И.А.Грунтов

Отдел урало-алтайских языков
Институт языкознания РАН
Москва

О создании электронной базы данных по памятникам монгольских языков

В своем докладе я бы хотел представить проект, который, на мой взгляд, может дать новый импульс для развития целого ряда отраслей монголистики: прежде всего лингвистики, но также и фольклористики, истории и этнографии. Речь идет о создании электронной базы данных, включающей в себя важнейшие средневековые тексты монгольских языков, а в перспективе и тексты иных эпох.

Общеизвестно, что для сравнительно-исторических исследований особое значение имеют данные старых письменных памятников. В частности, для монгольского языкознания изучение памятников среднемонгольского языка всегда было важной задачей. Тем не менее, как ни странно, до сих пор нет ни одного сводного труда, в котором бы рассматривалась лексика всех памятников среднемонгольского языка. Почти нет и обобщающих трудов по грамматике среднемонгольского (среди немногочисленных исключений можно назвать работы М. Вайерса и М.Н. Орловской). Для решения этих и многих других проблем важным инструментом может послужить электронная база текстов, с удобным интерфейсом и многофункциональной системой поиска.

Структура базы. Используемые тексты

На первых порах в базу будут включены основные тексты на среднемонгольском языке. Уже готовы предварительные электронные варианты «Сокровенного сказания монголов», монгольской части многоязычного словаря XIV века Хуа-и и-ю и императорских указов из этого же памятника, арабографичных монгольских словарей XV века Мукаддимат ал-адаб и словаря Ибн-Муханны, Лейденского манускрипта, известного также как Kitâb-i Majmû’ Tarjumân-î Turkî va ‘Ajamî va Muġalî va Fârsî (XIV в), а также первого письменного памятника на монгольском языке - надписи на Чингисовом камне (XIII век). Это стало возможным благодаря помощи японских монголистов Хитоси Курибаяси и Ёсио Сайто, а также российского ученого О.А.Мудрака.

В дальнейшем планируется максимально полно представить в базе все известные на сегодняшний день среднемонгольские памятники: тексты квадратного письма (XIII-XIV вв), синографичные памятники (в частности, Жи Юань и-ю) и арабографичные многоязычные словари (в частности, шестиязычный словарь известный как «Rasûlid Hexaglott»), письма монгольских властителей Ирана — Ил-Ханов — и другие уйгурографичные памятники преклассического монгольского языка, в том числе тексты, обнаруженные в Турфане и различную эпиграфику, записи монгольских слов в текстах западноевропейских путешественников, армянских, грузинских, арабских, персидских и корейских историков.

Большой интерес для истории монгольского языка представляют при этом не только оригинальные произведения, но и переводные тексты. В частности, в базу совершенно необходимо включить переводы на монгольский язык ряда буддийских текстов, выполненные в XIV в. Переводная литература дает важную информацию о процессах освоения заимствований, показывает пути проникновения в язык синтаксических и морфологических калек.

При этом все данные в базе будут представлены в оригинальной графике, будь то китайские иероглифы, арабская письменность, квадратное, уйгурское, армянское или грузинское письмо, снабжены транскрипцией и по возможности транслитерацией, а также различными версиями переводов на разные языки, если таковые имеются.

Так, к примеру, для «Сокровенного сказания монголов» будет дана фонетическая иероглифическая запись каждого предложения; разбивка данного предложения на морфемы; оригинальный китайский перевод; транскрипции С.А. Козина, Л. Лигети, Э. Хениша, И. де Рахевильца, Б. Сумьябаатара; переводы на русский, английский, немецкий, французский, бурятский и современный монгольский языки; будут даны ссылки на параллельные места из Алтан Тобчи. В перспективе предполагается включить и комментарии к отдельным фрагментам текста.

Каждый памятник, присутствующий в базе, должен быть аннотирован - указано время и место создания, дан развернутый историко-филологический комментарий. Помимо предполагаемой или точно известной даты создания документа необходимо указывать и датировку самого раннего дошедшего до нас списка, поскольку часто среднемонгольские документы доходят до нас в поздних копиях, в которые переписчики могли добавлять элементы языка современной им эпохи.

Функции электронной базы

Единицей хранения в базе является предложение, снабженное морфологической разметкой (глоссированием). Пользователь базы может делать поиск, как по словам или их сочетаниям, так и по грамматическим и словообразовательным морфемам. На выходе пользователь будет получать хронологически организованный список памятников (с указанием датировок текстов), в которых встречается искомое слово, словосочетание или морфема — так можно определить начало употребления слова или морфемы в текстах, а перейдя к нужному памятнику он увидит ряд контекстов – предложения, включающие его поисковый запрос, и их перевод. При этом всегда будет возможность посмотреть и более широкий контекст – абзац, главу, а при желании и весь текст полностью. Таким образом база будет существовать и функционировать в двух режимах:

— Во-первых, как собрание текстов, т.е в виде электронной библиотеки. Такой режим, очень удобен для изучения содержания текстов, но довольно мало пригоден для исследования их языка. Впрочем, и в таком виде многооконный режим работы базы позволяет одновременно выводить на экран два и более текста и извлекать информацию из такого сравнения.

