DOI: 10.37892/2218-1393-2022-16-1-20-38

Значимое отсутствие: лакуны в описании языков (на материале базы данных «Языки мира» ИЯз РАН)

А.К. Зотова, Д.И. Коломацкий, О.И. Романова

(Институт языкознания РАН)

Аннотация

В статье рассматриваются случаи отсутствия информации о языках на материале последней версии базы данных «Языки мира» ИЯз РАН. Анализируются причины и типы лакун, их связь с особенностями терминологических традиций и эквивалентами других терминосистем на основе в первую очередь энциклопедического издания «Языки мира» ИЯз РАН. Материал может быть интересен специалистам в области прикладной лингвистики, лингвистической типологии и баз данных.

Ключевые слова

Лингвистическая типология, лингвистическая терминология, базы данных, языки мира.

База данных «Языки мира» Института языкознания РАН была задумана и создаётся в качестве специализированного информационного ресурса, содержащего лингвотипологическую информацию об отдельных языках, группах и семьях языков в виде унифицированных формализованных описаний (рефератов) и дополнительных возможностей, предусмотренных пользовательским интерфейсом. За годы своего существования БД претерпела несколько этапов усовершенствования, последние из которых связаны с адаптацией к возможностям Интернета. Статья содержит материал, иллюстрирующий направление работы по теме «Совершенствование лингвистического и программного обеспечения Базы данных «Языки мира» и ее расширение: создание современного лингвистического ресурса для исследований по типологии и прикладной лингвистике». Одновременно она является продолжением проводившихся в предыдущие годы исследований (на материале имеющихся в

БД германских, романских, славянских и других языков), целью которых было изучение возможности использования БД для проведения качественных дескриптивных, контрастивных и сравнительно-исторических исследований языковых явлений в семьях и группах языков.

Актуальная четвёртая версия Базы данных «Языки мира» была написана на языке программирования *Java* (разработчик — м.н.с. сектора прикладного языкознания ИЯз РАН Е.А.Макарова) и опубликована на сайте Института языкознания РАН в 2020 г. Для использования базы данных её требуется скачать и установить.

В 2020 г. в Секторе прикладного языкознания Института начата разработка веб-интерфейса базы данных. Новый вариант интерфейса сам по себе не может считаться отдельной версией базы данных, поскольку изначально онлайн-версия запрограммирована так, чтобы использовать те же файлы исходных данных, что и десктопная версия. Вместе с тем переход к онлайн-формату позволил реализовать опции, которые были недоступны в десктопном варианте.

Так, непосредственно из веб-интерфейса пользователь получит возможность открывать тома энциклопедии «Языки мира», чтобы ознакомиться с развёрнутым описанием языка (и при желании сравнить его с формализованным вариантом, которым является реферат) или просмотреть языковые примеры.

Кроме того, список языков, указатели и мастер запросов будут снабжены интерактивной картой, на которой в виде маркеров будут отображаться найденные языки. По мере перемещения или изменения масштаба карты будет корректироваться список языков, которым соответствуют отображенные на выбранной области маркеры.

Другие, менее заметные изменения интерфейса помогут пользователю работать с базой данных быстрее и удобнее, без лишних манипуляций с признаками или языковыми семьями.

Что касается наполнения базы данных формализованными описаниями языков (рефератами), основным требованием к онлайн-версии было обеспечить возможность лёгкого

добавления новых признаков, значений признаков и языков. В третьей версии от 2013 г. такой возможности не было, в десктопном варианте четвёртой версии этот процесс затруднён необходимостью производить сложные технические манипуляции при изменении общего набора признаков или допустимых значений какого-либо признака. В коде онлайн-версии эти процессы максимально упрощены, в частности, добавление нового значения признака в общий перечень не приводит к необходимости править описания языков, для которых это новое значение признака проставлять не требуется. В процессе конвертации текстовых анкет (рефератов) в более современный и универсальный формат *CSV*, широко используемый для представления табличных данных[1], был выявлен ряд методологических вопросов, требующих изучения и обсуждения. Среди них одним из ключевых стал вопрос о лакунах в описаниях языков.

В анкетах языков можно обнаружить несколько типов лакун и формы их обозначения. Причины возникновения таких лакун разнообразны и требуют тщательного анализа при выявлении.

