Ю.В. Кнорозов

Забытые системы письма

Остров Пасхи, Великое Ляо, Индия

Москва
«Книга по Требованию»

УДК 81
ББК 81
Ю11

Ю.В. Кнорозов
Ю11 Забытые системы письма: Остров Пасхи, Великое Ляо, Индия / Ю.В. Кнорозов – М.: Книга по Требованию, 2023. – 296 с.

ISBN 978-5-458-28599-5

Материалы по дешифровке.В сборнике излагаются некоторые результаты дешифровки протоиндийских, киданьских и рапануйских текстов. Исследование велось с привлечением вычислительной техники. В статьях последовательно отражены основные этапы исследования недешифрованных текстов: математическая обработка, формальный анализ и переход к филологической обработке. Имеется большое количество иллюстраций с примерами текстов.

ISBN 978-5-458-28599-5

Как правило, в фонемных алфавитах единый принцип полностью не выдерживается; например, некоторые знаки передают две фонемы, и, наоборот, некоторые фонемы передаются группами знаков.

Количество знаков в разных системах письма резко различается. В иероглифическом (морфемно-силлабическом) оно доходит примерно до 400 синхронно употребляемых знаков, в силлабическом — до 100, в фонемном — до 40, в инфрафонемном — до 10. По количеству употребляемых знаков можно судить о характере письма, для чего, однако, требуется обширный текст, в котором представлена основная часть знаков.

При наличии только короткого текста о характере письма можно судить по убыванию появления новых знаков, наиболее медленному в иероглифической записи. Естественно, чем меньше знаков в письме, тем длиннее цепочка знаков, передающих слова; однако в иероглифике цепочки могут удлиняться за счет добавления детерминативов и звуковых подтверждений, т. е. в зависимости от правил орфографии.

Были найдены записи, сделанные полностью забытым письмом на неизвестном языке, в связи с чем возник вопрос об их изучении и по возможности дешифровке.

Термин «дешифровка» понимается весьма различно и требует уточнения. Прежде всего следует отметить, что дешифровка исторических систем письма и дешифровка секретных шифров не имеют почти ничего общего. В древних текстах знаки стоят в обычном порядке, но чтение их забыто; в шифрованных записях известные знаки заменены другими и порядок их смешан. В первом случае язык либо неизвестен, либо сильно изменился, во втором — известен.

Под дешифровкой в узком смысле следует понимать установление чтения забытых знаков. Однако чтение текста отнюдь не означает его понимания, так как язык мог полностью исчезнуть или же сохраниться в виде языков-потомков, отличающихся по грамматике и лексике. Некоторые древние тексты (например, этрусские) написаны известным письмом, но на вымершем языке. Таким образом, наряду с дешифровкой письма необходимо изучение языка неизвестных текстов. Наконец, если достаточно известны и чтения знаков, и язык, необходимо дать чтение, перевод и интерпретацию каждого конкретного текста со всеми его особенностями, что, собственно, относится уже к области филологии, но часто называется дешифровкой текста.

Чтобы прочесть текст, необходимо знание данного кода (т. е. чтения знаков и правил их употребления) и знание языка, на котором написан текст. Предполагается, что текст, подлежащий дешифровке, является записью человеческой речи. Человеческие языки сильно изменялись с течением времени, но во всех языках всех времен используются сходные способы передачи информации. Кроме того, предполагается, что подлежащий дешифровке текст не зашифрован умышленно. При наличии умышленной зашифровки необходимо предварительно восстановить нормальный порядок знаков, а затем уже вести дальнейшие исследования.

При дешифровке неизвестных текстов основным и решающим источником информации являются сами тексты. Однако о любом тексте всегда имеется дополнительная информация, которую можно использовать для целей дешифровки. Так, сведения о времени, к которому относятся тексты, дают возможность ограничить круг поисков лингвистических и графических аналогий и определить хронологический разрыв между изучаемым языком и языком-потомком. Сведения о месте находки и объекте, на котором начертан текст, могут дать указания о его содержании. Иногда тексты сопровождаются изображениями, которые могут оказаться важным источником дополнительной информации.

