АВТОМАТИЗАЦИЯ ПРОЦЕССА ПЕРЕВОДА ЭТИМОЛОГИЧЕСКОЙ ТРАНСКРИПЦИИ РУССКОГО СЛОВА В ЕГО ОРФОГРАФИЧЕСКУЮ ФОРМУ

Огаркова Н.В.

Язык является иерархически организованной структурой. Каждый из уровней иерархии содержит свойственные ему базовые элементы, например, морфемы, слова, предложения. Но сами по себе базовые элементы говорят только о морфологии данного уровня, а не о его структурной организации, вследствие чего появляется необходимость вводить правила для работы с этими базовыми элементами. Совокупность таких базовых элементов и правил на каждом уровне иерархии образует свою подсистему. Для исследования и максимальной формализации каждой подсистемы необходимо создавать программный инструментарий, реализующий процесс ее изучения путем выявления и проверки правил анализа и синтеза. Фиксация правил анализа и синтеза приводит к созданию анализаторов и синтезаторов для каждого уровня иерархии [1].

Подсистемы языка не являются изолированными друг от друга, и структура подсистемы одного уровня может влиять если не прямо, то косвенно на структуру подсистемы другого уровня. Поэтому исследование взаимосвязи между подсистемами, с одной стороны, может дать дополнительную информацию для формализации самих подсистем, а с другой стороны, оно является необходимостью, поскольку требуется обеспечить переход от базовых элементов одного уровня к базовым элементам другого уровня.

Одной из разработок Научно-методического центра по компьютерной лингвистике ВГУ является программа синтеза русского слова, которая исходя из имеющегося набора морфем, пропущенных через языковые фильтры, выдает слова, записанные в этимологической транскрипции, предложенной А.А. Кретовым [2]. Эта транскрипция имеет нетрадиционную форму. В качестве примера можно привести данные из "Большого морфемного русского словаря" А.А. Кретова.

Орфографическая форма	Этимологическая транскрипция
медлительность	мьд = ьл = И = тел = ьн = ост _ ь
мерзлый	мЬрз = Н =л _ ый
местоположение	мет = т _ о по –лог = и = Е = н = ій _ е
мечтательница	мек = ьт = А = тел = ьн =ик _ а

С одной стороны, это может затруднить работу пользователя, а с другой, синтезированные слова могут являться исходными данными для следующей подсистемы компьютерной программы анализа текста, которая использует информацию о морфемном составе слова для синтеза предложений. Таким образом, появляется необходимость преобразовывать слово из его этимологической транскрипции в более привычную для человека орфографическую форму. Важность этимологической транскрипции слова заключается в его глубинном представлении, в более четком и правильном проведении границ между морфемами, что дает возможность перехода от наблюдения и описания к пониманию, объяснению и интерпретации.

Выявление и формализация совокупности правил перехода от этимологической транскрипции к его орфографической форме требуют от исследователя глубокого погружения в предметную область. Для оформления этой совокупности в виде системы, определяющей связи между правилами и их взаимное влияние друг на друга, необходимо создание программного инструментария, который не только позволит освободить исследователя от рутинного процесса накопления информации, но и снимет вопрос о трудоемкости ее обработки.

И.А. Кретовым была предпринята попытка решить данную проблему. За основу схемы переходов была взята продукционная модель представления знаний, то есть модель, основанная на правилах типа: ***если*** *(условие),* ***то*** *(действие). Условием* в этой схеме являлся результат поиска заданной подстроки в слове, *действием* - замена найденной подстроки на другую подстроку. Каждому правилу приписывался определенный номер, который

определял порядок применения правил. Таким образом, все правила были выстроены в линейную систему и, соответственно, каждое правило применялось только один раз. Эта модель описана в нашей публикации [2]. Наполнение указанной модели происходило на основе созданного А.А. Кретовым "Большого морфемного русского словаря", который содержит около 165 тыс. слов в орфографической форме с указанием их этимологической транскрипции. Процесс наполнения происходил следующим образом: каждое из слов, записанных в этимологической транскрипции, проходило через машину вывода - программу, управляющую перебором правил. Если преобразование прошло успешно, то есть на выходе была получена орфографическая форма данного слова, то перечислялись те правила, которые были применены. В случае неудачи указывалось, что набор правил не полон или порядок применения этих правил не верен, и специалист предметной области должен внести изменения в исходный набор правил, чтобы получить положительный результат.