В первую очередь, лакуны естественным образом возникают в тех случаях, когда отсутствуют объективные данные о тех или иных фрагментах структуры языка. Это касается вымерших, малоизученных или бесписьменных языков. Для описания таких языков ещё на этапе планирования энциклопедии «Языки мира» были предусмотрены четыре типовые схемы от более подробных о семье языков, группе языков или группе диалектов (тип I), статье о языке (тип II) до самых кратких для описания диалектов (тип III) и отдельных недостаточно изученных языков (тип IV). При этом объём статей не всегда однозначно коррелирует со статусом (живой или вымерший) или степенью изученности языка (малоизученный язык, бесписьменный диалект).

В целом ряде разделов анкеты БД отсутствие явления предусмотрено в списке значений в составе признаков:

[1] Анкета является, по сути, таблицей, где каждой строке соответствует один признак, название которого указано в левом столбце, а в остальные столбцы вписываются значение и комментарий.

— A-5. Противопоставление гласных по лабиализации → Противопоставление гласных по лабиализации *отсутствует*

— A-6. Противопоставление гласных по назализации → Противопоставление гласных по назализации *отсутствует*

— A-7. Противопоставление гласных по фарингализации → Противопоставление гласных по фарингализации *отсутствует*

— A-8. Противопоставление гласных по продвинутости корня языка → Противопоставление гласных по продвинутости корня языка *отсутствует*.

В других позициях для констатации отсутствия некоторого явления в языке используются иные синтаксические конструкции. Так, в списке значений можно найти:

— E-3. Признаки агглютинативного строя → *Отсутствие* полисемантизма аффиксов

— F-1. Количество согласовательных классов → Согласовательные классы *отсутствуют*

— F-6. Согласование числительных в роде → *Отсутствует*

— G-3. Гонорифические формы числа → *Отсутствуют*

— G-7. Неопределенный артикль и числительное «один» → Неопределенный артикль *отсутствует*

— I-1. Способ выражения залоговых форм в глаголе → Категория залога *отсутствует*

— K-2. Типы артиклей → Артикли *отсутствуют*.

Подобные перечисленным выше явные способы указания на отсутствие явления или значения признака в языках также относятся к категории лакун по объективным причинам.

Иначе складывается ситуация при пропуске, т.е. не-заполнении позиций, нерелевантных для отдельного языка, семьи или группы языков. В качестве примера рассмотрим

признак «A-8. Противопоставление гласных по продвинутости корня языка». Этот параметр, относящийся к акустической фонетике, не используется авторами статей при описании, например, индоевропейских языков, поэтому во многих анкетах эта позиция игнорировалась.

На данном этапе работа по редактированию БД предполагает использование некоторых принципов, направленных на унификацию анкет. В частности, если первый признак в разделе опускается как неприменимый к данному языку, то остальные признаки автоматически становятся нерелевантными и тоже опускаются. При описании нетоновых языков признак «B-6. Количество уровней тонов» может быть незаполненным. Но из этого не следует, что дальнейшие признаки «B-7. Контурные признаки тона (восходящий, нисходящий и др.)» и «B-8. Носитель тона (слог, слово)» несущественны. Тоновые характеристики традиционно используются авторами статей при описании языков разных систем и во избежание лакун эта информация должна фиксироваться в анкетах БД. Сведения о количестве тонов традиционно относятся к описанию тоновых языков (языков Юго-Восточной Азии: некоторых сино-тибетских, мяо-яо, тайско-кадайских; языков Африки: афразийских, нило-сахарских, койсанских, нигеро-конголезских; североамериканских атабаскских языков и др.). В этих языках число тонов составляет их характеристические особенности. Производными от него становятся словоизменительные и словообразовательные функции тонов, маркирующие число или род существительных, время или отрицательную форму глагола и др. При этом во многих тоновых языках связь тонов и ударения не прослеживается, поскольку тон не всегда связан с просодикой и является элементом слога.

