ВВОДНЫЙ ДОКЛАД КРУГЛОГО СТОЛА «МАРИЙСКИЙ ЯЗЫК, РОДСТВЕННЫЕ ФИННО-УГОРСКИЕ ЯЗЫКИ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ – ЧТО ДАЛЬШЕ?»

Джереми Мос Бредли (Jeremy Miss Bradley)
Исследователь отделения финно-угристики
Венского университета

ПРОЕКТ ВЕНСКОГО УНИВЕРСИТЕТА «MARI WEB PROJECT» И ЕГО МАРИЙСКИЙ МОРФОАНАЛИЗАТОР

Сегодня Internet позволяет распространить любую информацию в огромном пространстве. Но народам, которые не имеют возможности полноценно использовать свои языки в компьютерных технологиях, приходится быть в изоляции своего региона. Среди таких языков и марийский язык.

Сейчас иностранцам очень трудно заниматься марийским языком. Дело не в трудности марийского языка, марийский язык не труднее других языков. Проблема в недостатке материалов. Материалы, которые есть, очень трудно найти, их слишком мало, они слишком старые, а по-английски вообще ничего нет. Нам хотелось, чтобы и в Западной Европе, и в Японии, и в Южной Африке – во всем мире – была возможность заниматься этим красивым языком. Пока это только мечта.

Так родилась идея: создать сайт, на котором было бы всё, что надо для самообучения марийскому языку, так что человек, знающий английский язык, мог бы его изучать. Это сайт – www.mari-language.com. Проект по созданию марийского сайта является некоммерческим проектом под руководством финно-угорского отделения Венского университета.

Мы пишем самоучитель марийского языка на английским языке. Первая часть в триста страниц уже в интернете, вторая тоже будет. На сайте размещена информация о том, как работать на компьютере с марийским языком. В будущем также будет марийско-английский словарь. Это сейчас наша самая большая работа – пятьдесят тысяч слов, а у нас только три года. Но сейчас уже пишу на букву «Й», так что пока всё идет хорошо. До конца две тысячи тринадцатого года словарь должен быть в интернете.

В контексте темы конференции наш проект самый интересный. Это всё-таки морфологический анализатор марийского языка. Есть его демо-версия. Можете попробовать и сами убедиться, как он работает. В программу уже введены лексемы, просто нет переводов.

Что он делает? Например ...

