КРАУДСОРСИНГ КАК СПОСОБ НАПОЛНЕНИЯ БАЗЫ ДАННЫХ «ЯЗЫКИ МИРА»

Макарова Елена Андреевна
М.н.с., ФГБУН институт языкознания Российской Академии Наук, Москва
antaresselen@mail.ru

Аннотация. В данной работе рассматривается 4-я версия базы данных «Языки мира». Ее главной особенностью стал переход от иерархического представления данных, при котором признаки были оформлены в виде бинарного дерева, к парадигматическому. Был разработан список из 124 признака, каждый из которых имеет, в среднем, восемь вариантов значений. Долгое время единственным источником информации для заполнения рефератов языков в базе данных служило одноименное энциклопедическое издание. Однако в ближайшее время, как только данные обо всех описанных в энциклопедии языках будут введены в базу, проблема поиска альтернативных источников информации встанет особенно остро. Решением данной проблемы может стать создание опросников и проведение краудсорсинг-проекта. Лингвистические опросники чаще всего используются специалистами при описании новых языков. Главной же особенностью предлагаемого краудсорсинга станет участие в нем не только специалистов по конкретным языкам, но и носителей, не имеющих лингвистического образования. Таким образом, возникает необходимость создания двух видов опросников. Первый вид предназначен для работы со специалистами и включает в себя, в основном, прямые вопросы, ответы на которые предполагают не только знания о языке, но и определенный уровень лингвистического образования. Второй вид предназначен для носителей языка, не имеющих профильного образования, и включает в себя, как правило, задания на составление словосочетаний и предложений по заданным критериям и задания на перевод. Главным преимуществом такого краудсорсинг-проекта станет возможность удаленно работать с большим количеством информантов одновременно и в обозримые сроки значительно расширить базу данных, добавив в нее новые языки.

Ключевые слова: база данных, языки мира, парадигматическое представление данных, краудсорсинг, опросник, носитель языка.

CROWDSOURCING AS A MEANS OF REPLENISHING "LANGUAGES OF THE WORLD" DATABASE

E. Makarova

Summary. The present work looks at the fourth version of the "Languages of the World" of IL RAS database. Its main distinctness is the shift from hierarchical data representation, when the features were organized in form of a binary tree, to paradigmatic. We developed a list of 124 features, each having, on average, eight possible values. For a long time the only source of information for the database was the encyclopedia of the same name. Nevertheless, as soon as all languages described in the encyclopedia are added to the database, search for alternative sources of information will become our pressing problem. The solution we suggest is to create questionnaires and conduct a crowd-sourcing project. Linguistic questionnaires are most frequently used by specialists when describing new languages. The main feature of the suggested crowdsourcing project is that not only specialists in certain languages, but also native speakers without any linguistic education will be able to take part in it. This brings us to the necessity of creating two types of questionnaires. The first type is designed for specialists and includes mainly direct questions that presuppose not only knowledge of the language, but also linguistic education. The second type is designed for native speakers who do not have a field-specific education. The second type of the questionnaire includes, as a rule, tasks to make phrases and sentences according to some given criteria and tasks to translate something. The main advantage of such crowd-sourcing project is the opportunity to work distantly with a big number of informants and to considerably replenish the database by adding new languages into it.

Keywords: database, languages of the world, paradigmatic data representation, crowdsourcing, questionnaire, native speaker.

Введение

База данных «Языки мира» ИЯз РАН — одна из немногих мировых универсальных баз данных. Среди других известных универсальных баз данных можно назвать ASJP (Wichmann et al., 2018) и WALS (Dryer & Haspelmath, 2013). Большинство же современных баз данных посвящены либо отдельным явлениям, например, типологическая база данных личных и указательных местоимений (Bliss & Ritter, 2009), PHOIBLE (крупнейшая фонологическая база данных) (Moran & McCloy, 2019), либо отдельным языкам и семьям, например, SAILS (база данных, посвященная индейским языкам Южной Америки) (Muysken et al., 2016).