— Во-вторых, как инструмент анализа, позволяющий искать примеры на любое лингвистическое явление за считанные секунды на огромном объеме текстов. При таком режиме пользователь сможет свободно изучать особенности словаря и грамматики соответствующих текстов и одновременно будет избавлен от необходимости трудоемкой и длительной росписи материалов вручную, что долгое время было неизбежно при отсутствии подобных электронных средств.

Перспективы развития

Каковы возможные дальнейшие варианты развития обсуждаемой электронной базы? Основных направлений может быть два:

а) распространение во времени. Включение в базу не только текстов среднемонгольского периода (13-16 вв), но и текстов классического письменно-монгольского языка, различных вариантов фольклорных и эпических текстов (в частности, эпоса «Гэсэр»), а также текстов современного монгольского языка, художественных, научных, газетных, официальных и т.п. Это позволит создать национальный корпус монгольского языка — информационно-справочную систему для широкого круга специалистов. Национальный

корпус представляет собой собрание разножанровых текстов некоторого языка и может служить, как для научных исследований, так и для практических целей обучения языку. Такие корпуса существуют уже для целого ряда языков, в том числе русского (http://ruscorpora.yandex.ru) и британского английского (http://sara.natcorp.ox.ac.uk/lookup.htm).

б) распространение в пространстве. Вполне возможно, включать в базу тексты не только на среднемонгольском языке, но и на других монгольских языках – старых и современных. В частности, сейчас уже вполне ощущается необходимость создания аналогичных баз данных по текстам на бесписьменных монгольских языках: дагурском, монгорском, могольском, ордосском, хамниганском, шира-югурском. Уже доступны необходимые материалы для создания большого корпуса ойратских и старокалмыцких текстов, имеются электронные варианты эпоса «Джангар». Есть возможность подготовки старобурятских памятников.

Оба этих варианта не исключают друг друга, однако выполнение столь широко поставленной задачи потребует больших временных затрат.

Что требуется для выполнения проекта?

Для работы над базой необходимо объединить усилия многих специалистов. Прежде всего требуются люди, которые будут сканировать, распознавать и вычитывать тексты. Количество среднемонгольских текстов велико, но конечно, а следовательно включение в базу всех известных среднемонгольских памятников является достижимой целью. Следует привлечь программистов для создания программной оболочки и удобного интерфейса базы, разработки дополнительных возможностей организации поиска. Наконец, была бы крайне желательна помощь в развитии проекта со стороны лингвистов, отвечающих за отбор и сверку текстов, их морфологическую разметку. Насколько мне известно, на кафедре тюркологии Школы монгольского языка и культуры при Монгольском государственном университете коллектив ученых, возглавляемый Ёндонжамцын Жанчивом и Цэндийн Баттулгой, уже ведет работы по созданию электронной библиотеки среднемонгольских текстов, которая хотя и отличается по концепции от данного проекта, но может быть интегрирована в него для достижения общей пользы.

Что подобная база данных может дать лингвистам?

В этом подразделе я хотел бы, хотя бы примерно, очертить круг тех возможностей, которые может дать электронная база по среднемонгольским данным исследователям монгольских языков. Возможные применения базы в других областях монголистики достаточно очевидны, но носят, с моей точки зрения, скорее вспомогательный характер, поскольку этот проект предназначен прежде всего для выполнения именно лингвистических задач.

1) Прежде всего и специалистам по лингвистической типологии, и монголистам окажется чрезвычайно полезной возможность поиска примеров на различные лингвистические явления. Так, например, известно, что в Сокровенном сказании и некоторых других памятниках присутствуют реликты согласования показателей глагола с субъектом по роду, однако насколько широко распространено это явление в среднемонгольском до сих пор точно не известно. Такая проблема достаточно быстро может быть решена с помощью электронной базы данных.

В идеале, наличие такой возможности требует полной морфологической разметки всех текстов (как это сделано, например, в базе данных по древнетюркским текстам,

расположенной по адресу http://vatec2.fkidg1.uni-frankfurt.de), однако поскольку такая разметка требует больших затрат времени, то на начальном этапе может оказаться достаточным опубликовать отдельным списком все аффиксы монгольского языка (со всеми возможными алломорфами) с указанием их значения. Это позволит проводить поиск по конкретным аффиксам.

2) Наличие электронного корпуса текстов позволяет с большой точностью отыскивать все употребления сходных грамматических форм и выявлять их дистрибуцию. Для монгольского языка по-прежнему сохраняют актуальность вопросы о точном семантическом различии между формами прошедшего времени, о влиянии категории определенности-неопределенности на выражение падежных морфем, о распределении форм дательного-местного падежа и т.п. Все эти вопросы и многие другие могут получить свое решение при использовании корпуса.