Неизвестный текст может сопровождаться параллельным известным текстом. Последний может оказаться либо независимым текстом на ту же тему (псевдобилингва), либо переводом неизвестного текста (билингва). Значение билингвы для дешифровки трудно переоценить. Дешифровщик как бы получает специально подготовленное учебное пособие (как известно, такие пособия — параллельное издание иностранного текста и перевода — широко используются и при изучении иностранных языков). Однако отсутствие билингвы вовсе не означает невозможности дешифровки.

Изучение текста требует его формализации. Прежде всего текст должен быть транскрибирован стандартными знаками. В качестве последних могут быть использованы стандартизированные знаки изучаемого письма, а также (для удобства обработки и публикации) общепринятые знаки (цифры, буквы). Эта работа требует не только большой точности, но и приобретения специальных навыков — овладения данным шрифтом и индивидуальным почерком. Составление транскрипции предусматривает опознание всех вариаций написания, а также полустертых и искаженных графем, восстановление утраченных мест, обнаружение ошибок и внесение конъектур. Эта работа обычно не бывает закончена к моменту дешифровки и продолжается по ходу дешифровки и после нее. Ошибки при опознании графем представляют значительную опасность, так как исправление их затруднительно.

По ходу составления унифицированной транскрипции возникает необходимость в составлении каталога графем (т. е. знаков и их аллографов). Составление каталога дает возможность начать планомерную работу по выявлению аллографов, выделить особые группы знаков (например, цифры) и «модель порождения» знаков, развернуть работу по сопоставлению изучаемого алфавита с известными, а иногда и по опознанию изображаемых знаками предметов, что может дать важную дополнительную информацию.

При формальном изучении текстов исследователь временно игнорирует всю дополнительную информацию, сосредоточиваясь исключительно на той, которую несут сами тексты. Дополнительная информация оказывается необходимой на более поздних этапах изучения текстов.

Для удобства исследования текст целесообразно рассматривать как ряд морфем, расположенных в последовательности, свойственной данному языку. Общее количество морфем в любом языке не зависит от количества фонем и не превышает синхронно 1500. Стабильность количества морфем определяется свойствами человеческого мозга. Превышение критического количества создает трудности для запоминания (оперативной памяти). В свою очередь, значительное уменьшение числа морфем повлечет за собой удлинение словоформ и создаст трудности для их распознания (т. е. для восприятия устной речи). Возможное число сочетаний фонем резко ограничено законами образования морфем в данном языке (фиксированные ограничения). Морфема—наименьшая семантическая единица языка, и поэтому она обычно является предельным референтом знака письма. Каждая группа тождественных морфем характеризуется позициями этих морфем в ряду (адресами) и частотой.

Все морфемы можно подразделить на корневые и служебные. Такое подразделение имеет относительный характер, так как в ряде случаев корневые морфемы могут употребляться в качестве служебных. Однако данная морфема, занимающая конкретную позицию в ряду, является альтернативно либо корневой, либо служебной.

С помощью служебных морфем образуются словоформы и осуществляется связь между словами в предложении. Количество морфем в словоформе обычно практически не превышает пяти. Общее количество служебных морфем, естественно, значительно меньше, чем количество корневых. Так как одна и та же служебная морфема соединяется с различными корневыми морфемами, частота наиболее употребительных служебных морфем должна намного превышать частоту корневых морфем. Однако в специфических текстах (где часто повторяются некоторые слова) рекордную частоту могут иметь и корневые морфемы. Чтобы избежать влияния специфики текста, целесообразно учитывать помимо абсолютной частоты также относительную, исключая повторяющиеся блоки. При этом рекордные по обсолютной частоте корневые морфемы займут соответствующее место независимо от специфики текста.

Ряд морфем, составляющий текст, может быть разделен на отдельные цепочки, соответствующие морфемам, словоформам и предложениям данного языка. Для целей дешифровки имеет смысл разделить текст на цепочки, соответствующие словоформам.

Общепринятой формой изучения языка является составление словарей и грамматик. Сведения по лексике сосредоточиваются именно в словарях. Поэтому при изучении языка неизвестного текста также целесообразно иметь упорядоченный набор цепочек, представляющих лексику данного языка.

