ОБЗОРЫ, СООБЩЕНИЯ, ЗАМЕТКИ

М.К. Григорьева

О технической стороне оцифровки «Большого китайско-русского словаря» под редакцией И.М. Ошанина. Воспоминания участника

От редакции

40-летие со времени выхода фундаментального Большого китайско-русского словаря (БКРС) под редакцией И.М. Ошанина заставляет нас обратиться не только к истории создания этого труда, но и к его трансформации в век информационных технологий. Оцифровка словаря имеет непреходящее значение, сейчас им пользуются сотни и тысячи китаистов нашего времени, а ведь еще недавно он существовал только в бумажном виде. Воспоминания Марии Григорьевой, одной из первых включившейся в эту нелегкую работу, содержат уникальную информацию начала 2000-х годов, когда нашлись десятки энтузиастов, поверивших в возможность появления в русскоязычном пространстве электронного словаря, работа над которым длилась с 2003 по 2006 г. Сегодня цифровой вариант БКРС стал современным инструментом, функционирующем на мировом уровне и, вероятно, превзошел своих авторов — советских ученых-китаистов.

Журнал «Российское китаеведение» • 3(4)2023 • С. 150—155 • © Григорьева М.К.

* * *

Я очень завидую тем, кто сегодня начинает учить иностранный язык, особенно такой сложный, как китайский. Теперь есть возможность скачать или купить любые учебники, в Интернете можно пообщаться с носителями языка в чатах или социальных сетях, смотреть фильмы и слушать песни, имеется много языковых курсов и школ. Современные технологии оказывают большую помощь в обучении: любой смартфон становится платформой для совершенствования произношения, чтения или письма. Раньше таких разнообразных ресурсов не было; когда мы 20 лет назад начинали изучать китайский язык, возможности студентов были очень ограничены. Но мы уже вступали в эпоху цифровизации, компьютеризации, социальных сетей, что дало возможность задумать и реализовать проект электронного китайско-русского словаря.

Получилось так, что будущие участники проекта в одно время оказались на интернет-форуме «Восточное полушарие», объединившем людей, которые интересовались языками, культурой и современной жизнью Китая, Японии, Кореи и других стран Азии. Именно там было положено начало работы над электронным словарем. Я попала на «Восточное полушарие» случайно: после двух лет учебы в Пекине мне было интересно продолжить обучение китайскому языку в России, познакомиться с новыми людьми. Один из участников «Полушария», Олег Панков[1], опубликовал на форуме объявление о том, что хочет отсканировать и провести электронное распознавание Большого китайско-русского словаря. Этот пост привлек мое внимание. К слову, на тот момент я ничего не знала о четырехтомном словаре, хотя училась уже несколько лет языку. Четырехтомник под редакцией И.М. Ошанина не был популярен среди молодежи и считался немного устаревшим.

Я написала Панкову, что готова присоединиться к проекту. Хочется заметить, что на тот момент в нашем распоряжении имелись только бумажные словари. И в России, и даже в Китае. Хотя в Китае, кажется, уже начали делать небольшие электронные устройства со словарями, но они были дорогие и в России еще не продавались. Печатными словарями было не всегда удобно пользоваться, но быстро развивались технологии, и уже появились компьютерные программы, такие как Lingvo, со словарями для европейских языков. Поэтому и возникла мысль сделать электронный словарь китайского языка. Чем китаисты хуже? Никто, правда, не предполагал, что работа будет довольно сложной и долгой.

Подчеркну, что начали мы не с нуля. На тот момент, когда я пришла в команду оцифровщиков словаря, там уже было несколько человек. И один из них, Алексей Дьяков, системный администратор из Санкт-Петербурга уже сделал электронную таблицу, в которую занес все иероглифы БКРС по их номерному знаку. То есть мы понимали, какие иероглифы уже имеются в системе Юникод, а каких еще нет. За 20 лет, которые прошли с тех пор, кодировка Юникод обновлялась несколько раз и сейчас в ней есть почти все знаки китайской письменно-

[1] Олег Панков оцифровал (т.е. организовал) электронную базу БКРС. Создатель интернет-сайта «Дао, выраженное словами» (daokedao), Telegram-канала «Китайская угроза» (URL: https://t.me/daokedao).

сти, а на тот момент очень многие иероглифы просто невозможно было набрать. Но так как в основном это были редкие знаки, мы не стали сильно беспокоиться и решили заменять их временными знаками.

Меня однажды заинтересовал вопрос: почему Алексей начал составлять эту таблицу задолго до начала нашего проекта. Он ответил, что просто ему был интересен Китай...

