фонологии. Например, для иранских [Языки мира 1999б], а также для нуристанских [Языки мира 1999а] и дардских языков принята принципиально иная классификация согласных, чем для большинства других индоевропейских языков. Так, классификация согласных в позиции «A-12: Инвентарь шумных согласных по способу артикуляции» предусматривает значение «A-12-3: Взрывные, фрикативные и аффрикаты». Термин *взрывные* является гипонимом по отношению к термину *смычные*. В описаниях иранских, дардских и нуристанских языков используется только термин *смычные*. Поэтому предлагается ввести этот термин в классификацию шумных согласных через тип значения *Custom*: «Смычные, фрикативные и аффрикаты». Но даже такое уточнение не отражает специфику и полноту классификации консонантизма в обсуждаемых языках. Поэтому она приводится в поле «Комментарий» с сохранением типичной для описаний этих языков схемы, например: «Смычные: чистые (неносовые и носовые), аффрикаты (одно- и двухфокусные) и щелевые: однофокусные (срединные и боковые) и двухфокусные» (кашмири).

Следует обратить внимание, что сонорные в дардских языках классифицируются как смычные и фрикативные (дамели, гарви и др.) или смычно-проходные (шина). Принятая в реферате и совпадающая с общепринятой классификация сонорных в позиции

[1] Адрес базы данных: https://grambank.clld.org/; https://github.com/grambank/grambank.

«A-19: Инвентарь сонорных согласных по способу артикуляции» включает в себя допустимое значение «A-19-6: Назальные, плавные, вибранты и глайды». В статьях о дардских языках сонорные интерпретируются иначе, что отражается в комментарии: «Назальные определяются как смычные чистые носовые, плавные — как щелевые боковые. Глайды /w/ и /y/ определяются как щелевые срединные однофокусные согласные» (гавар, гарви, калаша, кховар и подавляющее большинство других дардских языков).

Этот и целый ряд подобных примеров свидетельствуют о дилемме, которая встает перед составителями рефератов при переносе данных из оригинальных текстов статей энциклопедии в реферат: с одной стороны, не потерять термины, использованные в статье, а с другой стороны — интерпретировать языковые явления в терминах, принятых в типологической базе данных. При составлении очерков в энциклопедии авторы пользовались т. н. типовыми схемами описания языков и рекомендациями по использованию терминологии. Несмотря на это, унифицировать описания в достаточной степени не удалось: каждая из статей несет на себе следы авторских методики и стиля, времени написания, описательной традиции и школы. Для того, чтобы не терять и не искажать языковую информацию при составлении реферата, оптимальным представляется использовать поле «Комментарий». В этом поле может размещаться широкий диапазон данных различной степени детализации — например, ограничения при реализации языкового явления. Так, редкое употребление в языке шумашти форм множественного числа существительных с числительными в позиции «G-5: Форма имен после числительного» при выборе допустимого значения «G-5-3: Единственное и множественное» оговаривается в комментарии: «Существительные редко выражают мн. число». Подобное ограничение снимает риск артефакта в виде обобщения этого явления. «Комментарий» может также служить средством интерпретация специфической терминологии с целью «умеренной унификации» [Коломацкий 2023].

Комментарий сообщает необходимую гибкость переводу данных из энциклопедической статьи в формат реферата. Его содержание специальным образом не регламентируется, но может быть типизировано следующим образом.

— Комментарий общего типа, например, сведения о месте языка среди близкородственных идиомов в случае, когда полномасштабная лингвистическая характеристика языка не приводится, в первую очередь, из-за недостатка данных. Такова, например, ситуация с дардским языком нингалами, для которого дается лишь отсылка к представленным в соответствующем томе описаниям близкородственных языков шумашти, говар и, особенно, глангали.

— Специфический комментарий, который уточняет значение термина-гиперонима, используемого в наименовании признака. Таковы комментарии к позиции «F-1: Количество согласовательных классов» в языках, имеющих категорию грамматического рода. Например, большинство дардских языков имеют два рода, мужской и женский. При отмеченном допустимом значении признака «F-1-2: Два» комментарий содержит следующий текст: «Различаются мужской и женский род».

— Уточняющий, или ограничительный, комментарий. Комментарии такого типа содержат ограничения для выбранного значения из списка допустимых. Например, в позиции «G-4: Атрибутивное согласование по числу» при выборе значения «G-4-4: Предикативное и атрибутивное» комментарий выглядит следующим образом: «Атрибутивное согласование по числу ограничено указательными местоимениями» (гавар; прилагательные в этом языке не имеют категории числа). Или в позиции «G-5: Форма имен после числительного» для языка майян при выбранном допустимом значении: «G-5-3: Единственное и множественное» употребление форм множественного числа уточняется фразой: «Часть существительных не имеет категории числа» в комментарии.