В ходе исследований А.А.Кретова выявился ряд недостатков модели И.А.Кретова. Во-первых, ему приходилось перечислять все частные случаи для каждого из правил. Это привело к тому, что количество правил сильно возросло (на данный момент более 3 тыс.). Поскольку все правила были выстроены линейно, то пользователю при формулировке нового правила приходилось определять его местоположение в системе, что достаточно затруднительно при наличии большого количества правил. При этом возникали ситуации, при которых неправильное месторасположение приводило к сбоям в работе правил, ранее работавших корректно. Также было выявлено, что в тех случаях, когда два слова с разной орфографической формой имели одинаковую фонематическую транскрипцию, на выходе машины вывода можно было получить только одну из орфографических форм (для получения другой формы требовалось вручную менять местами ряд правил). И, как следствие, набор, включающий в себя все возможные правила, нельзя считать полной системой, поскольку машина вывода не всегда выдает корректные результаты. Также одним из недостатков явилось то, что линейность модели потребовала повторения одного и того же правила в наборе более одного раза для корректной работы машины вывода, так как для некоторых слов такое правило должно применяться раньше, а для некоторых – позже. Нельзя не отметить и достоинств системы И.А. Кретова. К ним можно отнести простоту формулировки правила: во-первых, не требуется написания специальной программы, которая позволяет вводить данные, а во-вторых, специалист предметной области, который формулирует правила, может быть знаком только со стандартными средствами MS Office; от него не требуется освоения какого-либо другого программного инструментария. Во-вторых, линейность модели, вследствие которой появляется так много недостатков, в то же время дает одно неоспоримое преимущество – значительно увеличивается скорость работы машины вывода.

С учетом всех недостатков и преимуществ модели И.А.Кретова, ведется разработка нового программного комплекса, предназначенного для решения задачи перехода от этимологической транскрипции слова к его орфографической форме. Процесс разработки состоит из трех этапов.

Основной задачей первого этапа, как и всего программного комплекса, является разработка структуры одного правила. В новой системе перехода каждое правило представляет собой список заменяемых значений (подстрок). При попытке применить данное правило к некоторому слову (к его этимологической транскрипции) в первую очередь осуществляется поиск заменяемой подстроки из списка в этом слове. Только в том случае, если такая подстрока найдена, становиться возможной проверка условия (условий) для замены, поскольку проверяемое слово уже можно разделить на три составляющие: левая проверяемая часть, заменяемая подстрока и правая проверяемая часть. Условие для замены, как видно, состоит из двух частей: левой и правой. Для удобства формулировки правил пользователю предоставляется возможность связывать их при помощи логических операций «И», «ИЛИ» и «НЕ», т.е. строить логическое выражение для проверяемых частей. Если условие выполняется, становится возможным применение одного из нескольких вариантов замен. Таких вариантов может быть несколько, поскольку всю систему правил можно условно разделить на три подсистемы: церковно-славянскую, русскую и общую. Для каждого из вариантов замены должно быть указано, к какой подсистеме он относится. Чтобы уменьшить количество правил в новой системе перехода, в первую очередь в формулировке условия замены вводиться понятие маски, которое позволит пользователю не указывать каждый частный случай, а ссылаться на некоторый набор элементов. Кроме этого пользователю предоставляется возможность формулировать более сложные условия, которые содержат основные логические операции («И», «ИЛИ», «НЕ»). Поскольку структура правила становится более сложной, требуется создание программного инструментария, который позволит пользователю не только формулировать правила, сохранять их, просматривать и распечатывать, но также предоставит дополнительные возможности. Например, в момент формулировки, пользователь может делать выборку из существующего набора правил, чтобы проверить правильность его формулировки [5]. Таким образом, в конце первого этапа будет получен набор переформулированных правил, причем количество их значительно уменьшится за счет усложнения структуры правила.

