Проблемы представления букв в диахроническом корпусе или метакорпусе

А. В. Коваленин

Институт систем информатики имени А.П.Ершова СО РАН, Новосибирск, Россия

Исследователи языка древних текстов чаще всего работают с однородным корпусом – текстами рукописей из одного археографического слоя, то есть локальной графико-орфографической традиции (далее – просто “традиции”). Используя привычные для себя технологии, они редко испытывают нужду в строгой постановке технологической задачи и ощущают лишь нехватку отдельных, характерных для этого слоя символов. С такой частной точки зрения решение проблемы видится только в добавлении в систему способа отображения этих недостающих символов.

В то же время, с одной стороны, создаются корпуса, собирающие различные по времени и месту написания тексты в своём особом целевом разрезе и не ограниченные одной традицией или одним временем, то есть являющиеся *диахроническими*, или просто *смешанными*. С другой стороны, стоит вопрос о создании *метакорпусов* – систем, позволяющих одновременно работать с единицами хранения независимо разработанных корпусов. Обе эти проблематики требуют учёта всех требований к системам представления письменности (СПП) отдельных традиций.

Но суммирование таких частных пожеланий существенно меняет ситуацию.

Во-первых, если в отдельной традиции роль каждого знака хорошо изучена и алфавит представлен известным числом знаков, то техническое решение для всей исторической совокупности традиций лишено возможности опереться на своё главное интуитивное основание – исторически сложившийся алфавит. Оказывается, что, с одной стороны, никогда нельзя быть уверенным в достаточной полноте множества учтённых знаков, а с другой стороны, само соответствие знака и означаемого меняется, причем за счёт обеих сторон.

Во-вторых, обнаруживается, что разные пожелания неявно подразумевают разные технологии работы; но решения, удачные для одной технологии, могут оказаться неудобными для другой.

В итоге, штурмуемая сегодня разными исследователями задача создания универсального представления символов оказывается не поставленной во всех важных аспектах.

1. Неопределённость *предмета*. Нет единого представления даже о перечне традиций, которые должна обслуживать искомая система. Диапазон предложений по представлению кириллических письменностей в этом отношении широк. На одном полюсе мы видим предложения, имеющие в виду разные изводы церковнославянского языка (распространяемые, в принципе, и на глаголическое письмо). На другом – поддержку всех видов письменности на кириллической основе, включая искусственно созданные уже в XX веке алфавиты неславянских языков народов СССР.

2. Неопределённость *деятельностного контекста*. Задача ставится вне конкретной дисциплины работы исследователя. Однако для каждой технологической платформы уже сама техническая постановка вопроса о СПП выглядит по-разному. Кроме того, каждая дисциплина работы включает в себя разные стороны взаимодействия человека и человека, человека и машины, внутримашинных протоколов, для которых разные СПП по-разному удобны или безразличны. Таким образом, сама постановка задачи об СПП для корпуса требует обзора всей необходимой деятельности с корпусом как при его наполнении, хранении и публикации, так и при решении на его материалах научных задач.

Информатики, видя неопределенность задачи, не берутся решать ее без необходимой информации от филологов, но сами не решаются ставить задачу филологам на их языке – в терминах фонологии, исторической и сравнительной лингвистики. Филологи, идущие по пути продвижения отдельных дополнений, не всегда видят проблему в целом и слабо представляют возможный спектр технических решений как в отношении платформ, на которых строятся корпуса, так и в области шрифтовых решений для конкретной платформы. В результате общая проблема об объединяющей СПП подменяется частной проблемой создания универсальной шрифтовой кодировки в рамках стандарта Юникод. В результате решение становится не только мало полезным для корпусов на целом ряде платформ, но и несвободным от ранее утверждённых решений этой частной задачи.

Следовательно, решение проблемы должно идти по пути совместного ее обсуждения лингвистами и программистами. В целом можно говорить о двух аспектах ее рассмотрения.

1. С одной стороны, необходимо показать сложность проблемы по сравнению с задачей кодирования одной традиции. Рассмотрим только наиболее очевидные соображения.

А. Внешне одинаковые знаки разных традиций могут выполнять разные функции, а внешне разные – идентичную функцию. Например, одна и та же (в интуитивном понимании) буква *i* может в разных традициях иметь в себе 0, 1 или 2 точки; но при этом украинская *ї* является другой буквой. Отсюда следует, что графическая идентичность – недостаточное основание для отождествления букв разных традиций, как и графическое различие – для их различения. Практически это означает, что СПП может различать несколько сущностей, которые при визуализации в конкретной традиции могут выглядеть одинаково.