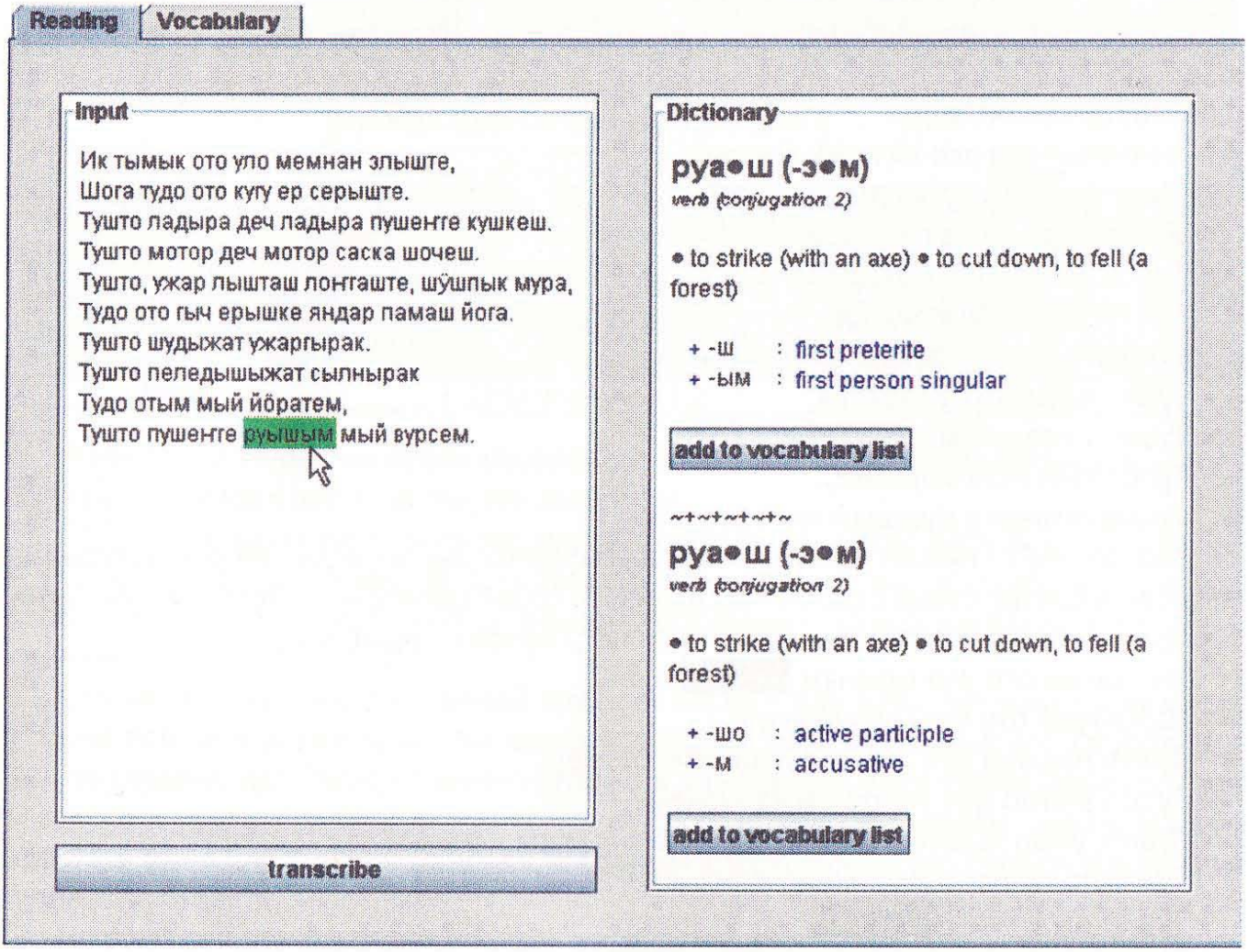
пӧрт	-ем	-влак	-лан	-ат
house	Px1Sg	Pl	Dat	Enc

'also to my houses'

пӧртем-влакланат → [ANALIZATOR] → пӧрт 'house'
+ Possessive Suffix First Person Singular '-ем'
+ Plural Marker '-влак'
+ Dative Case '-лан'
+ Enclitic '-ат'

Если вводим "пӧртем-влакланат" – "также к моим домам", то в результате получаем то, что видете сейчас – лексему и список суффиксов. Но, что мы можем делать с этим? Очень много.

Наша первая программа, для которой мы написали анализатор в первую очередь – это Помощник для чтения. Это программа для учеников, изучающих марийский как иностранный язык. Программа работает следующим образом:



Окно Помощника для чтения состоит из двух частей: первая содержит текст, а другая – морфологический анализ слов текста. При нажатии на слово в тексте его морфологический анализ появляется в правой части окна.

Очень важно иметь ввиду, что это недетерминистский анализатор. Работает только с морфологией, синтаксического анализа ещё нет. Поэтому в результате содержатся все возможные морфологические интерпретации. Ученик должен сам определить правильный в данном контексте вариант (здесь – второй).

Но: данная компьютерная программа даёт ученику быстрый доступ к словарю. Кроме того, нажав на суффикс из списка, он автоматически попадает в раздел учебника, в котором рассказывается об этом суффиксе.

Эта программа может также использоваться для создания глоссария незнакомых слов – только тех слов, которые ученику незнакомы.

Демо-версия сейчас содержит только лексемы. Например, программа знает, что есть слово „сылне", она знает, что это прилагательное, но не знает, что это означает „красивый". Без законченного словаря данная компьютерная программа не может быть опубликована, значит – примерно до 2013 года.