3) Электронная база облегчает поиск изоглосс в языках различных памятников. Это может, в частности, уточнить наши знания о диалектном состоянии среднемонгольского языка, которое в настоящее время практически не описано. Фактически, большинство специалистов связывают диалектное членение данного идиома с разделением памятников на синографичные (восточные) и арабографичные (западные), при этом не опираясь на собственно лингвистические данные. В то же время отдельные работы показывают, что деление среднемонгольского языка на два диалекта не совпадает в точности с делением памятников по системам письма и географической локализации. Электронная база данных может помочь выделить пучки изоглосс и окончательно решить вопрос о распределении памятников по диалектам, прояснив тем самым этот запутанный момент таксономии монгольских языков.

Кроме того, к настоящему времени все обобщающие работы по грамматике среднемонгольского языка безнадежно устарели, поскольку в последние годы было открыто множество новых монголоязычных текстов данного периода (преимущественно, текстов квадратного письма), которые по объективным причинам не могли быть приняты во внимание авторами соответствующих грамматик. Электронная база может стать удобным инструментом для построения новой полной грамматики среднемонгольского языка.

4) Довольно часто можно встретить жалобы специалистов по монгольскому языкознанию на отсутствие до сих пор сводного среднемонгольского словаря, который охватил бы материал всех или хотя бы многих памятников (ср. например, специальные указания на это в наиболее современной энциклопедии по монгольским языкам «The Mongolic Languages», 2003, стр.61). При наличии такой базы создать этот словарь будет гораздо проще, чем вручную. При этом электронный словарь среднемонгольского языка может обладать и свойствами конкорданса, то есть при описании значения каждого слова есть возможность привести **все** контексты его употребления во всех памятниках.

5) Лексикографическим целям может послужить и создание частотного словаря среднемонгольского языка, который можно сгенерировать в полуавтоматическом режиме (нужно лишь разграничить омонимы) на основе имеющихся текстов.

6) Электронная база открывает новые возможности для написания сравнительно-этимологического словаря монгольских языков, о необходимости которого разговоры ведутся с середины прошлого века. Для построения этимологического словаря важно знать дату первой письменной фиксации слова, а также то, в каких памятниках первоначально

встречалось слово. Если, скажем, некоторое слово впервые встречается в тексте, переведенном с тибетского языка в относительно недавнее время, то вероятность его исконного происхождения невелика.

7) Можно повысить лексикографическую ценность электронной базы, представив каждое слово в ней в виде гиперссылки, при нажатии на которую будет появляться словарная статья данного слова из разных источников, в частности толкового словаря Я. Цэвэла, монгольско-русско-французского словаря О.Ковалевского, четырехтомного Большого академического монгольско-русского словаря, Этимологического словаря алтайских языков (Starostin et al.) и, в перспективе, из сравнительно-этимологического словаря монгольских языков.

8) В перспективе, при расширении базы и включении в нее материала и других монгольских языков разных эпох, она может сыграть важную роль в написании новой сравнительной грамматики монгольских языков. Имеющиеся сравнительные грамматики Н. Поппе и Г.Д.Санжеева значительно устарели, а использование полной информации о языке памятников может помочь создать новую сравнительную грамматику на качественно новом уровне описания.

9) Широкие возможности открываются и для текстологических исследований. Благодаря электронной базе можно легко осуществлять сопоставление фрагментов различных памятников, что, в частности, удобно при анализе заимствований из одного памятника в другой или при анализе распространения какой-либо фольклорной формулы.

10) Сохранение оригинальной графики при электронной публикации памятников наряду с транслитерацией/транскрипцией позволяет упростить сравнительные графологические исследования. В частности, до сих пор остается не до конца проясненным вопрос о том, могли ли китайские писцы при иероглифическом транскрибировании монгольской речи XIII-XIV вв. отразить с помощью выбора иероглифов с определенным среднекитайским тоном какие-либо просодические особенности среднемонгольского языка, которые было невозможно передать в любой иной графической системе, когда-либо применявшейся для монгольского языка.

Итак, осуществление описанного проекта несомненно должно послужить к дальнейшему развитию монголистики, облегчит работу многих монголистов и, возможно, приведет к открытию многих языковых закономерностей, которые было невозможно выявить работая с разрозненными текстами.

ЛИТЕРАТУРА

Большой академический монгольский русский словарь, тома I-IV, Москва, 2001-2002.
Ковалевский О. Монгольско-русско-французский словарь. I-III, Казань, 1844,1846, 1849.
Орловская М.Н. Язык монгольских текстов XIII-XIV вв. Москва, 1999.
Санжеев Г.Д. Сравнительная грамматика монгольских языков, тома 1-2, Москва, 1953-1964.
Цэвэл Я. Монгол хэлний товч тайлбар толь. Улаанбаатар, 1966.
Juha Janhunen (eds.). The Mongolic Languages. London-New York, 2003.
Poppe N. Introduction to Mongolian Comparative Studies //Mémoires de la Société Finno-Ougrienne 110, Helsinki, 1955.

Starostin S.A. Mudrak O.A. Dybo A.V. (with assistance of I.Gruntov and V.Glumov) Etymological dictionary of Altaic Languages. Leiden-Brill, 2003.
Weiers M. Untersuchungen zu einer historischen Grammatik des präklassischen Schriftmongolisch. Wiesbaden, 1969.