Йорма Луутонен

Некоторые наблюдения по поводу построения компьютерного морфологического анализатора марийского языка.

(Originally published in: К. Н. Сануков et al. (ed.) *Congressus Decimus Internationalis Fenno-Ugristarum. Йошкар-Ола 15.08.-21.08.2005. Pars IV. Linguistica.* Йошкар-Ола: Марийский государственный университет, 2008. Pp. 480-484.)

Материалом данной статьи послужили наблюдения, касающиеся морфологии марийского языка и научного описания этой морфологической системы. Наблюдения были сделаны нами в процессе проектирования компьютерной модели, анализирующей словоформы марийского языка. Часть проблем, представленных в статье, связана с разработкой такого типа компьютерной модели вообще, часть касается общего характера марийской морфологической системы, часть представляет отдельные проблемные марийские словоформы. Также рассмотрим некоторые проблемы, касающиеся литературной нормы лугово-восточного литературного языка. Целью нашей статьи является не разрешение, а изложение проблем, возникших и возникающих в процессе построения компьютерной модели морфологии марийского языка.

Разрабатываемый нами марийский морфологический анализатор основывается на двухуровневой компьютерной модели (по-английски ”Two-level Morphology”), созданной Киммо Коскенниеми и опубликованной им в 1983-ем году. Данная модель является универсальной моделью автоматического морфологического анализа и генерирования словоформ. Компьютерная программа, использующая двухуровневую модель, работает следующим образом: сначала пользователь вводит в компьютер словоформы, после чего компьютерная программа анализирует данные словоформы и выдаёт результат анализа пользователю. Кроме анализа компьютерная программа способна генерировать словоформы. Для этого пользователю необходимо ввести в компьютер словарную форму слова и символы грамматических категорий. На основе заданной информации программа производит запрашиваемую пользователем словоформу. В 90-ые годы двадцатого столетия компьютерные программы, используемые при разрабатывании двухуровневых моделей разных языков, получили значительное развитие. Первоначальной модели Коскенниеми

была разработана альтернативная модель. Создателями данной модели являются Кенет Р. Бесли и Лаури Карттунен. К изданной ими в 2003 году книге ”Finite State Morphology” прилагается диск CD-ROM, содержащий программу, необходимую при разработке двухуровневой модели. Данная программа используется нами при построении компьютерной модели морфологии марийского языка.

Первый прототип марийского двухуровневого морфологического анализатора был создан нами в 1985-ом году. После опубликования книги Бесли и Карттунена нами была возобновлена работа над разработкой марийского анализатора. Нашей задачей является создание компьютерной системы, способной осуществлять автоматический анализ любого текста на лугово-восточном марийском литературном языке. На данный момент разработана часть модели, описывающая морфологию существительгого и глагола. Нашей следующей задачей является ввод в компьютерную модель парадигм других частей речи.

Двухуровневый морфологический анализатор состоит из двух компонентов. Первый компонет представляет собой формализованный словарь, содержащий слова и аффиксы языка. Формализованный словарь также содержит информацию о возможном порядке следования аффиксов в слове. Второй компонент морфологического анализатора представляет собой множество морфофонологических правил, которые функционируют между двумя уровнями модели, то есть между обыкновенными поверхностными словоформами и их абстрактными лексическими изображениями.

При построении двухуровневой модели вся морфологическая информация языка кодируется для того, чтобы компьютерная программа могла использовать данную информацию при анализе и генерировании словоформ. После ввода закодированной информации в компьютер нами проверяется, способна ли компьютерная программа на основе полученной информации производить анализ заданных ей словоформ, а также осуществлять генерирование, то есть производить грамматически правильные словоформы. В построении двухуровневой модели используется принцип двунаправленности: возможен и анализ, и генерирование. Как нам известно, данный принцип не используется при проектировании других компьютерных программах, осуществляющих автоматический анализ морфологии марийского языка.

При кодировании существующей грамматической информации для модели, а также при использовании данной модели компьютерными программами нами проверяется полнота, последовательность и точность данной информации. При обнаружении недостатка в грамматическом материале разработчик двухуровневой модели обрабатывает те правила, которые недостаточно освещены в грамматической литературе или формулирует

совершенно новые. При этом в качестве вспомогательного инструмента мы использовали компьютерный корпус марийских текстов, размещённый в сети Интернет. Главной лингвистической целью построения марийского двухуровневого анализатора является составление более полного и точного, по сравнению с прежним, описания морфологии марийского литературного языка.

