SUMMA TYPOLOGIAE: БАЗА ДАННЫХ «ЯЗЫКИ МИРА» И ДИЛЕММА ФОРМАЛИЗОВАННОГО ОПИСАНИЯ ЯЗЫКОВ

Д.И. Коломацкий
Институт языкознания РАН
dk@iling-ran.ru

Perhaps the most important and underappreciated class of linguistic generalizations that are used as data are the (usually implicit) claims that specific languages exist in the first place (Good 2022: 39–40)

Введение

Базу данных «Языки мира» можно смело назвать одним из старейших цифровых проектов Института языкознания. Трудный и разветвлённый путь её развития высвечивает целый ряд интересных проблем — от технических аспектов создания лингвистических баз данных до фундаментального вопроса об общей возможности создания всеобъемлющего унифицированного описания языков.

1. Исторический обзор

Планы по созданию печатного энциклопедического издания о языках мира появились уже в середине 1970-х гг. (Соловьёв, Кибрик 2015), было создано несколько десятков статей. Не позже середины 1980-х гг. началась разработка цифрового продукта, который должен был сопровождать энциклопедию[1] и даже заменять её: в частности, для англоязычных исследователей планировался машинный перевод описаний языков (Журинская и др. 1986). Формализованное описание одного языка или диалекта было названо *рефератом*, была

[1] База данных «Языки мира» не могла бы существовать без отредактированных и выверенных описаний языков — то есть без печатной энциклопедии. Тем, что в 1993 году всё же вышел первый том «Уральские языки», а за ним последовало ещё более двух десятков томов, научное сообщество, несомненно, обязано чествуемому юбиляру.

DOI 10.37892/978-5-6049527-2-6-65

разработана так называемая *модель реферата* (Журинская и др. 1986; Ярославцева 2001), представлявшая собой дерево признаков.

После М.А. Журинской, А.И. Новикова и Е.И. Ярославцевой разработкой базы данных занимался Ю.П. Скокан, затем проектом руководил В.Н. Поляков. К 2013 году было выпущено три версии, в последней из них было 315 рефератов языков и диалектов[2] (Поляков и др. 2019). Исследованиям, проводившимся на этих данных, посвящён целый ряд работ (Belyaev 2009; Polyakov et al. 2009; Solovyev, Polyakov 2013; Виноградов и др. 2003; Соловьёв, Кибрик 2015).

В 2019 году была предпринята попытка отхода от древовидной структуры признаков. Осенью 2020 г. на основе нового перечня признаков была выпущена 4-я десктопная версия базы данных, разработанная Е.А. Макаровой. Создание описаний языков началось заново, и к концу 2022 года по новой модели было заполнено 338 рефератов.

Хотя база данных «Языки мира» официально является проектом отдела прикладной лингвистики, её разработка была бы невозможна без тесного взаимодействия с сектором ареальной лингвистики (ранее — группа «Языки мира»), возглавляемым А.А. Кибриком. С конца 2022 г. в состав группы, которая на постоянной основе будет работать над улучшением базы данных, входят сотрудники и других подразделений Института, в том числе редакторы печатных томов энциклопедии и авторы энциклопедических статей[3]. Сам процесс работы над базой данных в новой итерации даёт толчок для новых исследований (Зотова и др. 2022)[4].

2. Особенности онлайн-версии 2023 г.

Практически все опции готовящейся к выпуску в 2023 году онлайн-версии в том или ином виде уже были реализованы в предыдущих версиях. В разные моменты прошлого существовали и вебсайт, и доступ к отсканированным версиям энциклопедических статей, и поиск по комбинациям различных параметров и географическим координатам (основанная на этих коорди-

[2] Наиболее подходящим термином для называния сущности, которую в формализованном виде описывает реферат, представляется *докулект* (Good, Cysouw 2013).

[3] На настоящий момент состав коллектива таков (в алфавитном порядке): О.И. Беляев, В.Ю. Гусев, В.В. Дьячков, А.А. Евдокимова, А.К. Зотова, Д.И. Коломацкий, Ю.Б. Коряков, Ю.В. Мазурова, Т.А. Майсак, Е.А. Макарова, О.И. Романова, Н.К. Рябцева. В заседаниях группы также участвует специалист по компьютерной лингвистике Т.О. Шаврина.