Работа над базой данных «Языки мира» была начата в Институте языкознания РАН в 80-е годы (Поляков и др., 2019). С тех пор было выпущено три полных версии базы данных: для MS DOS (в 1997 году) и две версии для Windows — 2-я, рабочая версия (в 2002 году) и 3-я, ин-

формационно-справочная (в 2013 году) (Anisimov et al., 2013). В 2018 году была выпущена четвертая, демо-версия базы данных «Языки мира» ИЯз РАН[1] (Makarova & Polyakov, 2018). Как и предыдущая, 3-я версия, она включает: социолингвистические сведения о языке (статус, число носителей, код ISO, область распространения, географические координаты и так далее), генеалогический и географический указатели, глоссарий и мастер запросов, позволяющий осуществлять поиск языков, задавая условия «отсутствует/присутствует» для грамматических признаков, а также генеалогического и географического указателя. В мастере запросов отсутствуют ограничения на количество указанных фильтров.

Самым важным новшеством 4-й версии базы данных «Языки мира» ИЯз РАН стал переход от иерархической структуры данных, при которой признаки были представлены в виде бинарного дерева, к парадигматическому представлению. На основе существующего дерева признаков и данных из энциклопедического издания «Языки мира» был предложен набор из 124 признаков, каждый из признаков может иметь от 2-х до 20-ти возможных значений. При этом для языка может быть выбрано только одно значение. Таким образом, одно значение признака описывает все имеющиеся в языке явления.

Начиная с первой версии базы данных «Языки мира» ИЯз РАН единственным источником информации о языках служило одноименное энциклопедическое издание. Это связано с тем, что дерево признаков, описывающее грамматический строй языка, и текстовая часть реферата, включающая социолингвистические сведения, были разработаны на основе структуры статьи энциклопедии.

На данный момент 4-я версия базы данных «Языки мира» включает описания порядка 200 языков, и ведется активная работа по ее наполнению. В ближайшее время будут добавлены все языки из опубликованных томов энциклопедии. В связи с этим возникает проблема поиска альтернативных источников информации, и краудсорсинг-проект может стать отличным решением данной задачи.

Краудсорсинг (Егерев, 2013) — привлечение к решению тех или иных задач большого количества людей для использования их опыта или знаний. Краудсорсинг применяется в различных отраслях для сбора идей, решения задач, создания крупных электронных ресурсов и так далее. В лингвистике одним из наиболее ярких примеров краудсорсинга можно назвать работу по оценке размера словарного запаса (Keuleers et al., 2015).

Для проведения краудсорсинг-проекта, нацеленного на пополнение базы данных «Языки мира» ИЯз РАН новыми языками, предполагается создать опросники, которые позволили бы удаленно собирать информацию о языках с привлечением как специалистов, так и носителей, не имеющих лингвистического образования.

В отличие от общедоступных лингвистических опросников[2], заполнить которые представляется возможным только специалисту, мы предлагаем создание опросника, заполнить который сможет любой носитель, опираясь исключительно на собственные знания о родном языке.

Материалы и методы

Для создания 4-й версии базы данных был разработан список из 124 признаков, каждый из которых, в среднем, может принимать 8 значений. Эти признаки охватывают следующие разделы: фонемная структура, просодические явления, фонетические процессы, слог, фонологическая структура, морфологический тип языка, именные классификации, число, падежные значения, глагольные категории, дейктические категории, части речи, структура словоформы, словообразования, простое предложение, сложное предложение.

Например, признак «I-7. Способ выражения временных категорий» может принимать следующие значения:

- Аффиксы;
- Служебные слова;
- Вспомогательный глагол и аффиксы;
- Аффиксы и служебные слова;
- Вспомогательный глагол, аффиксы и служебные слова.