Подтипом тоновых языков являются языки с музыкальным ударением, к которым относятся в том числе некоторые германские (шведский, норвежский) [Языки мира 2000a, 300—328; там же, 349—380], романские (португальский) [Языки мира 2001, 462—492] и славянские (сербохорватский) [Языки мира 2017, 151—212] языки. В этих языках количественные параметры (число тонов) менее важны, чем качественные — такие как контурные

признаки и единицы-носители тона. Кроме слога (долгого или краткого) или слова носителями тона могут быть более крупные синтаксические единицы, в которых элементы фразовой мелодики и интонационные контуры выполняют семиотические функции, например, маркируют типы вопросительных предложений в английском [Языки мира 2000a, 52—53], французском [Языки мира 2001, 208] или итальянском языках [Там же, 65].

Есть случаи, когда в описании языка в целом отсутствует то или иное явление, например, гармония гласных. Это отсутствие не фиксируется выбором значений в соответствующих позициях «B-10. Область гармонического уподобления гласных» или «B-11. Признак гармонического уподобления гласных». Явление сингармонизма, или точнее, прогрессивного сингармонизма, ассоциируется в основном с агглютинативными языками (в частности, алтайскими и уральскими). В то же время метафония в виде разновидности сингармонизма — регрессивного сингармонизма — свойственна также таким флективным романским языкам, как румынский, сардинский и, в более ограниченном масштабе, португальский. Компенсировать эту лакуну можно добавлением нового значения в признак «B-12. Уподобляющий элемент при гармоническом уподоблении гласных», добавив значение «гласный флексии» к имеющемуся там фиксированному списку «первый гласный слова», «первый гласный аффикса», «первый гласный корня», «последний гласный основы» и др.

Для размещения специфической для конкретного языка информации, которую не отражают типизированные значения в признаках анкеты, предусмотрена строка *Commentary*. Покажем, как эта строка заполняется, на следующих примерах.

Неполнота сведений о том или ином явлении не всегда обозначается авторами статей энциклопедии «Языки мира» как полное отсутствие данных. В частности, в описании готского языка отмечается, что «категория определённости/неопределённости имени не имела регулярного способа выражения» [Языки мира 2000a, 113]. В списке значений в составе признака анкеты «J-5. Морфологическое выражение определённости/неопределённости имени» есть вариант «отсутствует». Однако эти выражения не являются однозначно

эквивалентными и не таящими в себе лакун. Поэтому мы считаем более корректным дополнить предлагаемый анкетой выбор, включив формулировку автора статьи в строку *Commentary*. В позиции анкеты «J-6. Части речи, выражающие определённость/неопределённость имени» из предлагаемого списка значений выбрано «артикли». Компромиссом между точным определением частей речи, которые выполняли эту функцию, представляется строка *Commentary*, содержащая уточняющую информацию: «Артиклеобразное местоимение ‘этот’ употреблялось факультативно» [Языки мира 2000a, 113]. Такое внимание к тексту автора совершенно оправдано, поскольку речь идет о вымершем языке, относительно которого важна любая сохранившаяся деталь, и о знании об этом языке на момент написания статьи.

Близкое к этому решение было принято при описании ударения в старославянском (церковнославянском) языке. В позиции «B-1. Тип ударения» теоретически возможен выбор: «Ударение отсутствует». Тем не менее, учитывая фрагментарность и функционально-стилистическую специфику сохранившегося материала, более адекватным представляется сохранение формулировки автора статьи в строке *Commentary*: «Точных сведений об ударении в С.я. нет» [Языки мира 2017, 52].

Еще несколько примеров из описаний вымерших языков, где сохранение текстовых фрагментов статей энциклопедии позволяет избежать категоричности однозначных значений признаков и предупредить появление лакун:

— в древнеперсидском языке, признак «A-6. Противопоставление гласных по назализации», выбор: «Противопоставление гласных по назализации отсутствует» уточняется в *Commentary*: «Предположительно наличие живой или исторической назализации гласных» [Языки мира 1997, 39];

— в среднеперсидском языке, признак «B-2. Вид ударения» при имеющихся значениях «динамическое», «количественное», «музыкальное» или «смешанное», в строке *Commentary* уточняется: «Предполагается силовое ударение» [Языки мира 1997, 59];

— в языке велатру (северо-западные иранские языки) в признаке «I-1. Способ выражения залоговых форм в глаголе» выбор из имеющихся значений «категория залога отсутствует» корректируется в строке *Commentary*: «Пассивный залог в текстах не зафиксирован» [Языки мира 1999, 142].