Так мы стартовали и появилась основа для работы. Оценив объем статей и имея таблицу А. Дьякова, участника проекта, были уверены, что мечта о цифровизации БКРС вполне осуществима.

Для оцифровки первым делом необходимо было полностью отсканировать словарь. Для этого пришлось разобрать все тома на странички, чтобы было удобнее сканировать. На тот момент технические характеристики сканеров были хуже, чем сейчас. Поэтому мы старались «выжать» максимальное качество из того уровня техники, которым располагали. Ведь от качества сканирования зависело качество оцифровки. Какие-то процессы по распознаванию текстов планировали делать автоматически с помощью программ Adobe Finereader OCR. Частично мы реализовали эту идею, и здесь хорошее качество сканов было нелишним.

Для распознавания пробных отсканированных страниц попробовали программу OCR. Эта программа неплохо обрабатывала русский текст, но совсем не видела китайские иероглифы и фонетическую транскрипцию. Мы решили, что так и будем работать: автоматически распознавать русский текст, затем вручную набирать иероглифы и пиньинь.

Необходимо было не только набрать текст БКРС на компьютере, но и сделать из этого текстового файла электронный словарь, хотя бы в формате «Lingvo», подключить его к программе, чтобы заработал поиск и была возможность пользоваться словарем на компьютере. На тот момент еще не было планшетов или мобильных телефонов, поэтому мы подразумевали установку словаря на персональные компьютеры. Для форматирования словарных статей под Lingvo нужны были определенные стандарты, которые мы стали сами формулировать и создавать.

Было ясно, что работа над словарем потребует участия большого коллектива, поэтому помимо составления технического задания, мы стали публиковать новости проекта на форуме «Восточное полушарие». В результате к нам стали присоединяться все новые и новые участники. Работу с ними и раздачу материалов для оцифровки поручили мне. Благодаря этому за несколько лет я познакомилась с очень многими людьми.

Как была организована работа? У меня на руках были отсканированные странички БКРС, распознанные программой OCR файлы и правила форматирования. Такой комплект материалов я отсылала каждому участнику. В случае с новичками обычно мы начинали с пробных одной-двух страничек, чтобы человек мог освоиться с работой. Необходимо было вычистить «мусор» из распознанного текста, а его было много: нераспознанные иероглифы, пиньинь — все превращалось в случайные знаки, которые нужно было удалять. Затем все иероглифы набирались вручную; было решено, что мы не будем упрощать оригинал и станем «вбивать» как есть, традиционными китайскими иероглифами. Следующий

этап — добавление транскрипции пиньинь. Сначала предполагалось дать всем задание прописывать пиньинь «красиво», вводя буквы уже с черточками тонов. На практике оказалось, что это сложно, и в правилах мы указали, что проставляем тона цифрами.

Набранный текст участник должен был отформатировать по заданному стандарту. Необходимо было максимально упрощать наши правила, потому что сложные правила могли отпугнуть наших волонтеров. С другой стороны, регламент был составлен так, что позволял использовать набранный текст для дальнейшей работы по форматированию в программе Lingvo.

На выходе участники присылали мне готовые странички примерно вот в таком виде:

一點 yi1dian3	
1) легонько тронуть, едва коснуться; при [одном] прикосновении...	2) час (напр. пополудни; также — 一點鐘)
3) один пункт; одно, нечто, кое-что 在這一點 в этом отношении (пункте)	4) немного, чуть-чуть; легонько; перед отрицанием ничуть, ни капли, нисколько. 一點兒也不錯 без малейшей ошибки, совершенно правильно

На втором этапе эти страницы я собирала и отправляла Олегу Панкову, который уже с помощью встроенных в программу Word макросов добавлял специальные теги, дополнял текст упрощенными иероглифами (мы решили, что словарные статьи будут в двух вариантах: в традиционных и упрощенных иероглифах, а примеры только в традиционных, как в БКРС), переделывал пиньинь в привычный всем вид (например, yidian). И на выходе мы имели файлы в формате DSL, потом LSD[1], которые уже можно было загружать в словарь Lingvo и использовать их.

Промежуточные итоги проекта мы выкладывали сначала на «Восточном полушарии», потом на другой интернет-площадке «Профессиональном востоковедческом форуме», где мы продолжили работу. Это демонстрировало и наш прогресс в оцифровке словаря, и являлось мотивацией для новых энтузиастов.

Нашей задачей было не терять интерес к проекту, привлекать новых людей, удерживать старых участников, и продолжать работу. Кроме технической части, были определенные сложности психологического характера по принятию нашего проекта общественностью.

