О ДАННЫХ В ЛИНГВИСТИКЕ

А.Б. Шлуинский
Гамбургский университет
ashl@yandex.ru

Мне довелось многие и многие часы обсуждать с Андреем Александровичем, как устроена лингвистическая наука и как она должна бы быть устроена. Но о многом поговорить не хватало времени. Пользуясь известной вольницей жанра юбилейного сборника, я рад предложить здесь в свободной форме (в частности, сознательно отказавшись, в силу малого объема, от какой-либо библиографии) еще одну реплику в этой дискуссии. Возможно, Андрею Александровичу будут интересны мои соображения.

Исследовательская работа в эмпирической науке (какой, конечно, является и лингвистика) основана на **данных**. Данные, если огрубить, бывают двух типов — назовем их **коллекции** и **протоколы**.

Коллекции — это массивы естественных данных, которые исследователь собирает и далее тем или иным способом препарирует и анализирует. К такому типу данных можно отнести предметные этнографические коллекции или, скажем, коллекции-гербарии у ботаников. К нему же можно отнести и письменные памятники, анализом которых занимаются филологические традиции. В лингвистике же эту нишу составляют массивы естественных текстов — записанных, транскрибированных и аннотированных (в современных условиях, конечно, в электронной форме). Здесь, в принципе, возникает вопрос о том, где граница естественного: все-таки, в большинстве случаев то, что удается записать лингвисту, — не образцы «совсем настоящей» коммуникации, которая происходит сама собой, а нечто, что порождается по запросу исследователя или, по крайней мере, с учетом его присутствия и самого факта записи. Но собирая коллекцию текстов, лингвист стремится получить тот максимум естественного, какой окажется ему доступен.

Протоколами я условно назвал типы данных, которые исследователь получает, намеренно воздействуя на внешнюю среду, с тем чтобы получить именно то, что ему нужно, — то есть искусственные данные. Как правило, в естественных науках эту нишу занимает эксперимент — то есть получение данных в специальных условиях, фиксирующих все переменные, кроме тестируемой. Эксперименты *протоколируются*. В лингвистике эксперимент в строгом понимании термина не стал (и для многих задач в обозримом бу-

DOI 10.37892/978-5-6049527-2-6-63

дущем не станет) центральным способом получения данных, хотя и расширяет сферу своего влияния. Однако искусственные данные в лингвистике активно добываются при помощи элицитации (то есть запроса лингвиста к носителю исследуемого языка дать переводной эквивалент для искомого значения или оценить приемлемость определенной конструкции) или интроспекции (то есть, по сути, аналогичного самозапроса лингвиста к самому себе как носителю); но поскольку регламентирование интроспекции всё же пока сложно представить, будем говорить об элицитации. В противоположность сбору текстов (по сути, *каких-нибудь* текстов искомых жанров), в случае элицитации лингвист концентрируется на том, чтобы получить искомые конструкции.

Классификация типов данных может и должна уточняться, и тем более в рамках этих заметок я не вдаюсь в собственно методологические аспекты их получения — как записывать и транскрибировать тексты или как спрашивать у информанта, можно ли сказать так-то, и реагировать на суждения типа «можно, но что-то не очень хорошо звучит». И методология полевой лингвистики или языковой документации, и методология лингвистического эксперимента весьма активно обсуждаются, тогда как собственно культура обращения с данными получила в лингвистике минимальное внимание. Оба типа данных, и коллекции, и протоколы, в очень большой мере остаются внутренней кухней исследователя, которая как бы не касается никого, кроме его одного; и тут уже от личных качеств исследователя зависит, порядок у него или хаос. Более того, каждый сам решает, заслуживают ли вообще материалы хранения, если проведенное на них исследование уже опубликовано. После смерти исследователя его архив может быть выброшен, а может быть и передан коллегам; но в последнем случае выясняется, что разобраться в данных, которые исследователь структурировал лично для себя, становится отдельной кропотливой задачей. Про примеры, фигурирующие в лингвистических работах, неизвестно, *откуда что взялось*, в том числе и в особенности про примеры, подвергшиеся вторичному цитированию. Становление документации как самостоятельного направления, конечно, привело к осознанию значения метаданных, но вплоть до настоящего времени обычное дело, что в статье, рассматривающей определенный круг явлений в таком-то языке, вообще не сказано, откуда взят материал, или «информативно» сообщается, что «примеры получены от информантов — носителей языка», в лучшем случае — с поименной благодарностью или с указанием на населенный пункт, в котором данные были собраны. Эта практика делает лингвистику едва ли не уникальной (не в хорошем смысле) среди эмпирических дисциплин, от самых образцово-гуманитарных до самых образцово-естественных.

Преодоление сложившейся ситуации, естественно, не делается в один момент. В идеале нужно получить общепризнанную международную конвен-

циональную систему практик, определенным образом регламентирующую описание и коллекций, и протоколов, такую чтобы любой специалист мог без затруднений работать с данными другого специалиста (естественно, вынося за скобки и компетенцию в языке-объекте, и владение языком-посредником). Едва ли не важнейшим качеством этой системы должна быть простота аннотирования: чем больше сил и времени будет требовать от лингвиста описание каждого конкретного фрагмента данных, тем больше вероятность, что эта работа будет откладываться на потом и в конце концов не делаться вовсе. (В то же время стандартизация описания должна способствовать и его лаконичности.) Но на начальных этапах можно и нужно обсуждать хотя бы базовые принципы организации данных.