По мере накопления и редактирования массива данных стала очевидна необходимость там, где возможно унифицировать формулировки, используемые при составлении анкет БД, в том числе формулировки об отсутствии данных в разных позициях. Так появилось несколько типизированных вариантов строки *Commentary*, не воспроизводящих формулировки авторов статей энциклопедии, но и не искажающих реальное положение вещей.

Формулировка «данных нет» используется в случаях, когда в описании вымерших и/или малоизученных языков автор не приводит соответствующей информации:

— B-1. Тип ударения *Commentary*: Данных нет (Авесты язык)

— F-5. Атрибутивное согласование по роду *Commentary*: Данных нет (парфянский язык)

— I-1. Способ выражения залоговых форм в глаголе *Commentary*: Данных нет (корнский язык)

— N-1. Линейный порядок компонентов в сложном предложении *Commentary*: Данных нет (древнерусский язык).

Есть варианты авторских формулировок об отсутствии данных, которые также помещаются в строку *Commentary*. Например, о наличии фонологических противопоставлениях единиц и категорий «D-4. Различия между знаменательными и служебными словами» *Commentary*: Систематизированных данных нет (каталанский язык) [Языки мира 2001, 501] или о характерных типах сложного предложения «N-1. Линейный порядок компонентов в сложном предложении» *Commentary:* Надежных данных о структуре сложного предложения нет (лепонтийский язык) [Языки мира 2000a, 457].

Формулировка «Других данных нет» используется в случаях, когда составители и редакторы анкет констатируют наличие некоторых данных и фиксируют их в определённых позициях БД с помощью имеющихся в списке значений признака, но считают необходимым отметить, что эти данные не представлены явным и непротиворечивым образом при описании языка в энциклопедии. Информация извлечена из примеров или почерпнута из обзорной статьи к соответствующему тому. Например:

— в списке значений признака «G-3. Гонорифические формы числа», выбираем «отсутствуют» и уточняем в строке *Commentary*: Других данных нет (церковнославянский язык);

— в позиции «H-2. Маркирование субъекта и объекта» добавляем значение «падежные аффиксы, глагольное согласование и порядок слов» и уточняем в *Commentary*: Других данных нет (шотландский язык);

— в позиции «K-1. Склонение личных местоимений» добавляем значение «Шестичленная падежная парадигма» в сочетании с *Commentary*: Других данных нет (кельтиберский язык);

— в списке значений признака «N-5. Тип связи между элементами сложного предложения» есть варианты «преобладает союзная», «преобладает бессоюзная» и «союзная и бессоюзная». Поскольку в источнике можно зафиксировать только примеры союзной связи, добавляем значение «союзная» и уточняем в *Commentary*: Других данных нет (древненовгородский язык/диалект).

В процессе редактирования анкет неоднократно пересматриваются и уточняются формулировки, маркирующие отсутствие языковых явлений. Так формулировка «категория отсутствует», предусмотренная в списке признаков в других случаях, не была представлена в признаке «H-1. Количество падежей». Но поскольку возможны несколько вариантов состава граммем категории падежа, в том числе у местоимения, во избежание лакун было решено добавить информацию в строку *Commentary* в следующем виде: «Категория

падежа у имени существительного отсутствует» (языки африкаанс, гасконский, франко-провансальский, хазара и др.).

Значительное число лакун в описании языков можно объяснить расхождениями в терминологии, используемой авторами статей энциклопедии, и той, которая применяется в анкетах БД. На данном этапе можно констатировать, что терминологические источники лакун, неточностей и ошибок связаны с адаптацией схем описания языков в энциклопедии к требованиям унифицированного формата БД. Поскольку анкета представляет собой сочетание классифицирующих (названия признаков) и терминологических (значения признаков) элементов, на этапе соотнесения устоявшихся терминов авторов статей и формулировок БД обнаруживается, что некоторые из них не всегда однозначно понимаются специалистами разных школ и/или не воспринимаются как эквиваленты в тех или иных разделах анкеты. Например, в позиции «B-2. Вид ударения» для португальского языка указано значение «динамическое с элементами количественного, качественного и тонического», при том что в наборе предлагаемых для выбора значений есть опция «смешанное», которая, однако, не позволяет адекватно описать природу ударения в португальском языке [Языки мира 2001, 467].