В словоформах служебные морфемы располагаются обычно в начале и конце, а в некоторых языках и в середине слова (внутренняя флексия). При разбивке текста на цепочки практически целесообразно включать в состав словоформы не только корневые, словообразующие и словоизме-

няющие морфемы, но также и все остальные служебные морфемы (например, предлоги и послелоги, частицы, союзы). Дело в том, что дешифровщик, ставящий задачу не допускать присоединения к словоформе служебных морфем (выделяемых в грамматиках в качестве самостоятельных частей речи), должен иметь критерии, позволяющие отделить словообразующие и словоизменяющие морфемы от остальных служебных морфем. Такие критерии вообще дать очень трудно, а до начала исследования текста просто невозможно. Наоборот, именно изучение текста может дать основания для классификации служебных морфем. Однако если даже и удалось бы отделить словообразующие и словоизменяющие морфемы от остальных служебных, то изучение последних от этого только затруднилось бы. В самом деле, грамматические функции служебных морфем можно определить, только изучая их сочетания со знаменательными словами, а в этом случае служебные морфемы оказались бы от них изолированными.

Иногда к словоформе приходится присоединять и определение. Это целесообразно в тех случаях, когда определение выражено неизменяемыми словами, лишенными своих служебных морфем и именно поэтому практически неотличимыми от них в неизвестных текстах.

Таким образом, неизвестный текст целесообразно разделить на отдельные цепочки, в общем соответствующие словоформам, к которым присоединены также служебные морфемы непосредственного окружения и неизменяемое определение. Такие цепочки получили название б л о к о в.

Технически разделение текста на блоки может вестись следующим образом.

Регистрируются все цепочки, повторяющиеся в тексте дважды и чаще. Такие цепочки могут соответствовать словоформе (если она встретилась более одного раза), основе словоформы (если ненулевой словоизменяющий показатель встретился один раз), корню словоформы (если ненулевые словообразующий и словоизменяющий показатели встретились один раз), сочетанию части одной словоформы с частью другой (случайно повторенному не менее двух раз), сочетанию двух или нескольких словоформ (повторенному не менее двух раз).

Ориентируясь на расчетную длину блока и на практическую длину, полученную в результате выделения блоков, легко устранить слишком длинные повторяющиеся цепочки, сильно превышающие среднюю длину блока.

Все зарегистрированные цепочки упорядочиваются в порядке убывания частоты. Кроме этого их целесообразно упорядочить по составляющим знакам на основе принятого каталога (например, в порядке нарастания номеров знаков при числовой транскрипции). Упорядоченный набор зарегистрированных цепочек может рассматриваться как исходный материал для составления словаря блоков.

Многие тексты имеют уже в оригинале разделение на цепочки знаков. Такие цепочки во всех текстах, имеющих разбивку, соответствуют,

как правило, словоформам, к которым могут присоединяться различные служебные морфемы и неизменяемое определение.

Ближайшей задачей после установления последовательности знаков в ряду и составления стандартной транскрипции является изучение состава блоков с целью выяснить морфологию изучаемого языка.

Состав блоков целесообразнее всего изучать с помощью словарей, построенных в порядке возрастания номеров знаков. Эти словари могут быть упорядочены по первому знаку (прямой словарь), по последнему знаку (обратный словарь) и по внутренним знакам (глубинные словари). Прямой и обратный словари необходимы во всех случаях. Необходимость составления глубинных словарей определяется спецификой данного языка.

Комбинируя данные прямого и обратного словарей, можно легко получить набор микропарадигм (грамматических показателей, употребляемых с данными корнями), а затем свести их в парадигмы.

Имеется некоторая опасность спутать знаки, входящие в состав устойчивой (корневой) группы блока, с переменными, передающими грамматические показатели. Однако если сомнительные знаки действительно входят в состав устойчивой группы, то они не должны встречаться в качестве перемещных перед другими заведомо устойчивыми группами и не должны заменяться заведомо переменными знаками.

Выделение переменных знаков (инициальных и финальных) дает возможность исключить из словаря блоков многие случайные цепочки (случайно повторенные сочетания конца одной словоформы и начала другой), а также приступить к расчленению на блоки оставшихся нерасчлененными частей текста.