Оценивая традиционные грамматики, составителю двухуровневой модели необходимо иметь ввиду, что многие детали, кажущиеся при построении компьютерной модели недостаточно освещёнными в грамматической литературе, являются незначительными, если данные грамматики используются не компьютером, а человеком. Говорящий или пишущий человек обладает обильной информацией об окружающем мире, о языке вообще и о языковой ситуации. Он имеет способность оформлять свою речь на основе такой информации. Компьютерные программы обычно не обладают такими способностями. Поэтому, например, при генерировании словоформ компьютерная программа применяет правила без ограничений, и, следовательно, часто производит большое количество форм, кажущихся совсем неожиданными и непригодными. Таким образом, двухуровневая модель легко становится сверхпродуктивной. Поскольку компьютерная модель основывается на правила традиционных грамматик, то отмеченная сверхпродуктивность касается также правил, изложенных в этих традиционных грамматиках. Сверхпродуктивность правил, представленных в грамматиках нельзя заметить, так как никто, кроме компьютера, не использует данные правила механически, без ограничения. При разработке двухуровневой модели марийского языка большая часть времени была потрачена на ограничение продуктивности данной модели. В качестве обобщения констатируем, что большая часть информации, содержащейся в человеческом языке, вероятно, связана с ограничениями правил на разных языковых уровнях, а также с тенденциями и ограничениями комбинирования лексических единиц.

А теперь поподробнее остановимся на некоторых проблемах, касающихся марийского языка. Обратимся к литературным нормам марийского языка. Разрабатываемая нами двухуровневая модель основывается на кириллической орфографии, что затрудняет нашу работу. Первая, правда, лингвистически тривиальная проблема, связана с использованием кириллических букв *е* и *я* при изменении формы слова. Например, в спряжении по лицам глагола *шӱяш*’сгнить’: *шӱям, шӱят, шӱеш, шӱйына, шӱйыда, шӱйыт* буквами *е* и *я* обозначаются звукосочетания *йе*, *йа*. В данных формах глагола *шӱяш* компонент *й* относится к основе слова, а гласные *а* и *е* к суффиксу. Таким образом, использование кириллицы затрудняет сегментацию подобных словоформ на морфемы. При построении компьютерной

модели техническим разрешением данной проблемы стало искусственное создание двух вариантов основы глагола, один из которых содержит компонент *й* (*шӱй-*) а другой не содержит (*шӱ-*). К данным двум основам прикрепляются разные суффиксы. Таким образом созданная система не совсем соответствует морфемному анализу.

Вторая проблема, касающаяся орфографии, является с лингвистической точки зрения более значительной. Данная проблема связана с чередованием звонких и глухих консонантов. Правописание звонких и глухих согласных марийского языка не всегда соответствует их действительному произношению. С исторической точки зрения, орфографические правила не всегда соответствуют действительным фонетическим процессам. В марийском литературном языке в некоторых суффиксах (напр., суффикс комитатива *-ке/-үе*), а также во многих корневых основах (напр., *βӱт:βӱд-* ’вода’) при написании всегда используется звонкая согласная, тогда как в традиционном произношении данная согласная представляет глухой вариант. К примеру, при написании слов *вӱд: вӱдна* ’вода: наша вода’ используется звонкий консонант *д*, который при произношении данных слов является глухим: *βӱт:βӱtna*. Другой пример: при написании слова *кид:кидге* ’рука: вручную’ используется также звонкий согласный, а при произношении глухой *kit: kitke*. (СМЯФ 148; Alhoniemi 1985: 33-34.) Отсюда следует , что морфофонологические правила, заложенные в основанную на кириллице двухуровневую модель, описывают большей частью орфографию марийского языка, чем реально происходящие в языке морфофонологические процессы.

Третья проблема, касающаяся норм литературного языка, состоит в том, что в нормативных грамматиках представлены полные парадигмы склонения и спряжения только беспроблемных слов. Слова, в склонении которых наблюдаются трудности, представлены в качестве примера лишь в нескольких формах. Во время совместной работы с информантами нами было замечено, что парадигмы спряжения некоторых обычных глаголов могут являться не совсем ясными для владеющих в совершенстве марийским языком. К таким проблемным словам относятся глаголы, основа которых состоит из консонанта и вокала (CV-), например, *муаш* ’найти’, *пуаш* ’дать’, *шуаш* ’доходить; достигать (I спр.); бросать (II спр.)’. Парадигмы данных глаголов являются вполне регулярными. Однако, в процессе спряжения внутри глаголов образуются сочетания гласной полного образования с редуцированной гласной *ы,* не характерные для многих диалектов марийского лугово-восточного языка, и поэтому являющиеся затруднительными. К примеру, *шуына* ’доходим; достигаем’ ср. *шуна* ’дошли; достигли’; *шуынем* ’хочу бросить’. (Ср. Alhoniemi 1985: 107-108.) Следует отметить, что в грамматиках марийского языка необходимо подробнее рассматривать проблемные глаголы,

чтобы грамматики могли в достаточной мере поддерживать нормы литературного языка, а пользователи могли правильно использовать данные формы.