Термины, употребляемые в реферате для определения категорий и значений признаков, чаще всего имеют форму множественного числа, поскольку грамматическая категория как правило имеет более одной граммемы. Комментарий позволяет избежать иска-

жения информации в случаях, когда в языке присутствует лишь один элемент соответствующей категории. Так, при выбранном значении «A-14-2: Билабиальные и лабиодентальные» из числа шумных губных согласных для языка гарви комментарий позволяет уточнить, что в нем есть только один, глухой, лабиодентальный согласный. Для языка катаркалаи при отмеченном значении «A-19-6: Назальные, плавные, вибранты и глайды» в позиции «A-19: Инвентарь сонорных согласных по способу артикуляции» оговаривается наличие единственного глайда /у/. То же самое справедливо по отношению к артиклям в ряде кельтских (валийский, ирландский) и славянских (болгарский и македонский) языках, в которых есть только определенный артикль, и некоторых дардских (тирахи) и иранских (белуджский, талышский) языках, в которых есть только неопределенный артикль.

— Комментарий относительно употребления и синонимии терминов. Применяется для характеристики языкового явления и представления терминологии, принятой в соответствующей статье энциклопедии в случае, если она отличается от принятой в реферате. Примером синонимии и уточнения специфической терминологии, которая варьирует в различных школах, является термин *церебральные* (согласные) в описаниях иранских и дардских языков и его синонимы *ретрофлексные* в традиции описания дравидийских (тамильский, малаялам, телугу, колами и др.) [Языки мира 2013] и некоторых германских (шведский) [Языки мира 2000] языках. В германистике этот термин употребляется наряду с термином *какуминальные* (норвежский). Также, интерпретация назальных сонорных как смычных, а латеральных и глайдов — как фрикативных, принятая для дардских и иранских языков, поясняется в комментарии к признаку «A-12: Инвентарь шумных согласных по способу артикуляции». Без подобного комментария очевиден риск появления артефактов, связанных с набором допустимых значений в реферате, где к шумным согласным относятся смычные, фрикативные и аффрикаты, а сонорные выделяются в отдельную позицию «A-19: Инвентарь сонорных согласных по способу артикуляции» с традиционной классификацией на назальные, плавные, вибранты и глайды.

— Комментарий, поясняющий причину отсутствия данных. Такой тип комментария приводится в комбинации с типом значения *Explicit gap*, который используется, когда в статье явно сказано, что данных нет. Комментарий имеет вид цитации в формулировке, приведенной автором статьи. Например, для позиции реферата «N-1: Линейный порядок компонентов в сложном предложении» комментарий становится единственным полем, где можно сохранить имеющиеся в тексте статьи сведения, которые могут оказаться существенными: «Сложное предложение не изучено» (гарви) или: «В зафиксированном тексте нет сложных предложений. Предполагается, что сложное предложение развито слабо» (торвали).

Основной целью данной статьи является выявление ситуаций, в которых вероятнее всего возможно появление типичных артефактов — искажения языковых реалий при необходимости выбирать то или иное значение в формате реферата, где они сформулированы в виде дихотомии или дизъюнкции. Предлагаемые варианты решений можно использовать в качестве прототипов при заполнении рефератов однородных массивов текстов, например, статей одного автора или описаний конкретной группы/семьи языков на основе единой методологии. Это поможет унифицировать процедуру составления рефератов.

Подводя итоги, следует отметить, что создание рефератов не сводится к автоматическому переносу данных из статей энциклопедии «Языки мира» в формат типологической базы данных. Он включает в себя расширенный поиск данных во вводных статьях каждого из томов, анализ терминологии, принятой в конкретных описательных традициях и школах, вплоть до индивидуальных особенностей стиля авторов статей. Не всегда грамматическая информация лежит на поверхности и находится в ожидаемом разделе типовой схемы статьи. Нередки случаи, когда, например, при перечислении состава частей речи некоторые из них не называются, но характеризуются с точки зрения формы и функций в дальнейшем описании. Особенно ответственной задачей для заполняющих реферат является необходимость извлекать данные с помощью анализа примеров. Так, в

реферате на основании статьи «Древненовгородский диалект» [Языки мира 2017, 469 — 475] в позиции «N-5: Тип связи между элементами сложного предложения» нет возможности выбрать ни одно из допустимых значений *Listed*, но на основании одного из примеров следует, что союзная связь в диалекте существовала. В реферате это отражается с помощью дополненного значения в тип *Custom*: «Союзная». Это убеждает нас в том, что языковые примеры должны подвергаться тщательному анализу, чтобы выявлять сведения о не описанных явным образом грамматических явлениях с достаточным основанием и без опасений порождения артефактов.

Степень стандартизации представления языковых данных в разных томах и статьях энциклопедии, несмотря на усилия редакторов, бывает различной. В ряде случаев своеобразие представления данных и используемой при описании языка терминологии обусловлены конкретной описательной традицией. Эти обстоятельства, на наш взгляд, не должны рассматриваться как недостаток энциклопедии. Напротив, такие случаи иллюстрируют своеобразие конкретной описательной традиции и могут служить источником данных для историков языкознания. Описания, которые на сегодняшний день могут показаться устаревшими, остаются материалом для изучения конкретных лингвистических традиций [Коломацкий 2023]. Для лингвистов-типологов информация в базе данных, равным образом как ее отсутствие, может указать пути новых направлений изысканий.