Характеристика блоков должна включать также сведения о позициях, которые они могут занимать в предложениях. Как известно, во многих случаях связь между словами обеспечивается только порядком слов, без морфологических показателей. Учет позиций блоков особенно важен для языков, которым свойствен твердый порядок слов. Впрочем, и в тех языках, синтаксис которых не требует твердого порядка слов, обычно имеется предпочтительный порядок (зависящий от литературного стиля).

Определение порядка слов значительно облегчает установление функций грамматических показателей. Появляется возможность расположить последние в синтаксическом порядке («синтаксическая сетка») и дать их классификацию (например, выделить словообразующие, словоизменяющие и иные служебные морфемы).

С другой стороны, определение порядка слов дает возможность развернуть классификацию блоков по частям речи (условным или фактическим) или по иным группам, удобным для изучаемого материала.

Выявление служебных морфем и их функций открывает возможность развернуть сопоставление изучаемого языка с известными языками. Первой задачей является отыскание ближайшего языка-потомка (или группы таких языков). При установлении генетического родства решающую роль играет, конечно, общая характеристика морфологии и синтак-

сиса. Для детального сопоставления грамматических показателей необходимо подвергнуть тексты на языке-потомке такой же обработке, какой подвергались древние тексты. Кроме того, в зависимости от величины хронологического разрыва между изучаемым языком и языком-потомком в последнем должны быть восстановлены старые грамматические формы методами внутренней реконструкции и сравнительно-историческими (эта работа может быть выполнена только профессиональными лингвистами). Сопоставление служебных морфем изучаемого языка и языка-потомка дает возможность приписать соответствующим знакам условное чтение (которое не следует смешивать с фактическим чтением).

Изучение морфологии и синтаксиса и классификация блоков дают возможность развернуть изучение лексики неизвестных текстов. При переходе к фонетическому чтению решающую роль могут сыграть условные чтения знаков, установленные при сопоставлении грамматических показателей изучаемого языка и языка-потомка. Однако фонетическое чтение слов во многих случаях не дает возможности определить их смысл. Для успешного изучения неизвестной лексики необходимы специальные морфемные словари языка-потомка и детальное изучение фонетических изменений. Кроме того, даже в тех случаях, когда перевод вполне возможен, текст остается непонятным по причине полной невразумительности. Для того чтобы придать древним текстам смысл, кроме грамматического перевода необходим широкий и всесторонний комментарий. Составление такого комментированного перевода уже выходит за рамки формального изучения текстов и, несомненно, требует привлечения всей возможной дополнительной информации.

М. А. Пробст

ИССЛЕДОВАНИЕ НЕИЗВЕСТНЫХ ТЕКСТОВ

Одним из важнейших вопросов развития средств информации на современном этапе человеческой культуры является проблема автоматизации преобразований информации, сжатия информации, извлечения из текста заданной информации. Это особенно важно сейчас, когда поток информации буквально захлестывает человечество. Изучение исторических систем письма можно рассматривать прежде всего как задачу извлечения из неизвестного нам текста информации о структуре самого текста, что является основой дешифровки исторических систем письма.

Имеется осмысленный текст, записанный на неизвестном языке. Нужно, исходя в первую очередь из самого текста, выяснить свойства неизвестного языка и уже затем путем сопоставления с известными языками и привлечения с большой осторожностью внетекстовой информации (археологических, исторических, филологических и иных сведений) передать смысл неизвестного текста. Возможность исследовать текст формальными методами перерастает в необходимость, если мы хотим максимальным образом исключить субъективный анализ текста. Нас интересует именно этот аспект дешифровки, поскольку он допускает точную постановку задачи, хотя это только одна из подзадач при изучении неизвестных текстов.

Прежде чем перейти к уточнению задачи дешифровки исторических систем письма, сделаем два замечания.

Часто дешифровка исторических систем письма ассоциируется с дешифровкой секретных шифров, что приводит к мысли о применении методов последней для исследования неизвестных письменных источников. От этого необходимо всячески предостеречь. Задачи, решаемые криптографией и дешифровкой исторических систем письма, почти противоположны, и методы первой могут весьма ограниченно применяться во второй.