Но если бы мы могли иметь марийско-русский словарь в подходящем цифровом формате, то можно было бы опубликовать данную программу для учащихся, изучающих марийский язык в школах и вузах Республики Марий Эл и других регионов проживания марийцев. Насколько нам известно, такие словарные базы имеются, поэтому считаю, что надо обсудить возможность скорейшей реализации такого проекта. Это было бы первый электронный инструмент для изучения марийского языка как марийскоязычными, так и русскоязычными учащимися.

Но другие возможности, которые дает анализатор несомненно интересны. Прежде всего это возможности его использования в качестве проверочника орфографии (спеллера). Вот пример.

Ик тымык ото уло мемнан елыште,
Шога тудо ото кугу ер серыште.
Тушто ладыра деч ладыра пушеҥге кушкеш.
Тушто мотор деч мотор саска шочеш.
Тушто, ужар лышташ лоҥгаште, шӱшпык мура,
Тудо ото гыч ерышке яндар памаш йога.
Тушто шудыжат ужарграк.
Тушто пеледышыжат сылнырак
Тудо отым мый йӧратем,
Тушто пушеҥге руышым мый вурсем.

Анализатор понимает только правильные формы. Значит: если программа понимает слово, в соответствии с своей лексикой и со своей схемой марийской морфологии, программа говорит: «правильно», а если нет, говорит: «ошибка».

Ик тымык ото уло мемнан елыште,
Шога тудо ото кугу ер серыште.
Тушто ладыра деч ладыра пушеҥге кушкеш.
Тушто мотор деч мотор саска шочеш.
Тушто, ужар лышташ лоҥгаште, шӱшпык мура,
Тудо ото гыч ерышке яндар памаш йога.
Тушто шудыжат ужарграк.
Тушто пеледышыжат сылнырак
Тудо отым мый йӧратем,
Тушто пушеҥге руышым мый вурсем.

К сожалению, это только демо-версия: интерфейс ещё плохо и слишком медленно работает, и пока нет очень важной интеграции с текстовыми программами MS Word и OpenOfficeOrg.

У нашей программы имеется еще одна возможность: Анализатор помогает нам создавать современный словарь. Если орфопроверочник (спеллер) читает текст и не понимает слово, то говорит: «ошибка». И в данном случае мы используем программы несколько иным образом: если она не понимает слова, то говорит: "я не знаю это слово. Может его надо внести в словарь?"

Таким образом мы анализируем современный корпус и находим много новых слов, которые используются в современном марийском языке, но которые отсутствуют в старых словарях. Особенно если мы находим слово в разных материалах (и в детской книге, и в тексте о технологии), тогда понимаем, что это слово важно (эти слова отмечены в программе зеленым цветом)

Word	Context	Interpretation
сайт	Тендам www.mari-language.com сайтышке пагален ӱжына!	site
блог	Пытартыш вашталтыш нерген уэмдалт шогышо Блогышто да мемнан Facebook лышташыште пален налын кертыда.	blog
финно-угристика	[...] Вена университетысе финно-угристика пӧлкажын [...]	Finno-Ugric studies. Maybe it should be угристике?
ончыкшым	Ончыкшым, моло туныктыш тӧнеж-влак дене кылдалтын. [...]	strange form ончыко used all over the place
аудио-материал	Ешартыш семын тыгакак тиде тунемме книгалан аудио-материалым ямдылыме.	audio materials
веб-сайт	Марий веб-сайтым ышташ [...]	website
операционный	[...] операционный системе ден [...]	operational
компьютерный	Ты ужашыште йылме символ-влакым компьютерный системыште, [...]	computer
кодироватлаш (-ем)	Кодироватлаш лийше [...]	to encode
кодироватлыме	Тиде кодироватлыме системе почеш [...]	encoded
кодироватлымаш	Символ-влакым кодироватлымаш	encoding
битан	[...] 5 битан символ [...]	bit (adjective)

А мы так не только находим русские и английские слова, но и марийские: – всех в старых словарях нет, а так мы это заметили.