Список сокращений

БРЭ —Большая Российская Энциклопедия (в 30 т.). М.: Научное издательство «Большая российская энциклопедия», 2005. Т. 2. Анкилоз — Банка. — С. 283 // https://old.bigenc.ru/biology/text/1832035.

Литература

Зотова А.К., Романова О.И. «Свой» среди «чужих», или Потенциал термина в системе (на материале терминологии новой версии базы данных «Языки мира» ИЯз РАН) // Лингвистика и методика преподавания иностранных языков. Периодический сборник

научных статей. Электронное научное издание. М.: ИЯз РАН, 2021. Выпуск 14, №1. С. 27—48. DOI: 10.37892/2218-1393-2021-14-1-27-48.

Зотова А.К., Коломацкий Д.И., Романова О.И. Значимое отсутствие: лакуны в описании языков (на материале базы данных «Языки мира» ИЯз РАН) // Лингвистика и методика преподавания иностранных языков. Периодический сборник научных статей. Электронное научное издание. М.: ИЯз РАН, 2022. Выпуск 16, №1. С. 20—38. DOI: 10.37892/2218-1393-2022-16-1-20-38.

Коломацкий Д.И. Summa Typologiae: База данных «Языки мира» и дилемма формализованного описания языков. М., 2023 (в печати).

Языки мира 1999а — Языки мира: Дардские и нуристанские языки. М.: Индрик, 1999.

Языки мира 1999б — Языки мира: Иранские языки. II. Северо-западные иранские языки. М.: Индрик, 1999.

Языки мира: Германские языки. Кельтские языки. М.: Academia, 2000.

Языки мира: Дравидийские языки. М.: Academia, 2013.

Языки мира: Славянские языки (изд. 2-е, испр. и доп.). СПб.: Нестор-История, 2017.

Lesage J., Haynie H.J., Skirgård H., Weber T., Witzlack-Makarevich A. Overlooked Data in Typological Databases: What Grambank Teaches Us About Gaps in Grammars // Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). Marseille, 2022. Pp. 2884—2890.

Skirgård H., Haynie H.J., Blasi D.E., Hammarström H., Collins J., Latarche J.J., Lesage J., Weber T., Witzlack-Makarevich A., Passmore S., Chira A.M., Maurits L., Dinnage R., Dunn M., Reesink G., Singer R., Bowern C., Epps P.L., Hill J., Vesakoski O., Robbeets M., Abbas N.K., Auer D., Bakker N.A., Barbos G., Borges R.D., Danielsen S., Dorenbusch L., Dorn E., Elliott J., Falcone G., Fischer J., Ate Y.G., Gibson H., Göbel H.-P., Goodall J.A., Gruner V., Harvey A., Hayes R., Heer L., Miranda R.E.H., Hübler N., Huntington-Rainey B.H., Ivani J.K.,

Johns M., Just E., Kashima E., Kipf C., Klingenberg J.V., König N., Koti A., Kowalik R.G.A., Krasnoukhova O., Lindvall N.L.M., Lorenzen M., Lutzenberger H., Martins T.R.A., German C.M., van der Meer S., Samamé J.M., Müller M., Muradoglu S., Neely K., Nickel J., Norvik M., Oluoch C.A., Peacock J., Pearey I.O.C., Peck N., Petit S., Pieper S., Poblete M., Prestipino D., Raabe L., Raja A., Reimringer J., Rey S.C., Rizaew J., Ruppert E., Salmon K.K., Sammet J., Schembri R., Schlabbach L., Schmidt F.W.P., Skilton A., Smith W.D., de Sousa H., Sverredal K., Valle D., Vera J., Voß J., Witte T., Wu H., Yam S., 葉婧婷 J.Y., Yong M., Yuditha T., Zariquiey R., Forkel R., Evans N., Levinson S.C., Haspelmath M., Greenhill S.J., Atkinson Q., Gray R.D. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss // Science Advances, Vol. 9, Issue 16 (19 Apr 2023). DOI: 10.1126/sciadv.adg6175.

Description of Linguistic Data: Facts and Artefacts ("Languages of the World" Database, Institute of Linguistics, Russian Academy of Sciences)

A.K. Zotova, O.I. Romanova

(Institute of Linguistics of the Russian Academy of Sciences)

The article examines the possibilities of specific language phenomena presentation and the causes of inaccuracies when transferring linguistic data from the encyclopedia edition "Languages of the World" to the web-version of the "Languages of the World" typological database. The material is intended for those interested in linguistic typology, data formalization and the development of linguotypological databases.

Keywords: languages of the world, linguistic data formalization, linguotypological databases.