Действительно, в криптографии предполагается, что исходный, незакодированный текст написан на известном языке, причем обычно либо этот язык известен, либо круг кандидатов на эту роль из числа хорошо известных языков весьма невелик. Последним обстоятельством объясняется успех американской разведки, которая для передачи секретных со-

общений использовала редкий язык американских индейцев, мало кому известный, так что сообщения передавались без зашифровки.

Хотя язык текста в криптографии предполагается известным, зашифрованный текст может быть весьма сложным образом связан с исходным текстом. Цель криптографии состоит в восстановлении исходного текста по зашифрованному тексту, в котором всякого рода статические распределения элементов текста могут очень сильно отличаться от соответствующих распределений элементов в исходном тексте.

В задачах криптографии известен язык исходного текста, но неизвестно преобразование, благодаря которому возник исследуемый текст. В задачах дешифровки исторических систем письма неизвестен язык, на котором написан текст, но сам текст не подвергался специальной обработке, имеющей целью затруднить чтение текста; текст записан в соответствии с нормами орфографии данного языка.

Из всего сказанного видно, что методы криптографии могут быть применены для целей исследования исторических систем письма весьма ограниченно. Они могут быть рассмотрены лишь как набор статистических методик обработки текста. В дальнейшем мы не будем касаться вопросов криптографии, и поэтому употребление термина «дешифровка» будет иметь однозначный характер.

Так как при дешифровке объем статистической обработки текста весьма велик, то естественно возникает вопрос, можно ли применить вычислительную технику.

С конца 50-х годов специалисты в области дешифровки исторических систем письма стали придавать большое значение машинной обработке изучаемых текстов. В широкой прессе появилось выражение «машинная дешифровка». В некоторых популярных статьях утверждалось даже, что вычислительная машина, получив неизвестный текст, может «выдать» его транскрипцию (например, латинскими буквами), а заодно и перевод.

Следует отметить, что широко разрекламированные возможности и надежность «машинной дешифровки» исторических текстов сильно преувеличены. В настоящее время не существует программ, на основе которых электронные вычислительные машины (ЭВМ) могли бы устанавливать чтение знаков неизвестного текста, а тем более его переводить.

Тем не менее использование вычислительной техники при дешифровке неизвестных текстов имеет большой смысл, так как позволяет осуществить очень громоздкую обработку, которая «вручную» заняла бы много времени. Разбивка нерасчлененного текста на блоки, составление прямых и обратных словарей, выявление формальной грамматики, безусловно, могут и должны (если позволяет размер текста) вестись с помощью вычислительной техники.

Неизвестный текст вводится в вычислительную машину в цифровой транскрипции, которая составляется вручную. Ошибки, допущенные при составлении цифровой транскрипции, могут сильно исказить результаты машинной обработки и привести к ложным заключениям. Результаты,

полученные при машинной обработке неизвестного текста, целиком зависят от программ, которые составляет не сама машина, а специалист-программист. Недостатки или ошибки программы, естественно, отражаются в полученных данных. До настоящего времени целью машинной обработки было получение исходных материалов для филологов. Разумеется, ошибки филолога, сделанные при лингвистической интерпретации машинных материалов, никоим образом не дают основания отвергать машинную обработку текстов вообще, если она правильно проведена.

В дешифровке, как и в других отраслях знаний, бытуют мифы об «озарениях», приведших к открытиям. Здесь и наитие Г. Гротефенда при дешифровке древнеперсидских клинописных знаков, и озарение Б. Грозного при работе над хеттскими надписями, так же как история о яблоке Ньютона вкупе с ванной Архимеда. Но при этом забывают, что знаменитое восклицание «Эврика!» означает «Нашел!», т. е. что Архимед искал и имел цель поиска. Ньютон, наверное, видел не раз, как падают яблоки с яблони, и трудно предположить, чтобы открытию им законов тяготения не предшествовала предварительная большая работа. Также трудно поверить в озарение Б. Грозного и в наитие Г. Гротефенда.