Каждое из возможных значений описывает все допустимые в языке явления. Пример заполнения признаков из разделов «Именные классификации» и «Число» для македонского языка представлен в Таблице 1.

Для заполнения анкет языков в данный момент используется энциклопедическое издание «Языки мира», выпускаемое Институтом языкознания РАН. Однако скорость заполнения данных о языке значительно превышает сроки выхода новых томов, таким образом, задача поиска альтернативных источников данных стоит

[1] Полная версия еще не была представлена официально. Демо-версию можно скачать по адресу: https://cloud.mail.ru/public/LpMu/gd6kKwTBV

[2] Typological tools for field linguistics. Max Planck Institute for Evolutionary Anthropology. URL: https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaires.php

Таблица 1. Пример заполнения признаков для македонского языка

Признак	Значение
F-1. Количество согласовательных классов	Три
F-3. Синтаксические способы выражения согласовательных классов	В прилагательном, числительном, местоимении и глаголе
F-5. Атрибутивное согласование по роду	Только в единственном числе
F-6. Согласование числительных в роде	Для числительного «один»
F-8. Противопоставление по личности/ не личности	Лексическое и синтаксическое
G-1. Структура числа	Единственное и множественное
G-4. Согласование по числу	Предикативное и атрибутивное
G-5. Форма имен после числительного	Единственное и множественное

особенно остро. В связи с этим было принято решение прибегнуть к методу краудсорсинга.

Для успешного краудсорсинг-проекта в рамках поставленной задачи необходимо разработать опросник, основываясь на 124 грамматических признаках, выделенных для описания языков в базе данных. Опросники давно используются лингвистами для сбора данных о языке. Существует большое количество опросников, как общих, так и посвященных отдельным разделам языка, а также виды вопросов в них.

Так как применение краудсорсинга в задаче пополнения базы данных «Языки мира» ИЯз РАН предполагает работу как с лингвистами — специалистами в конкретных языках, так и работу с носителями, возникает необходимость создания двух различных опросников, ориентированных на разных уровень лингвистической подготовки информантов.

Для лингвистов наиболее приемлемыми являются прямые вопросы. Например:

- Существует ли в языке N паукальное число?
- Совпадает ли форма неопределенного артикля с числительным «1»?
- Существует ли в языке N атрибутивное согласование прилагательных?

В то же время вопросы для носителей языка должны быть сформулированы таким образом, чтобы человек, не имеющий лингвистического образования, смог ответить на них, опираясь только на собственные знания о родном языке. Как правило, для носителей языка используются прямые вопросы и задания на составление фраз/предложений и перевод, например:

- Есть ли в вашем языке двойственное число (т.е. особая форма существительного, обозначающая, что речь идет о двух предметах?)
- Напишите 4 словосочетания «прилагательное + существительное» в единственном числе, где 2 существительных обозначают людей, и 2 существительных обозначают предмет. Поставьте эти словосочетания во множественное число.

Вне зависимости от уровня профессиональной подготовки информанта (лингвист или носитель, не имеющий профильного образования), предполагается удаленная работа с опросниками. Работа с привлечением человеческих ресурсов лингвистических отделений университетов и колледжей по всему миру позволит значительно сократить время на наполнение БД «Языки Мира» ИЯз РАН.

Для опросников, заполненных лингвистами — специалистами в области конкретных языков, будет написано программное обеспечение. Оно позволит автоматически конвертировать полученную информацию в реферат, который можно будет добавить в базу данных «Языки мира» ИЯз РАН. Что касается опросников, предназначенных для носителей, то их результаты будут подвергнуты ручной обработке лингвистами, работающими над базой данных.

Результаты

На текущий момент разработаны вопросы для следующих разделов: именные классификации, число, падежные значения, дейктические категории, части речи. Рассмотрим пример вопросов для признака «G-4. Согласование по числу». Данный признак имеет следующие возможные значения:

- Согласование по числу отсутствует,
- Предикативное,
- Атрибутивное,
- Предикативное и атрибутивное.