Вторым этапом разработки новой модели перехода от этимологической транскрипции русского слова к его орфографической форме является реализация машины вывода. Поскольку в модели И.А. Кретова было выявлено значительное количество недостатков и неудобств, обусловленных линейностью системы, машина вывода новой модели должна работать независимо от того, в каком порядке расположены правила перехода, то есть должна

осуществлять полный обход дерева решений. При реализации машины вывода надо соблюдать особую осторожность, поскольку возможно зацикливание в процессе применения правил. Чтобы избежать этого, во время прохождения слова через машину вывода придется хранить дополнительную информацию о том, какие правила уже были применены к заданному слову и в каком состоянии находилось оно само во время их применения. Признаком конца работы машины вывода в данном случае должна являться невозможность применения ни одного правила к заданному слову. Таким образом, для некоторого заданного слова может быть получено несколько вариантов его преобразования, причем только часть из них будут являться правильными. Если для каждого слова, пропущенного через машину вывода, сохранять номера правил, которые были применены к нему, то можно получить цепочки, в которых имеется порядок следования правил, то есть определены отношения порядка: какое-то правило было применено раньше, а какое-то позже. Если воспользоваться информацией "Большого морфемного русского словаря" А.А. Кретова, в котором задана этимологическая транскрипция слова и его орфографическая форма, то все цепочки можно будет разделить на положительный материал и на отрицательный. Это осуществляется следующим образом: если в ходе применения некоторой цепочки правил к слову, заданному в виде этимологической транскрипции, была получена соответствующая орфографическая форма, то цепочка является правильной, и, как следствие, относится к положительному материалу, в противном случае – к отрицательному. Полученный таким образом набор цепочек является результатом второго этапа создания системы правил. Помимо вышеизложенного, еще одной дополнительной возможностью, облегчающей работу пользователя, должна стать процедура, которая позволит прогонять слово через машину вывода не только в автоматическом режиме, но и пошагово. В таком режиме работы пользователь сможет и обнаружить ошибки, допущенные в процессе формулирования правил, и контролировать процесс преобразования слова.

Целью третьего этапа разработки модели перехода от этимологической транскрипции слова к его орфографической форме является улучшение скорости работы машины вывода. Это становится возможным, если частично упорядочить правила. Например, разбить их на группы, внутри которых возможно применение правил в любом порядке, и установить строгий порядок следования групп друг за другом. Частичное или полное упорядочение правил может быть достигнуто в ходе исследования материала, полученного на втором этапе разработки. К этому материалу, как к положительному, так и к отрицательному, будут применены статистические методы, которые позволят создать некоторую надстройку над набором правил, которая будет сохранять информацию о порядке следования правил друг за другом. Структура надстройки еще не разработана, поскольку она явным образом зависит от данных, полученных на втором этапе: либо она даст возможность проверять условия применения правил последовательно, либо будет отличаться сложной иерархической структурой. Отметим, что создание и заполнение надстройки над правилами должно осуществляться автоматически, поскольку изменение набора правил в ходе лингвистических исследований может повлечь за собой и изменение статических оценок, на основе которых была построена и просчитана модель. Именно поэтому пользователю должен быть предоставлен программный инструментарий, благодаря которому можно будет пересчитывать статистические оценки и автоматически менять наполнение модели.

Структура программного комплекса, предназначенного для решения задачи перехода от этимологической транскрипции слова к его орфографической форме, представлена на рис. 1.