Отличие открытий с помощью «озарений», «наитий» от стандартных открытий лишь в том, что в первом случае не был опубликован «алгоритм открытия» ни до, ни после него, а при стандартных открытиях так или иначе алгоритм хоть частично публиковался — например, в виде сведения проблемы к серии подпроблем, допускавших более легкое решение, чем вся проблема.

Все это говорится лишь для одной цели: ждать «озарения» без разработки программы нельзя — не будет озарения. «Озарение» может возникнуть лишь иногда при случайном переборе. Нужен общий план поиска, который может уточняться, изменяться в процессе решения, в процессе отбрасывания путей, не ведущих к решению. Кстати, всякое негативное решение задачи, т. е. указание тех случаев, где задача не может быть решена, имеет огромную ценность, так как тем самым сужается область поиска решения задачи. Все эти рассуждения становятся особенно актуальными теперь, когда пытаются иногда фетишизировать вычислительную технику, а иногда, что почти то же, полностью отрицать разумность применения ее при решении задач дешифровки. Все зависит от человеческого разума и его применения.

Используя средства вычислительной техники, нужно знать, для чего, для какой цели применяются машины и где и как будут использоваться данные машинной обработки. Делать «просто так», «когда-нибудь пригодится кое-что из» не только не экономично, но приводит к такому обилию материалов, что разобраться в них труднее, чем работать без них. Например, если исследуются знакосочетания в тексте, то бессмысленно получать сведения о всех возможных знакосочетаниях по три знака. Ведь даже при алфавите в 30 знаков количество тройных сочетаний будет оцениваться тысячами (их 27 000, но не все могут быть реализованы

в тексте). Из них могут быть нужны лишь знакосочетания с некоторыми знаками, знакосочетания определенной структуры и т. п.

Перед тем как выполнить некоторую работу с помощью средств вычислительной техники, нужно твердо знать, как будут использоваться полученные результаты, иначе они могут оказаться ненужными или даже могут затруднить работу исследователя. В этой связи вспоминается высказывание Б. Паскаля, который, сравнивая математику с жерновом на мельнице, заметил, что от того, как поставлен жернов, зависит, что получится — мука или труха.

Выше отмечалось, что одна из основных задач дешифровки текста состоит в выяснении структуры текста, принципов его построения и построения формальной грамматики языка текста. Отсюда ясно, как велико значение дешифровки в общем цикле семиотических проблем, таких, как формализация перевода текста с одного языка на другой, формализация процессов реферирования, сжатия текста и информационного поиска и т. п.

Под формальной грамматикой текста мы понимаем набор структурных элементов, выделенных в тексте, подобных знакам алфавита, морфемам, словоформам, и выведение законов взаимодействия внутри наборов и между наборами, правил преобразований и построений.

При исследовании текста формальными методами можно получить формальное описание структуры текста и формальную грамматику (насколько позволяют объем и характер текста), но нельзя установить смысл текста. Это можно сделать, привлекая материалы известных языков.

Прежде чем перейти к изложению общих принципов изучения неизвестных систем письма, нужно уточнить, что мы понимаем под терминами «знак текста» и «текст».

Первой задачей исследования лингвистического текста является составление каталога знаков текста, выявление аллографов и т. п. При этом возникает вопрос: что же считать знаком текста? Для того чтобы сформулировать формальное понятие «знак текста», выясним, что мы обычно вкладываем в это понятие. Интуитивно мы предполагаем, что текст является последовательностью некоторых частей, причем «самые мелкие» части текста, из которых состоят другие конструкции текста (морфемы, словоформы, предложения), и есть знаки текста. Но эти «мелкие части» текста еще достаточно велики, чтобы появляться в тексте самостоятельно, без постоянного сопутствующего набора других таких же частей текста.

Таким образом, знаком текста мы будем называть элемент такого разбиения всего текста, при котором будут выполняться два условия:

а) каждый элемент разбиения текста имеет самостоятельное распределение в тексте, т. е. появление знака в тексте не может однозначно предсказать появление в тексте соседних с ним других знаков;

б) если разбить текст на более мелкие части, то последние не обладают самостоятельным распределением (иначе говоря, при фактическом самом