Далее коснёмся проблемы, связанной с изменением порядка расположения морфем. Данное явление широко наблюдается в склонении существительных и значительно усложняет построение компьютерной модели. Возникновению данного явления способствует ряд причин. Первая причина – самостоятельный характер некоторых суффиксов, являющихся похожими на отдельные слова. Второй причиной является изменение порядка расположения падежного и притяжательного суффиксов в зависимости от падежа. Третья причина связана с возможностью притяжательных суффиксов располагаться до и после морфемы множественного числа, что в некоторой степени влияет на функцию притяжательного суффикса. В диалектах, являющихся основой марийского литературного языка, можно наблюдать разные тенденции в последовательности суффиксов. (Luutonen 1997: 149-156.) В двухуровневой модели словоподобные морфемы *шамыч* и *влак* относятся, с точки зрения их склонения, к словам. При составлении модели мы сталкиваемся с проблемами, когда пытаемся обычным способом установить порядок расположения суффиксов. Созданная модель производит множество грамматически невозможных форм. Причиной этому является то, что на основе установленных правил, определяющих порядок расположения морфем, один и тот же суффикс может встречаться в слове много раз. На самом деле такая многократная суффиксация возможна лишь в определённых условиях с притяжательными суффиксами, реже с морфемами множественного числа и падежными окончаниями. По этой причине мы вынуждены сильно ограничить продуктивность морфологической модели. Для этого нами были созданы фильтры, удаляющие из двухуровневой модели нежелаемые словоформы.

Наибольшую сложность в формализации морфологии марийского языка представляет проблема подвижности притяжательных суффиксов и их многофункциональность. При склонении существительных подобные суффиксы могут выражать притяжательное и определённо-указательное значения, могут служить вокативным элементом, а также присоединяться к инфинитивной форме глагола. Существует тенденция, когда притяжательный суффикс в определённой функции занимает определённую позицию в словоформе, однако, не существует абсолютного правила, описывающего порядок расположения притяжательного суффикса в определённом значении.

Далее сделаем замечания по поводу недостатков в описании некоторых явлений, наблюдаемых нами в грамматиках марийского языка. Большая часть замечаний связана с тем, что в грамматиках марийского языка не дана достаточно точная информация о

дистрибуции некоторых суффиксов. Примером являются энклитические частицы. В грамматиках данные частицы обычно просто перечисляются, а также приводятся немногочисленные примеры. Такими суффиксами являются *-ла, -лай, -я, -ян, -огыла, -можо, -мочет, -ыс, -с*. Иностранному исследователю, не носителю языка, приходится выяснять дистрибуцию данных элементов с помощью информантов, а также используя электронный корпус марийских текстов.

В грамматической литературе также неясно дана информация о дистрибуции алломорфов некоторых суффиксов. Примером является алломорф притяжательного суффикса 3 лица единственного числа, оканчивающийся на согласный *ж*, прикреплённый к суффиксу дательного падежа. Например, в формах *кышаж(ы)лан* 'на его след' и *корныж(ы)лан* 'на его путь, дорогу' по опросу трёх информантов, луговых марийцев, возможны два варианта: с гласной и без гласной *ы*. Тогда как информант, говорящий на восточном диалекте марийского языка, использует короткую форму, т.е. без гласной *ы*: *кышажлан* и *корныжлан*. В то же время все информанты, независимо от диалекта, были одного мнения по поводу слова *тер* 'сани', имеющего лишь одну форму: *тержылан*, а не **терыжлан*. До сих пор нами не найдено правдоподобного объяснения, почему слова *кыша* и *корно* склоняются в данном случае иначе чем слово *тер*. (Ср. СМЯМ 57-59; Alhoniemi 1985: 74, 76.)