Так мы находим не только заимствованные русские или английские слова, но и собственные марийские. Вот, например, какие слова, отсутствующие в старых словарях, мы нашли при помощи данной программы: йӧршӧ, касвелне, туге-гынат, садикте, тошкешташ, туддеч.

Сейчас, мы переходим к самой трудной части – как работает наш анализатор. У меня есть авторские права на данный продукт. Но я не против его свободного использования для "малых" финно-угорских языков (т.е. всех кроме финского, эстонского и венгерского). Я отказываюсь от своих прав и вы можете делать все что хотите, я готов для этого поделиться исходным кодом, можно это сделать и в рамках будущего проекта.

В своей основе анализатор имеет словарь и схему марийской морфологии.

да	со	and
и•мн'е	no	horse
йӱ•штӧ	ad	cold
нерге•н	po	about
ни•не	pr	they
толаш	vb1	to come
толаш	vb2	to steal
ше•рге	ad	expensive
шерге•	no	comb
э•ркын	av	slowly

Словарь программы около так – это упрощение. Очень важно, что программа знает, какие слова, чтобы знать, что может делать с словом. Глаголы может спрягать, другие слова может склонять, у послелога может быть притяжательный суффикс – мый денем, с мной – а падежа не будет, и так далее.

Словарь программы пока примитивен. Очень важно, что программа знает к какой части речи относится слово, для того, чтобы знать как слово может изменяться. Глаголы

могут спрягаться, другие части речи – склоняться, к послелогам может прицепляться притяжательный суффикс (денем – с мной), но не падежные суффиксы, и т.п.

Перевод важен Помощнику для чтения. Для орфопроверочника (спеллера) это не очень важно, орфопроверочнику не надо знать, что слово значит, надо только знать, что оно существует. Всем программам важно, чтобы были обозначены ударения и палатализация. Ударение важно, потому что оно влияет на образование других форм данного слова.

ше•рге > ше•ргым
шерге• > шерге•м

Например, безударный е в конце слова в винительном падеже заменится на ы. Ударный е в конце слова в винительном падеже сохранится. Ударение необходимо знать, чтобы знать что произойдет в винительном падеже, но орфография его не показывает. Поэтому оно специально маркировано.

/ne/ → не
/n'e/ → не
/nə/ → ны
/n'ə/ → ньы

Маркирование палатализации в марийском языке очень далеко от оптимального. Например, /ne/ и /n'e/ пишутся одинаково, но есть разница между /ne/ и /n'ə/. Для компьютерной программы это огромная проблема, когда в именительном падеже в конце слова стоит /n'e/. Мы знаем, что безударный е в винительном падеже должен превратиться в ы. Но когда /n'e/ превращается в /n'ə/, появляется мягкий знак.

и•мн'е > и•мньым
ни•не > ни•ным

Что значит „метод суффиксации"? В марийском языке различные суффиксы примыкают к различным основам разными способами. Например, притяжательный суффикс первого лица единственного числа обычно "-ем", а если в конце слова ударный а, это будет только –м. Если в конце слова неударный о, то о исчезает.

пӧрт	'house'	>	пӧртем	'my house'
ава	'mother'	>	авам	'my mother'
тумо	'oak tree'	>	тумем	'my oak tree'

Надо сказать: Хотя генерация марийской морфологии детерминистская, анализ таковым не является. Если, например, рассматриваем слово „юмем", мы не можем знать является ли для него базовая лексема „юмо" – „бог" или „юм" – „волосатик". „Юмем" может быть и „мой бог", и „мой волосатик". Детерминистский анализ марийской морфологии невозможен.

У другого суффикса будет другая методика. В нашем программе пять методов суффиксации.