Например, мы столкнулись с тем, что нам задавали вопросы, связанные с интеллектуальной собственностью составителей БКРС; имеет ли право наша команда заниматься его оцифровкой; есть ли среди участников проекта профессиональные китаисты, лингвисты, которые контролируют процесс, проверяют

[1] DSL (Dictionary Specification Language) — язык описания словарей. Разработан специалистами российской компании ABBYY. LSD (Lingvo System Dictionary) — основной формат локальных словарей. Используется в Lingvo (начиная с версии 6.0).

ошибки и т. д. Многие выражали сомнения относительно необходимости нашей работы в целом, упирая на то, что БКРС устарел и для многих практических целей уже не подходит, в нем плохо отражена современная лексика, мало кто сегодня из потенциальных пользователей словаря знаком с традиционными иероглифами и т. д.

Этот трудный этап сомнений нам тоже пришлось преодолеть. Когда работа наполовину была уже проделана, несмотря на значительное снижение энтузиазма, мы все-таки были преисполнены решимости довести ее до конца. Среди наших участников остались только те, кто многое уже сделал и готов был продолжать, и мы всячески поддерживали этих людей, старались как-то морально поощрять их, помогать всем чем можем. И я очень благодарна тем, кто выполнил вместе с нами эту задачу и завершил проект.

В целом, работа по оцифровке БКРС заняла у нас около трех лет: О. Панков опубликовал свою идею в декабре 2002 г., работа началась в первые месяцы 2003 г. и продолжалась примерно до июня 2006 г. Когда мы завершили последнюю страницу, конечно, очень обрадовались. Было решено, что техническая часть закончена, хотя, конечно, на тот момент мы уже мечтали о сайте, о публикации словаря онлайн, о возможности добавлять новую лексику и т. д.

Однако технических возможностей у нас для этого не было. Продолжение проекта в подобном ключе потребовало бы и материальных вложений, и оплаты услуг программистов, и урегулирования вопросов с авторским правом. Подумав, мы решили, что выложим все полученные файлы в открытый доступ, чтобы люди могли пользоваться словарем и всей базой данных уже на свое усмотрение. Таким образом, мы отблагодарили всех за участие, за то, что провели вместе несколько прекрасных лет нашей жизни... Хотя мы все это время то ссорились, то мирились, но в результате остались друзьями на много-много лет, и главное, реализовали важный для себя и профессионального сообщества проект.

После его завершения результаты работы зажили своей жизнью. Данные были опубликованы в формате открытой энциклопедии наподобие Википедии, кто-то форматировал базу под другие словари, которые стали выходить на рынок. При появлении проекта e-БКРС и сайта bkrs.info, было понятно, что они созданы на основе именно нашей базы. Я была очень рада, потому что это было реализацией мечты, которую мы не осуществили: чтобы появился удобный сайт, чтобы имелась открытая база с возможностью пополнения новой современной лексикой.

Я благодарна создателям bkrs.info за то, что они сделали БКРС доступным для многих и китаистов, и японистов, и вообще для всех, кто как-то связан с китайскими иероглифами. К сожалению, словарь частично утратил дух академичности, некоторой даже архаичности в хорошем смысле. Как я поняла, в онлайн-словарь не вошли многие статьи БКРС, которыми редко пользуются; впоследствии все традиционные иероглифы были упрощены, но, похоже, что этот шаг был вызван технической необходимостью и удобством для пользователей.

Однако полная база, конечно, существует: она имеется и в формате DSL, и в формате вордовских файлов. Также, насколько я знаю, созданная нами словарная база практически полностью доступна на сайте jardic.ru, где размещены сло-

вари японского языка. Там сохранено традиционное написание иероглифов, за что хочется сказать спасибо автору сайта Виталию Загребельному.

Заканчивая рассказ об оцифровке БКРС, не могу не выразить благодарность всем, кто придумал и помог реализовать проект, кто вдохнул в него новую жизнь после того, как мы завершили работу.

Когда я слышу, что наш проект стал прорывом в китаеведении, мне очень лестно и я горжусь своей причастностью к этому начинанию. Во время работы над оцифровкой мы занимались интересным, новым делом в свое удовольствие. Мы были молоды, у нас было сообщество единомышленников, площадка для общения, для обмена мнениями, и, я думаю, без этой важной социальной составляющей проект не был бы реализован.

Лично для меня эта история оказалась очень важной как для человека, который приобщился к сообществу востоковедов. Мне очень приятно осознавать, что и академическая наука не воспротивилась любительскому порыву, и проявляет интерес к нашему проекту до сих пор.

Поступила в редакцию: 29.08.2023. Received: 29 August 2023.
Принята к публикации: 01.10.2023. Accepted: 1 October 2023.