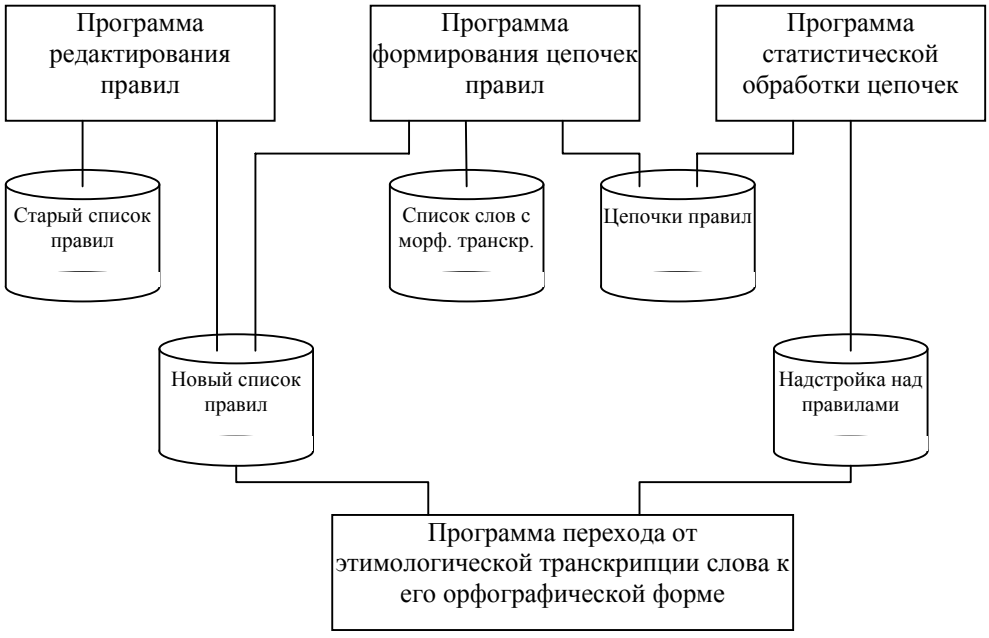

Рис.1 Схема программного комплекса, предназначенного для решения задачи перехода от этимологической транскрипции слова к его орфографической форме

В заключение отметим некоторые возможные варианты дальнейшего использования и развития программного комплекса. Одной из важных задач компьютерной лингвистики является не только переход от этимологической транскрипции слова к его орфографической форме, но и наоборот, от орфографической формы слова к его этимологической транскрипции. Решение этой задачи пока не формализовано, и существует несколько вариантов алгоритмов для такого преобразования. Один из вариантов - это использование системы правил из рассмотренного программного комплекса. Существует вероятность создания специальной процедуры, которая по заданной орфографической форме слова, путем специального перебора правил, будет выдавать один или несколько вариантов этимологической транскрипции этого слова. Но это предположение должно быть проверено программно и оценена эффективность такого способа преобразования. Другой способ использования этого программного комплекса – изучение самой системы правил перехода. Возможно, для этого потребуется создание дополнительных процедур, которые позволят оценить востребованность правила, то есть насколько часто данное правило используется, не является ли данное правило частным случаем какого-либо другого и т.д.

Литература

1. Воронина И.Е., Кретов А.А. Метод последовательной фильтрации при разработке лингвистического обеспечения информационных процессов // Математическое обеспечение ЭВМ: Межвуз. сб. науч. тр., Вып. 1. Воронеж: Изд-во ВГУ, 1999. С. 17-21.
2. Кретов А.А. Четвертая транскрипция // Аванесовские чтения: Международная научная конференция: Москва, 14-15 февраля 2002 г.: Тезисы докладов /Под ред. М.Л. Ремневой и М.В. Шульги. – М.: МАКС Пресс, 2002, С.157-160.
3. Кретов А.А. Теоретические и практические аспекты создания морфемного словаря // Вестник ВГУ. Серия лингвистика и межкультурная коммуникация, 2002, № 2, с.55-61. – 1,75 п.л.
4. Огаркова Н.В. Моделирование процесса перехода от глубинной формы к поверхностной на примере русской фонетики // Математическое обеспечение ЭВМ: Межвуз. сб. науч. тр., Вып. 4. Воронеж: Изд-во ВГУ, 2002. С. 128-134.
5. Огаркова Н.В. Моделирование условия в задаче перехода от этимологической транскрипции слова к его орфографической форме // Информатика: проблемы, методология, технологии: Материалы второй региональной науч.-метод. конф. Воронеж: Изд-во ВГУ, 2003. С. 117-119.