В опроснике, предназначенном для лингвиста, для заполнения данного признака требуется всего два вопроса:

1. Существует ли в языке N предикативное согласование по числу?
2. Существует ли в языке N атрибутивное согласование по числу?

В опроснике, предназначенном для носителя языка, предусмотрены следующие вопросы-задания:

1. Напишите 4 словосочетания «прилагательное + существительное» в единственном числе, где 2 существительных обозначают людей, и 2 существительных обозначают предмет.
2. Поставьте эти словосочетания во множественное число.
3. Подберите подходящие по смыслу глаголы для придуманных вами словосочетаний. Напишите предложения по принципу «прилагательное + существительное» из пп. 1 и 2 + глагол из п. 3 (должно получиться 8 предложений). Используйте привычный для вашего языка порядок слов.

Приведем еще пример вопросов для заполнения признака «F-9. Способ выражения одушевлённости/ неодушевленности», который может иметь следующие значения:

- Лексический,
- Морфологический,
- Синтаксический,
- Лексический и синтаксический,
- Лексический и морфологический.

В опроснике для эксперта предусмотрены для вопроса:

1. Имеет ли категория одушевленности/неодушевленности формальное выражение (различия в парадигме словоизменения)?
2. Имеет ли категория одушевленности/неодушевленности синтаксическое выражение?

В опроснике, заполнять который будет не специалист, есть следующие задания, при анализе, который также можно будет заполнить признак F9:

1. Придумайте пары существительных X и Y, относящихся к одному роду (классу, если возможно): слово X обозначает человека, а слово Y — неживой предмет. Запишите эти слова. Переведите следующие фразы:
2. Я вижу X / Y.
3. Это X / Y.
4. Это (прил.) X / Y (требуется придумать подходящее по смыслу прилагательное)
5. X/ Y (глагол) (требуется придумать подходящий по смыслу глагол)
6. В пунктах 1–5 в скобках напишите, какое местоимение использовалось бы, чтобы заменить X и Y.
7. Если в Вашем языке есть артикли: какие артикли могут употребляться при существительных X и Y?

Выводы

Главным достоинством применения краудсорсинга является возможность удаленной работы с информантами. Отсутствие необходимости личной встречи с информантами и относительная неограниченность во времени заполнения опросника позволит одновременно работать с большим количеством языков.

Очевидно, специалисты в области конкретного языка являются более предпочтительными информантами. Во-первых, в этом случае уровень достоверности ответов значительно выше. Кроме того, как было сказано выше, опросники, предназначенные для лингвистов, включают вопросы, на которые предполагается ответ «да» или «нет». Таким образом, заполненные опросники могут быть автоматически конвертированы в анкету по данному языку.

Что же касается опросников, которые будут предложены носителям языка, то их автоматическая конвертация не представляется возможной, и потребуется дополнительная ручная обработка данных. Еще одним недостатком опросников, предназначенных для носителей, является время их заполнения. Как видно из примеров, приведенных в разделе «Результаты», для заполнения одного и того же признака специалисту необходимо просто ответить «да/нет», а носителю языка требуется придумывать и записывать примеры. Кроме того, сохраняется вероятность ошибки при заполнении анкеты не специалистом в тех случаях, когда вопросы предполагают прямой ответ «да/нет». Следовательно, для получения наиболее достоверных сведений об одном языке необходимо обращаться к нескольким информантам, что значительно увеличивает время на последующую обработку данных.

Среди недостатков краудсорсинга можно также отметить языковой барьер. На данный момент опросники разрабатываются параллельно на английском и русском языках. В будущем возможно будет перевести их на некоторые другие языки. Тем не менее, важно будет искать информантов, в достаточной степени владеющих одним из языков опросника.