[4] Любопытно в этой связи упомянуть аналогичную по тематике работу Lesage et al. 2022 на материале типологической базы данных Grambank.

натах[5] интерактивная карта, реализованная в новой версии, весьма условна, поскольку каждому языку соответствует одна пара координат — то есть одна точка[6]). Тем не менее, на этот раз все полезные опции собраны в одной версии, а удобству пользовательского интерфейса уделено особое внимание. Визуализации данных создаются интуитивно понятными и максимально наглядными.

Что касается самой базы данных как набора связанных отношениями таблиц, то в стремлении соответствовать принципам FAIR в целом (Nordhoff 2021) и цифрового грамматического описания в частности (Nordhoff 2008; Nordhoff, Hammarström 2014) все данные были преобразованы в текстовые форматы CSV и YAML и размещены онлайн. Из них автоматически формируется датасет стандарта CLDF (Forkel et al. 2018).

Основной технической проблемой стало хранение и представление признаков, содержащих «элементарные» значения в разных комбинациях (например, атрибутивное согласование может быть: по роду; по числу; по определённости; по роду и числу, по роду, числу и определённости и т. д.). Есть основания утверждать, что эта проблема возникла на самых ранних этапах создания базы данных в 1980-х гг. и не была должным образом решена[7]. Однако больший интерес для нашего обсуждения представляет концептуальный вопрос выбора общей стратегии формализованного описания языков.

3. Дилемма формализованного описания языков

Кратко проблему можно сформулировать так: авторы энциклопедических статей не только используют собственную терминологию[8], но и при-

[5] Часть координат, которых не было в предыдущих версиях, была взята из базы данных Glottolog (Hammarström et al. 2022).

[6] Это выглядит особенно странно для самых крупных языков. Однако идеального решения проблемы не существует: даже крайне трудоёмкие в создании карты ареалов не обеспечивают полного и точного представления информации о языках (Drude 2018). Поэтому такие системы, как Glottolog и WALS, продолжают обозначать языки точками на карте. Задание координат этих точек — отдельная сложная задача (Loginova 2009). См. также комментарии о карте языков как продукте имплицитных обобщений о языках в Good 2022.

[7] Актуальность проблемы стала особенно заметной именно после появления интерактивной карты: оказалось, что иногда на ней нужно отобразить несколько десятков разных маркеров. С похожим вызовом ранее столкнулись создатели атласа WALS, и выбранное ими решение (ввести значение “Mixed” или “Other”, чтобы ограничить число маркеров) оказалось, по их собственному признанию, не оптимальным (Haspelmath 2009).

[8] См. пример с термином *топик* в работе Соловьёв, Кибрик 2015, а также работу Зотова, Романова 2021, полностью посвящённую терминологии в базе данных «Язы-

держиваются своих традиций в классификации явлений — разные языки (особенно принадлежащие к разным языковым семьям) временами описаны будто бы в разных измерениях. При заполнении реферата специалист вынужден оставлять пропуски или «читать между строк». В итоге, как справедливо отмечает Е.И. Ярославцева (2001: 356), субъективный взгляд автора энциклопедической статьи на описываемый язык дополнительно преломляется субъективным взглядом референта[9].

По-видимому, изначально создатели базы данных планировали унификацию терминологии: указания на это можно найти в монографии Журинская и др. 1986. Спустя десятилетия подход изменился в пользу сохранения авторских терминов (Polyakov et al. 2009), что неизбежно привело к появлению фактически эквивалентных значений признаков.

Полная унификация терминологии и классификаций неизбежно приведёт к потерям, которые могут оказаться слишком серьёзными. Отказ же от неё будет означать, что база данных «Языки мира» не сможет именоваться типологической базой данных. Но стоит ли претендовать на создание таковой в условиях, когда, во-первых, само печатное издание не охватывает все языковые ареалы, а во-вторых, цифровой продукт в любом случае будет использоваться лишь как отправная точка для типологического исследования[10]? С другой стороны, сохранение в неизменном виде авторской терминологии и принципов описания языков может привести к неконтролируемому росту числа значений признаков и полной потере информативности базы данных[11].