В признаке «L-1. Способы словообразования» лакуны может вызвать нетипизированное представление модели глагольного словообразования вида *V* + *Adv* (фразовые глаголы), характерной для ряда германских (английский, африкаанс) и романских (ладинский) языков. Терминологические обозначения этой модели в статьях энциклопедии могут быть самыми разными.

В признаке «J-8. Способы выражения отрицания» используются термины «отрицательные частицы» и «отрицательные аффиксы». При этом в статьях энциклопедии частицами могут называться и аффиксы («префиксальные частицы» — нидерландский, исландский), и собственно частицы, которые не входят в состав словоформы (словенский, македонский).

Начиная с первых версий, БД «Языки мира» разрабатывалась на основе оригинальных статей, написанных авторами для энциклопедии «Языки мира». В совокупности оба источника представляют собой результаты исследований нескольких поколений специалистов Института языкознания РАН. Помимо ценности содержащейся в них фактической информации, эти исследования иллюстрируют широкую палитру школ, направлений и традиций описания языков — славистики, иранистики, германистики, романистики и др., что представляет значительный самостоятельный, в том числе и исторический, интерес.

В качестве примера можно привести традицию использования термина «непредметное местоимение» в описании дейктических средств в романских языках для обозначения неодушевленного местоимения среднего рода. Местоимение среднего рода в дейктической функции употребляется также в германских языках, но в германистике этот термин не является частью традиции описания дейксиса, а правила употребления этого местоимения описываются в разделе синтаксиса. В результате возникает лакуна в описании способов выражения дейктических категорий (признак J-2 анкеты БД), например, в английском языке.

Уточнение значений признаков происходит по мере накопления материала, т.е. количества рефератов. Принимается во внимание частотность формулировок и терминов. Предполагается, что в итоге будет возможна количественная и качественная коррекция терминологии в составе анкет, в частности, замена некорректных формулировок на более систематизированные и соотносящиеся с терминологией статей энциклопедии.

Приведём примеры формулировок, требующих коррекции во избежание лакун. Для признака «I-1. Способ выражения залоговых форм в глаголе» в списке значений используются термины «вспомогательный глагол», «аффиксы» и «служебные слова». Во многих индоевропейских языках вспомогательный глагол участвует в выражении залоговых значений в составе конструкции «вспомогательный глагол + причастие» либо залоговые значения выражает та или иная форма причастия. Термин «служебные слова» в этом списке оказывается слишком общим и не позволяет назвать части речи и разряды служебных слов:

возвратное местоимение (польский), возвратное местоимение-частица (старославянский и церковнославянский язык, древнерусский язык), возвратный постфикс (русский язык), лично-возвратное местоимение (сербохорватский язык), возвратные частицы (белорусский язык). Если употреблять конкретные термины для наименования разрядов этих слов, то одновременно появляется возможность отразить грамматическую информацию и о способах их употребления (пре- или постпозиция, неизменяемость или изменяемость по лицу и числу).

Похожую ситуацию можно проиллюстрировать с помощью списка значений в признаке анкеты «H-6. Способ выражения пространственных отношений». Список содержит множество терминов и их комбинаций, в том числе таких, где одновременно фигурируют термины-гиперонимы и термины-гипонимы, например, «именные аффиксы» и «наречия места». С одной стороны, учитывая общий контекст анкеты, к числу именных аффиксов могут относиться падежные флексии, в том числе в именных конструкциях с предлогами. С другой стороны, термин «аффиксы» выступает в качестве гиперонима по отношению к терминам в признаке «L-2. Аффиксы, участвующие в словообразовании», который включает в себя префиксы, суффиксы, инфиксы, постфиксы, конфиксы, трансфиксы, супрафиксы или циркумфиксы. Некоторые из этих аффиксов могут быть словоизменительными и использоваться в составе предложных конструкций, в том числе и для выражения пространственных отношений. Второй элемент пары — наречия места, так же, как и местоименные наречия, можно рассматривать в качестве гипонимов по отношению к наречиям в целом. В этом же списке значений используется термин «местоимения», к которым при выражении пространственных отношений чаще всего относятся указательные местоимения. Таким образом, во многих анкетах в этом признаке наиболее частотными становятся комбинации терминов одного уровня: «местоимения, наречия и предлоги» (чешский язык, итальянский язык, меглено-румынский, фарерский, французский язык, фриульский язык), а также «указательные местоимения, наречия и предлоги» (истро-романский/ истриотский язык, окситанский язык), которые требуется включить в список значений.