Заключение

В статье была рассмотрена 4-я версия базы данных «Языки мира» ИЯз РАН, главной особенностью которой стал переход от иерархической к парадигматической

форме представления данных. Долгие годы единственным источником информации для базы данных служило одноименное энциклопедическое издание, однако вскоре все описанные в нем языки будут введены, и вопрос поиска альтернативных источников встанет особенно остро.

В качестве возможного решения проблемы был предложен краудсорсинг-проект, который позволил бы удаленно собирать информацию о новых языках. Для проведения краудсорсинга разрабатываются два вида опросников: предназначенный для заполнения специалистами и предназначенный для заполнения носителями. Первый вариант является более предпочтительным и удобным и включает в себя, как правило, прямые вопросы, которые предполагают ответ «да» или «нет». Второй вариант опросника чаще всего включает задания на составление словосочетаний и предложений по заданным критериям и перевод.

Несмотря на такие недостатки как необходимость последующей ручной обработки опросников, заполненных носителями, а так же задачу поиска информантов, в достаточной степени владеющих русским или английским языками, успешное выполнение краудсорсинг-проекта позволит удаленно работать с большим количеством информантов одновременно и в обозримые сроки значительно расширить базу данных, добавив в нее новые языки.

Благодарность

Исследование было поддержано грантом РФФИ № 19–012–00476 А.

ЛИТЕРАТУРА

1. Егерев С. В., Захарова С. А. (2013) Краудсорсинг в науке // Альманах «Наука. Инновации. Образование» / Российский научно-исследовательский ин-т экономики, политики и права в научно-технической сфере (РИЭПП). — Языки славянской культуры, 2013. — № 14. — С. 175–186. — ISSN1996–9953
2. Поляков В. Н., Соловьев В. Д., Макарова Е. А. (2019) База данных «Языки Мира»: история и перспективы. — Москва, Казань: Институт языкознания Российской Академии Наук, Казанский (Приволжский) Федеральный Университет. 370 стр.
3. Anisimov, Ivan, Polyakov, Vladimir & Solovyev, Valery. (2013) Database "Languages of the World". New Version. New Research Horizons. Collection of Papers of the First International Forum on Cognitive Modeling (14–21 September, 2013, Italy, MilanoMarittima). Part 1. P. 27–34. ISBN918–5–87872–731–0
4. Bliss H., Ritter E. (2009) A typological database of personal and demonstrative pronouns. In "The Use of Databases in Cross-Linguistic Studies". Everaert M., Musgrave S., Dimitriadis A. (eds.). Berlin: Mouton de Gruyter. Pp. 77–116.
5. Dryer, Matthew S. & Haspelmath, Martin (eds.). (2013) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.
6. Keuleers et al. (2015) «Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment». Quarterly Journal of Experimental Psychology. 68 (8): 1665–1692. Doi:10.1080/17470218.2015.1022560
7. Makarova Elena, Polyakov Vladimir. (2018) «Languages of the World» database: Paradigmatic representation // Cognitive Modeling: Proceedings of the Fourth International Forum on Cognitive Modeling (30 September — 7 October, 2018, Tel Aviv, Israel). In 2 parts. / Edited by S. Masalóva, V. Polyakov, V. Solovyev. — Rostov-on-Don: Science and Studies Foundation. Pp. 116–122. ISBN978–5–907125–21–6
8. Moran, Steven & McCloy, Daniel (eds.). (2019) PHOIBLE2.0. Jena: Max Planck Institute for the Science of Human History. (Available online at http://phoible.org, Accessed on 2019–10–28.)
9. Muysken, Pieter, Harald Hammarström, Olga Krasnoukhova, Neele Müller, Joshua Birchall, Simon van de Kerke, Loretta O'Connor, Swintha Danielsen, Rik van Gijn & George Saad. (2016) South American Indigenous Language Structures (SAILS) Online. Leipzig: Online Publication of the Max Planck Institute for Evolutionary Anthropology.
10. Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (eds.) (2018) The ASJP Database (version 18).

Журнал «Современная наука: актуальные проблемы теории и практики»