Промежуточный подход может заключаться в умеренной унификации[12] и создании синонимических рядов терминов, что в перспективе превращается

ки мира». П.М. Аркадьев в личном сообщении первым указал автору на тот факт, что даже в рамках одного тома используемая в статьях терминология может быть непоследовательной.

[9] См. также замечания о «цепочках» обобщений данных о языках в Good 2022.

[10] О том, что подобная база данных может служить для исследователя только отправной точкой, в личной коммуникации с автором сообщали О.И. Беляев, Д.О. Жорник, Н.В. Сердобольская и другие коллеги. Существует типологическая база данных Grambank, создатели которой проводят квантитативные исследования непосредственно на её материале, но она основана почти на четырёх тысячах грамматических описаний почти двух с половиной тысяч языков, причём все её 195 признаков строго тернарные («да», «нет», «неясно») и ограничиваются только морфосинтаксисом; её наполнением в течение многих лет занимаются десятки специалистов (Lesage et al. 2022; Skirgård et al. 2022).

[11] См. в этой связи интересные рассуждения об излишней унификации и «профилях языка» в Virk et al. 2020.

[12] Например, скорее очевидно, **что посессивные прилагательные** и **притяжательные прилагательные** описывают одно и то же явление. Возможно, то же самое можно сказать и о терминах **род** и **именной класс**. В личной беседе с автором

в разработку их онтологии. Если строго ограничить её область терминами, встречающимися в энциклопедии, эту задачу можно считать выполнимой, хотя и весьма трудоёмкой [13].

Заключение

База данных «Языки мира» может стать своего рода «путеводителем» по энциклопедии, который, с одной стороны, демонстрирует все её достоинства, а с другой — не скрывает несовершенств печатного издания и объективной неполноты данных о языках, тем самым побуждая исследователя к новым самостоятельным изысканиям.

Благодаря простоте и наглядности веб-интерфейса база данных способна выполнять и популяризаторскую функцию, знакомя более далёкого от лингвистики пользователя не только с языковым разнообразием, но и с различными лингвистическими традициями, которые, несмотря на все усилия редакторов, нашли своё выражение в энциклопедии. Это может быть её слабостью — и одновременно её силой.

Литература

Виноградов В.А., Новиков А.И., Ярославцева Е.И. 2003. База данных «Языки мира» как инструмент лингвистического исследования. *Вопросы языкознания,* №3, 3–14.

Журинская М.А., Новиков А.И., Ярославцева Е.И. 1986. *Энциклопедическое описание языков: Теоретические и прикладные аспекты*. М.: Наука.

Зотова А.К., Коломацкий Д.И., Романова О.И. 2022. Значимое отсутствие: лакуны в описании языков (на материале базы данных «Языки мира» ИЯз РАН). *Лингвистика и методика преподавания иностранных языков,* № 1(16), 20–38.

Зотова А.К., Романова О.И. 2021. «Свой» среди «чужих», или Потенциал термина в системе (на материале терминологии новой версии базы данных «Языки мира»

Д.И. Эдельман указала на то, что, например, принятый в иранистике термин *аорист* (означающий в этой традиции настояще-будущее время сослагательного наклонения) не нужно механически включать в реферат: благодаря усилиям В.С. Расторгуевой, статьи посвящённых иранским языкам томов обычно содержат пояснения терминов, из которых вдумчивый лингвист вполне может сделать верный вывод относительно того, какое значение проставить для признака в реферате, не полагаясь исключительно на использованный термин.

[13] Опыт создания всеобъемлющих онтологий показал, что это практически невозможно осуществить на практике. См. напр. Langendoen 2020 об известной онтологии GOLD. Интересный «экуменический» подход к типологическим базам данных в целом и онтологиям в частности демонстрирует проект TDS (Dimitriadis et al. 2009; Windhouwer et al. 2017).

ИЯз РАН). *Лингвистика и методика преподавания иностранных языков,* № 1(14), 27–48.

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

Соловьёв В.Д., Кибрик А.А. 2015. Чем компьютерные технологии могут помочь лингвистической типологии? *Вестник Российской академии наук,* № 1(85), 32–38.

Ярославцева Е.И. 2001. Грамматикон и база данных «Языки мира». В кн.: А.И. Новиков (ред.). *Scripta linguisticae applicatae. Проблемы прикладной лингвистики — 2001*. М.: Азбуковник, 339–357.