В признаке «I-8. Синкретическое выражение нескольких глагольных значений» наблюдаются частотные комбинации значений признаков, которые во избежание лакун должны быть включены в список значений, например, «время, вид и модальность (наклонение)» (германские языки), «лицо, число, время, вид и модальность (наклонение)» (романские языки) и «лицо, число, род, время, вид и модальность (наклонение)» (многие славянские языки).

К особому типу лакун относятся количественные данные, которые нельзя интерпретировать в существующем формате признаков анкеты. Этот тип лакун связан с особенностями описания в энциклопедии ряда малоизученных или бесписьменных языков и языков с пограничным статусом (язык/диалект), когда авторы статей при описании фонетических систем ограничиваются лишь списками фонем, не сопровождая их таблицами или терминологической интерпретацией. Значительное количество таких примеров обнаруживается в описании иранских языков. Например, в признаке «A-1. Количество степеней подъема» нет возможности выбрать ни одного из предлагаемых значений, поскольку в описаниях указывается только количество гласных фонем и их вариантов, которые терминологически не интерпретируются (фаризанди [Языки мира 1999], сурхеи язык/диалект [Языки мира 1999, 154], осетинский язык [Языки мира 2000б, 313]). То же самое распространяется и на классификацию согласных в признаке «A-13. Инвентарь шумных согласных по месту образования», где для многих языков списки фонем терминологически не интерпретируются — например, велатру язык/диалект [Языки мира 1999, 141].

Вопрос о лакунах в рефератах не был должным образом разрешён в третьей версии и десктопном варианте четвёртой версии, поскольку кодирование отсутствия значения было недостаточно последовательным и/или не предполагало достаточной дифференциации типов лакун.

В третьей версии для языков, «типологическое пространство которых описано очень слабо... отсутствующие ветви дерева бинарной части реферата помечаются знаком ".O" (ИНФОРМАЦИЯ ОТСУТСТВУЕТ)» [Поляков и др. 2019, 77–78]. Можно заключить,

что такая маркировка используется для случаев, когда в энциклопедической статье содержится информация об отсутствии релевантных данных, что сравнительно часто встречается в энциклопедических описаниях мёртвых языков. Вместе с тем, «в некоторых слабоописанных языках [статус признака] ЛОЖЬ (FALSE) означает “НЕИЗВЕСТНО” <…> [Н]екоторые признаки со статусом ЛОЖЬ (FALSE) в описании такого языка можно интерпретировать как НЕИЗВЕСТНО» [Там же, 78, 81]. Таким образом, в ряде случаев отсутствие данных о значении признака в третьей версии выражено имплицитно; более того, из этого описания трудно понять, идёт ли речь об отсутствии данных в энциклопедической статье или о наличии в статье явного указания на отсутствие данных.

В десктопном варианте четвёртой версии анкета заполняющего представляет каждый признак в виде блока строк. Если для данного языка возможен выбор значения из фиксированного списка, напротив соответствующего значения ставится «1», остальные значения помечаются флагом «0». Если ни одно из фиксированных значений не подходит, у всех таких значений ставится «0», а в поле *Other* вписывается текстовое значение. Кроме того, заполняющий имеет возможность отметить все возможные значения признака знаком «x». Это сигнализирует об отсутствии информации о данном признаке.

При этом у заполняющего нет способа указать тип отсутствующей информации. Теоретически можно было бы разработать конвенцию, согласно которой, например, знак «0» у каждого значения обозначал бы ситуацию «нет данных в статье», а знак «x» — ситуацию «в статье указано, что данных нет». Кроме того, можно пользоваться полем *Commentary* и разработать систему помет. Но ни одно из этих решений не является достаточно надёжным, т.к. вместо предоставления пользователю ясного выбора опций требует соблюдения договорённостей всеми участниками проекта.