Б. Различение символов, необходимое только для какой-то одной традиции, влияет на общий алфавит символов и заставляет пересматривать реализацию уже представленных традиций. В качестве примера можно привести “вопрос о земле”. Из разделяемого всеми принципа “внутреннее представление не должно зависеть от почерка” следует, что слово “земля” должно быть одинаково зафиксировано в электронных представлениях текстов рукописи XI века и настоящего сборника. Однако добавление в корпус текста традиции, в которой различаются буквы “земля круглая” и “земля хвостатая”, потребовало расширить алфавит – и то, что до этого было вопросом почерка, теперь стало разницей букв. Но для уже созданных однородных корпусов это внутренне необоснованно и поэтому неубедительно. Напрашивается вывод, что место в искомой СПП необходимо связывать не с формой буквы “земля”, а только с самим её свойством дополнительности по отношению к “основной букве”.

В. Но в разных традициях такая дополнительность к одной и той же букве графически разная. Рассматривая одну традицию, мы могли пользоваться *достаточным критерием различения* букв: если графемы встречаются в одном тексте в разной роли, то их нужно различать. Этот критерий сам по себе расплывчат и противоречив, но применение его к буквам из разных традиций просто стирает различия: сербская “другая н”, согласно критерию, не отличается от якутской “другой н”, значит, их графическое различие становится вопросом почерка. Очевидно, такой логичный вывод не устроит тех практиков, которые хотят иметь право набрать одним-единственным шрифтом Arial Unicode сербский текст с комментариями на якутском языке.

Подобные противоречия, возникающие именно в связи со смешением традиций, требуют фундаментального разговора о том, какие сущности должны различаться и каково должно быть их соотношение с видимыми графическими элементами, чтобы иско-

мая СПП была внутренне непротиворечивой и устойчивой к добавлению новых традиций. Это требует тщательного формулирования принципов. Язык для такого разговора удобно строить на основе используемых в ряде однородных корпусов транслитерационных представлений, в которых свои обозначения получают, например, такие “сущности”, как ударение и “заглавность” буквы, и которые позволяют легко вводить новые обозначения как для фонологических и грамматических аспектов, так и просто для графических различий.

2. С другой стороны, подготовка междисциплинарного разговора состоит в построении и системном структурировании деятельностной схемы работы с текстами корпусов. Это поможет увидеть различие в актуальности и содержании отдельных требований к СПП: а) на разных этапах работы с корпусом, б) в корпусах на разных технических платформах, в) для разных задач, решаемых на материалах корпуса.

В результате становится ясно, что задача адекватной *визуализации* текста – не единственная и даже не самая важная задача, для которой создаётся корпус. Например, для задач контекстного поиска и сравнительного сопоставления фрагментов более удобными оказываются формы *хранения* текста, способные гибко удерживать не только необходимые для визуализации детали, но и разные аспекты обобщённого содержания. Для задач, связанных с изучением языка и графики конкретной традиции, наоборот, становится важным отразить все детали, причём даже более подробно, чем на это способно любое фиксированное кодировочное решение. Это разделение может быть эффективно проведено в ситуации метакорпуса, когда наиболее подробная информация о тексте хранится и обрабатывается процедурами первичного однородного корпуса, а процедуры межкорпусного обмена получают от первичных корпусов необходимые фрагменты в уже преобразованной форме. Но это в корне по-другому ставит вопрос о способе стандартизации представлений письменности.

Результатом междисциплинарного обсуждения должна быть корректная постановка задачи о различении символов *при их хранении* в корпусе, на решение которой, не связанное с деталями реализации, могли бы опираться все технологические платформы. Было бы полезно, если не необходимо, совместно выработать такую форму представления сводной информации о традициях, которая бы, с одной стороны, вбирала в себя текущие достижения языкознания обо всех необходимых для создания СПП аспектах каждой традиции, а с другой стороны, служила бы основой для технического задания IT-специалистам в каждой из существующих платформ. Простой перечень необходимых изображений символов ещё не может служить такой формой, являясь лишь частным, производным от неё результатом.

Problems of representing letters in a diachronic corpus or metacorpus

Alexander V. Kovalenin

A. P. Ershov Institute of Informatics Systems of the Siberian Branch
of the Russian Academy of Science, Novosibirsk, Russia

To devise a character representation system for texts in a diachronic corpus is a more intricate problem than to represent characters of a certain written tradition. Both philological and technical difficulties appear. Proposed solutions often ignore this difference and implicitly have in mind a certain operational environment. Consideration of these issues is important for discussing the correct target setting and proper allocation of philologists’ and IT specialists’ responsibilities.