Особо надо сделать акцент на том, что применительно к устному модусу (единственному или доминирующему у большинства языков мира) первичной формой данных надо осознать аудиозапись. В нынешних условиях, когда хотя бы любительская техника аудиозаписи доступна каждому, фиксация данных только в письменной форме должна быть признана исключительной. Таким образом, с точки зрения модуса, данные могут быть трех основных типов: (1) письменный текст; (2) устная речь; (3) старые архивные или опубликованные материалы, фиксирующие устную речь в графической форме. Другие типы возможны (как, например, элицитация в мессенджерах в графической форме), но редки. Сосредоточимся на данных второго типа: с одной стороны, именно с ними ситуация особенно тяжелая, а с другой, именно с ними и возможно изменение практики. Если для уже опубликованных текстов, изначально записанных в устной форме, остается неизвестным, что же на самом деле было произнесено, и приходится с ними работать как они есть (а редкая возможность сопоставить публикацию с архивной аудиозаписью может дать весьма неожиданные результаты), то практика работы с новыми данными может и должна быть изменена. Соответственно, будем исходить из ситуации, когда для элементов и коллекции, и протокола, существует аудиозапись. Более того, именно аудиозапись собственно и представляет собой элемент лингвистических данных в строгом смысле слова, тогда как транскрипция любого уровня — уже их препарирование. Пока дополнение опубликованной статьи доступными аудиоматериалами остается экзотическим решением (в основном для работ, специализирующихся на звуковой стороне языка), но уже легко представить себе, как pdf-файл, включающий в себя аудиозаписи всех приводимых примеров, станет нормой для лингвистической статьи.

В области коллекций текстов и идея общедоступной аудиозаписи, и идея необходимости ссылки на конкретный пример, в принципе, уже носятся в воздухе, и задача состоит в том, чтобы метаданные и аннотации приобрели наконец какой-либо общепринятый формат. Наиболее сложной проблемой

для коллекций текстов представляется редактирование на этапе транскрибирования. Как правило, аудиозапись расшифровывается с информантом — носителем языка, который часто предлагает правку, уменьшающую речевые сбои, оговорки, а порою и заменяющую неразборчивые места на альтернативные фрагменты. Стремление к максимально точному транскрибированию устной речи (принципам которого Андрей Александрович уделил немало сил) вступает в противоречие как с необходимостью транскрибировать сколько-нибудь представительные объемы данных в короткий срок, так и с тем, что информант-неспециалист редко готов принять установку на дотошное воспроизведение сказанного в ущерб внятности результирующего текста. Более того, когда примеры из коллекции текстов используются для иллюстрации грамматических явлений, точность транскрипции устной речи оказывается усложняющим фактором для собственно лингвистических целей: куда как удобнее, если пример на управление определенного глагола содержит только его актанты и ничего больше, а не дополняется несколькими фальстартами, плейсхолдерами и оговорками. Таким образом, здесь есть запрос на стандартизацию практичного соотношения между идеальной точностью и репрезентативным удобством.

В области же данных-протоколов дело обстоит гораздо хуже: элицитированные языковые примеры, приведенные лингвистом в опубликованной работе, фактически занимают нишу исходных, а не вторичных данных для всех других специалистов, кроме автора (крайне редко — небольшой рабочей группы). Ни узнать, какие еще примеры были получены, ни послушать звучание примера, ни даже уточнить, идет ли речь о примере, полученном в результате перевода с языка-посредника или сконструированном лингвистом, возможности нет.

Опыт полевой работы с носителями разных во всех отношениях языков убедил меня в том, что записи в полевой тетради или ее аналоге должны рассматриваться именно как *протокол* сессии элицитации. Сессия работы фиксируется как она есть в своей фактической последовательности. Естественно, что выбор между традиционной бумажной тетрадью и компьютерным файлом уже давно стал делом вкуса, причем тетрадь смотрится как выбор ретроградов, который неизбежно уйдет; но именно тетрадь технически обязывает исследователя сохранять последовательную фиксацию сессии с минимальными дополняющими комментариями, тогда как компьютерный файл склоняет к тому, чтобы сразу преобразовывать его в аналитический. Между массивом аудиозаписей сессий элицитации и их протоколами должно быть создано однозначное соответствие, причем чем проще оно устроено, тем удобнее дальнейшая работа с данными (например, сквозная нумерация гораздо проще в обращении, чем новая нумерация внутри каждой даты работы). Сами аудио-

записи должны быть как можно более краткими (не более 3 минут), что может позволить в буквальном смысле за секунды найти как фрагмент, соответствующий нужному примеру, так и комментарии информанта к нему. Понятно, что такая система сама по себе не делает работу с данными для внешнего человека простой (всё же речь о рабочих материалах), но делает ее принципиально возможной.

Создание унифицированной системы нотации для данных обоих типов, естественно соотносящейся с метаданными, будучи, казалось бы, очень частным вопросом, станет методологическим прорывом в лингвистике, повышающим эмпирическое качество работы.