В качестве решения этой проблемы в обновлённый вариант анкеты заполняющего решено ввести дополнительный параметр, который можно назвать «типом значения» или «статусом признака». Он может принимать пять значений (в скобках указываются условные кодовые наименования для последующей программной обработки):

1. Значение признака доступно в заранее утверждённом списке допустимых значений (‘*listed*’);

2. Значение признака можно указать, но оно отсутствует в фиксированном списке (‘*custom*’);

3. Для этого языка не может быть выбрано ни одно значение признака, т.к. этому противоречит выбранное ранее значение другого признака (‘*not_applicable*’);

4. В энциклопедической статье нет информации о значении признака (‘*not_stated*’);

5. В энциклопедической статье явно указано, что данных об этом признаке нет (‘*explicit_gap*’).

Таким образом, если представлять данные о языке в табличной форме, то значение признака будет выражаться двумя столбцами — типом значения и (если тип это позволяет) текстовым значением.

Эта кажущаяся сложность заполнения анкеты компенсируется средствами проверки на этапе заполнения и этапе обработки. В частности, даже при использовании для анкеты заполняющего таких простых средств, как таблица *Microsoft Excel*, по мере заполнения столбцов таблицы можно подсвечивать возможные ошибки. Кроме того, при программной обработке реферата несложно организовать проверку на наличие ошибок заполнения (например, если тип значения не позволяет вводить текстовое значение, а оно было указано, это может означать, что тип значения был выбран неверно).

Особого внимания требует заполнение признаков, связанных друг с другом имплицитными связями. В тех случаях, когда для описываемого языка не может быть выбрано ни одно из значений, поскольку этому противоречит выбранное ранее значение другого признака, все признаки блока, за исключением первого, маркируются кодовым наименованием ‘*not_applicable*’ (пункт 3 списка выше).

К случаям такого рода относятся следующие блоки признаков, в каждом из которых для первого признака в блоке есть возможность отметить значение «Категория/явление *N* отсутствует»:

— A-9. Наличие дифтонгов и трифтонгов, A-10. Типы дифтонгов.

При выборе в признаке A-9 значения «дифтонги и трифтонги отсутствуют» ни одно из доступных для выбора значений признака A-10 не может быть выбрано;

— B-1. Тип ударения, B-2. Вид ударения, B-3. Носитель ударения, B-4. Фиксированность ударения, B-5. Носитель фиксированного ударения в слове.

Выбор в первом из этих признаков значения «ударение отсутствует» влечет за собой маркирование последующих четырех признаков пометой ‘*not_applicable*’. При этом возможны ситуации, когда признаки B-1 и B-3 заполнены, но ударение не фиксировано. В этом случае признак B-4 получает значение «Нефиксированное», а следующий признак B-5 маркируется как ‘*not_applicable*’;

— D-7. Функциональный тип чередований, D-8. Вид чередований.

При выборе значения «чередования отсутствуют» в первом признаке для второго справедлива помета ‘*not_applicable*’;

— F-1. Количество согласовательных классов, F-2. Морфологические способы выражения согласовательных классов, F-3. Синтаксические способы выражения согласовательных классов, F-4. Другие способы выражения согласовательных классов; F-5. Атрибутивное согласование по роду.

Блок признаков, представляющих информацию о системе падежей:

— H-1. Количество падежей, H-3. Падежное оформление именного сказуемого, H-4. Падежное оформление посессивного отношения, H-8. Падежное оформление одушевленных и неодушевленных имен, H-9. Наличие вторичных падежей также связаны отношениями импликации. В случае отсутствия в языке категории падежа в системе имени существительного это указывается в признаке H-1. Тогда при заполнении признаков H-3, H-4 и H-8 и H-9 применима помета ‘*not_applicable*’.

Аналогичным образом, если в признаке «I-1. Способ выражения залоговых форм в глаголе» отмечено значение «отсутствует», то для следующих признаков: «I-2. Возможность наложения нескольких залоговых форм», «I-3. Двойные залоговые формы» и «I-4. Совпадение залоговых форм» применима кодировка '*not_applicable*'.

Если в признаке «I-5. Основные временные противопоставления» отмечено значение «отсутствует», то для следующих признаков «I-6. Выражение вида и времени» и «I-7. Способ выражения временных противопоставлений» применима кодировка '*not_applicable*'.