Belyaev O. 2009. Temporal stability of features in Jazyki Mira. A talk presented at The Swadesh Centenary Conference at Max Planck Institute for Evolutionary Anthropology, Leipzig. https://www.eva.mpg.de/lingua/conference/09_SwadeshCentenary/files/program.html (дата последнего доступа 13.03.2023)

Dimitriadis A., Windhouwer M., Saulwick A., Goedemans R., Bíró T. 2009. How to integrate databases without starting a typology war: The Typological Database System. In: M. Everaert, S. Musgrave, A. Dimitriadis (eds.). *The Use of Databases in Cross-Linguistic Studies*. Empirical Approaches to Language Typology [EALT]. Berlin, New York: Mouton de Gruyter, 155–207.

Drude S. 2018. Why we need better language maps, and what they could look like. In: S. Drude, N. Ostler, M. Moser (eds.). *Endangered Languages and the Land: Mapping Landscapes of Multilingualism*. Presented at the FEL XXII/2018 (Reykjavík, Iceland), London: FEL & EL Publishing, 33–40.

Forkel R., List J.-M., Greenhill S.J., Rzymski C., Bank S., Cysouw M., Hammarström H., Haspelmath M., Kaiping G.A., Gray R.D. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data,* n. 1(5).

Good J. 2022. The Scope of Linguistic Data. In: A.L. Berez-Kroeker, B. McDonnell, E. Koller, L.B. Collister (eds.). *The Open Handbook of Linguistic Data Management*. Cambridge, MA: The MIT Press, 27–48.

Good J., Cysouw M. 2013. Languoid, Doculect and Glossonym: Formalizing the Notion ‘Language.’ *Language Documentation & Conservation,* v. 7, 331–359.

Hammarström H., Forkel R., Haspelmath M., Bank S. 2022. *glottolog/glottolog: Glottolog database 4.7*.

Haspelmath M. 2009. The typological database of the World Atlas of Language Structures. In: M. Everaert, S. Musgrave, A. Dimitriadis (eds.). *The Use of Databases in Cross-Linguistic Studies*. Empirical Approaches to Language Typology [EALT]. Berlin, New York: Mouton de Gruyter, 283–299.

Langendoen D.T. 2020. Whither GOLD? In: A. Pareja-Lora, M. Blume, B.C. Lust, C. Chiarcos (eds.). *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. Cambridge, MA: The MIT Press, 19–24.

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

Loginova E. 2009. Technique of definition of geographical coordinates at the identification of the area of distribution of language (on the material of DB "Jaziki Mira"). *Text Processing and Cognitive Technologies*, v. 17.

Nordhoff S. 2008. Electronic Reference Grammars for Typology: Challenges and Solutions. *Language Documentation and Conservation*, v. 2, n. 2, 296–324.

Nordhoff S. 2021. Open publication of texts, open publication of data. Talk presented at Emerging Topics in Typology. International Workshop Series 25 October — 22 November 2021. https://github.com/langsci/lsp-presentations/blob/master/2021emergingtopics/emergingtopics.pdf (дата последнего доступа 13.03.2023)

Nordhoff S., Hammarström H. 2014. Archiving grammatical descriptions. *Language Documentation and Description*: Special Issue on Language Documentation and Archiving, v. 12, 164–186.

Polyakov V.N., Solovyev V.D., Wichmann S., Belyaev O. 2009. Using WALS and Jazyki mira. *Linguistic Typology,* v. 13, n. 1, 137–167.

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. 2022. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss (preprint). SocArXiv.

Solovyev V.D., Polyakov V.N. 2013. Database "Languages of the World" and its application. State of the art. In: *Computational Linguistics and Intellectual Technologies*. Papers from the Annual International Conference "Dialogue" (2013). М.: Издательство РГГУ, 748–758.

Virk S.M., Hammarström H., Borin L., Forsberg M., Wichmann S. 2020. From Linguistic Descriptions to Language Profiles. In: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*. Presented at the Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020. Marseille, 23–27.

Windhouwer, M., Dimitriadis, A., Akerman, V., 2017. Curating the Typological Database System. In: J. Odijk, A. van Hessen (eds.). *CLARIN in the Low Countries*. London: Ubiquity Press, 123–132.