При разработке двухуровневой модели большую разъяснительную работу потребовали такие случаи, когда к слову, оканчивающемуся на сибилянт, присоединяется суффикс, начинающийся с сибилянта. Первый случай касается присоединения суффиксов направительного падежа *škV* и местного падежа *štV* к словам, оканчивающимся на *Vš* (*V* обозначает гласный звук). В этом случае одновременно с полной формой может использоваться её укороченный вариант. Например, комбинация слова *мучаш* 'конец' и суффикса направительного падежа *шке* может быть представлена или в полной форме *мучашышке* или в виде укороченного варианта *мучашке*, аналогично комбинация этого же самого слова и суффикса местного падежа *ште* может быть представлена или в полной форме *мучашыште* или укороченной *мучаште*. В укороченных формах наблюдается процесс слияния сибилянтов основы и суффикса. (Alhoniemi 1985: 45-46.) В процессе разработки модели нами были протестированы комбинации других корневых основ, оканчивающихся на сибилянт и суффиксов, начинающихся с сибилянта. Данное тестирование было сделано для того, чтобы определить, является ли процесс слияния сибилянтов в подобных комбинациях обычным явлением. Для этого нами было составлено морфофонологическое правило, описывающее данный процесс. Согласно этому правилу

компьютерная модель осуществляет генерирование укороченных форм следующим образом: основа *мучаш* + *шт* ’притяж. суф. 3 лица множ. числа’ → **мучашт* (ср. полная форма *мучашышт* ’их конец’); *корно* ’дорога’ + *еш* ’суф. обстоят. падежа’ + *шт* → **корнешт* (ср. *корнешышт* ’на их дорогу’). По мнению информантов, такие сокращённые варианты являются неграмматическими. Таким образом, процесс слияния сибилянтов на границе двух морфем касается, очевидно, только формы направительного и местного падежей.

Второй случай представляют комбинации корневой основы, оканчивающей на *Vš*, и притяжательного суффикса 3 лица единств. числа, на стыке которых появляется соединительный гласный. Согласно Алхониеми (1985: 74), возможны два варианта формы слова *мучаш* ’конец’: *мучашыже* и *мучашше*. Наши информанты одобрили оба варианта. Обе формы употребляются также в фольклорных текстах (Kokla 1963: 45). По мнению Тужарова (1987: 45-46) форма *мучашыже,* в которой используется соединительная гласная, является литературной нормой, однако, в художественной литературе, а особенно в поэзии, встречаются формы, подобные форме *мучашше,* не использующие соединительной гласной. В процессе разработки компьютерной модели мы заметили, что возможность отсутствия соединительной гласной касается скорее тех случаев, когда первый сибилянт представлен в конце основы слова. Если первый сибилянт относится к суффиксу, использование соединительной гласной является обязательным. Если к концу слова, заканчивающегося суффиксом обстоятельственного падежа *(е)ш*, присоединим притяжательный суффикс без использования соединительной гласной, образуются формы, являющиеся, по мнению информантов, неграмматическими. К примеру, присоединив к основе *муно* ’яйцо’ суффикс обстоятельственного падежа *еш* и притяжательный суффикс 3 лица единств. числа *же* , получаем форму *мунешыже*, но не **мунешше.* Обобщая два выше изложенных случая сокращения форм (*мучаште, мучашке; мучашше*), можно утверждать, что подобные сокращения используются только в некоторых строго определённых контекстах. Другими словами, такое сокращение форм не является общим морфофонологическим процессом, оказывающим воздействие на язык.

Марийский язык обычно считается относительно несложным, имеющим чёткую морфологическую структуру языком, однако, особенности марийского языка, описанные выше, затрудняют разработку двухуровневой модели в той степени, что компьютерная модель, являющаяся результатом данной работы имеет довольно многосложный характер.

Литература

Alhoniemi, Alho 1985: *Marin kielioppi.* Helsinki: Suomalais-Ugrilainen Seura. 172 c.

Beesley, Kenneth R. & Karttunen, Lauri 2003: *Finite state morphology.* Stanford: CSLI Publications. 510 c.

Kokla, Paul 1963: *Possessiivsufiksid mari keeles.* Väitekiri filoloogiakandidaadi teadusliku kraadi taotlemiseks. Tallinn. 307 c. (Рукопись.)

Koskenniemi, Kimmo 1983: *Two-level morphology: A general computational model for word-form recognition and production.* Publications of the Department of General Linguistics, University of Helsinki. No. 11. 160 c.

Luutonen, Jorma 1997: *The variation of morpheme order in Mari declension.* Suomalais-Ugrilaisen Seuran toimituksia 226. Helsinki: Suomalais-Ugrilainen Seura.

СМЯМ = Пенгитов, Н. Т. и другие (ред.): *Современный марийский язык. Морфология.* Йошкар-Ола 1961: Марийское книжное издательство. 324 с.

СМЯФ = Галкин, И. С. и другие (ред.): *Современный марийский язык. Фонетика.* Йошкар-Ола 1960: Марийское книжное издательство. 162 с.

Тужаров, Г. М. 1987: *Грамматические категории имени существительного в марийском языке.* Йошкар-Ола: Марийское книжное издательство.142 с.