И наконец, если в признаке «K-2. Типы артиклей» отмечено значение «артикли отсутствуют», то для следующих признаков: «K-3. Расположение артикля» и «K-4. Постановка артикля в именной группе» и «K-5. Грамматические категории, выражаемые артиклем» применима кодировка '*not_applicable*'.

Введение нового параметра (тип значения или статус признака) позволит добавить в онлайн-версию новые возможности для пользователя базы данных. Отсутствие информации о том или ином признаке, точно так же как и её наличие (или даже в большей степени), может служить отправной точкой для самостоятельных исследований.

Если взглянуть на проблему лакун с точки зрения пользователя базы данных, то приходится констатировать, что на настоящий момент у пользователя нет надёжного способа получить о них систематизированную информацию. В десктопной версии базы данных действительно информативными можно считать только те значения признаков, которые входят в фиксированный перечень: лишь они отображаются в указателе признаков и мастере сложных запросов. В остальных случаях приходится обращаться к описанию конкретного языка. Более того, если информация о признаке отсутствует, в текущем варианте БД данный признак просто не будет отображаться в этом описании. Это означает, что пользователь должен очень хорошо представлять себе перечень всех признаков, чтобы понять, что информация о каком-либо из них отсутствует, и даже в этом случае не будет никакого средства выяснить, к какому типу принадлежит найденная лакуна.

Появление в формализованном представлении языка чётко обозначенных лакун позволит особым образом маркировать их на странице каждого языка. Это само по себе позволит быстро оценить, насколько хорошо конкретный язык описан (в пользовательском интерфейсе можно ввести и маркер полноты заполнения реферата, показывающий, например, процент заполненных признаков от их общего числа), но этим открывающиеся возможности не исчерпываются. Указатель признаков можно будет настроить так, чтобы он показывал перечни языков, для которых по выбранному признаку нет данных, а в общем перечне языков можно будет ввести маркировку полноты описания каждого языка.

Энциклопедия «Языки мира» относится к тем авторитетным изданиям, с которых многие лингвисты начинают изучение выбранной ими проблемы. Формализованное представление энциклопедических данных с быстрым доступом к самим томам энциклопедии, несомненно, облегчит исследователям этот первый шаг. Однако не менее важным для них, как представляется, будет формализованная информация о пробелах в описании. Сравнение числа заполненных признаков с их общим количеством даст более ясное представление о том, на каких языках в зависимости от типа работы следует сосредоточиться исследователю. В этом смысле можно утверждать, что формализованные данные о наличии информации по той или иной области описания языка лишь облегчают учёному работу над материалом, который в остальном он и так смог бы обнаружить в энциклопедии. В частности, помета и комментарий о недостатке данных и неизученности некоторого языкового явления в конкретном языке — основание для выделения следующего этапа исследований.

Указание на отсутствие данных, с другой стороны, предлагает исследователю важнейшую информацию, которую он мог бы не воспринять при простом чтении энциклопедической статьи. Таким образом, описанные нововведения в базе данных «Языки мира» могут превратить недостаток информации в наиболее ценную информацию для лингвиста.

Литература

Поляков В.Н., Соловьев В.Д., Макарова Е.А. База данных «Языки мира»: история и перспективы. Москва; Казань: Институт языкознания РАН, 2019.

Языки мира: Иранские языки. I. Юго-западные иранские языки. М.: Индрик, 1997.

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

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

Языки мира: Иранские языки. III. Восточноиранские языки. М.: Индрик, 2000б.

Языки мира: Романские языки. М.: Academia, 2001.

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

Meaningful Absence: Lacunae in the «Languages of the World» Database

(Institute of Linguistics, Russian Academy of Sciences)

A.K. Zotova, D.I. Kolomatskiy, O.J. Romanova

(Institute of Linguistics,

Russian Academy of Sciences)

The paper outlines lacunae in the latest version of the «Languages of the World» Database (Institute of Linguistics, RAS). Lacunae of various origins and types are analysed along with linguistic terminology representing different descriptive traditions, primarily based on the «Languages of the World» encyclopedia (Institute of Linguistics, RAS). The topic can be of interest for those dealing with applied linguistics, linguistic typology and databases.

Keywords: Linguistic typology, linguistic terminology, databases, languages of the world.