	-влак (A)	-н (B)	-лан (C)	-ем/-эм/-м (D)	-ат/-ят (E)
пӧрт	пӧрт-влак	пӧртын	пӧртлан	пӧртем	пӧртат
тумо	тумо-влак	тумын	тумылан	тумем	тумат
ава	ава-влак	аван	авалан	авам	ават
изи	изи-влак	изин	изилан	изиэм	изиат
...	...	...	...	...	...

Если наш суффикс попадает в группу D, то мы знаем, что он ведет себя как вышеупомянутый притяжательный суффикс первого лица единственного числа, и, таким образом, программа знает, как работать с ним.

```
Nominal stem +  [comp][gen][poss][plur][case-g1][p3][enc][red]

Nominal stem +  [comp][gen][poss][plur][case-g3][p3][enc][red]

Nominal stem +  [comp][gen][plur][poss][case-g1][p3][enc][red]

Nominal stem +  [comp][gen][plur][poss][case-g3][p3][enc][red]

Nominal stem +  [comp][gen][plur][case-g2][poss][p3][enc][red]

Nominal stem +  [comp][gen][plur][shILL][p3][enc][red]

Nominal stem +  [comp][gen][plur][case-g3][poss][p3][enc][red]

Verbal stem +  {tmp}[comp][p3][enc][red]

Verbal stem +  {infger}[comp][poss][p3][enc][red]

Postposition stem + [poss][p3][enc][red]

Any stem + [p3][enc][red]
```

Это упрощение актуальных схем – наша программа также работает со словообразовательными суффиксами, но здесь в графике их нет, так как это бы было бы слишком сложно, нет времени говорить о них.

Знаю, что и без словообразовательных суффиксов объяснение сложно, поэтому хочу привести пример – анализ слова „пӧртем-влакланат" – «также к моим домам».

Первый вопрос возникающий у анализатора: "пӧртем-влакланат – это слово?". Он сверяется со словарем, но такого слова (в таком виде) в словаре нет. Таким образом анализатор делает вывод, что либо это слово, написанное с ошибками, либо оно должно иметь один или несколько суффиксов, а значит должно подвергнуться анализу.

Когда анализатор начинает анализировать это слово, он ничего о нем не знает. Является ли оно глаголом? Или это послелог? Ничего. По этой причине он не знает какие суффиксы искать. Он перебирает опробывая все что имеет и находит несколько вариантов интерпретации.

пӧртем-влаклана	-т	пӧртем-влаклан	-ат
?	Px2Sg	?	Ind2Sg
'your пӧртем-влаклана'			'you пӧртем-влаклан'

пӧртем-влаклан	-ат	пӧртем-влаклан	-ат
?	Ind3Pl	?	Ind2Sg
'your пӧртем-влаклана'			'also пӧртем-влаклан'

Например,-т может быть притяжательным суффиксом второго лица единственного числа. Тогда может быть наше слово "твоя пӧртем-влаклана"?

А –ат может быть суффиксом второго лица единственного числа глагола первого спряжения или третьего лица множественного числа глагола второго спряжения. Тогда это значит, "ты пӧртем-влаклан-ешь", или "они пӧртем-влаклан-ут"?.

Это всё абсурд, но анализатор этого не знает – он же ничего не знает о слове, которое обрабатывает, поэтому должен испробовать все варианты. Он находит эти абсурдные варианты, и ещё другие абсурные варианты, но в конечном итоге находит правильный ответ – "-ат" является энкилитикой означающей "также".

Теперь программа имеет много вариантов – абсурдные, и один правильный вариант. Она ещё не может знать, какой из них правильный. Анализ должен быть продолжен.

Анализатор обрабатывает абсурдные, но ничего не находит. Со временем доходит до правильного варианта – пӧртем-влаклан с энклитикой оначающей «также». Пока мы назовем этот вариант гипотезой.

Анализатор сверяет «пӧртем-влаклан» со словарем, не находит. Но марийский язык – агглюнативный язык, поэтому суффиксов может быть больше, чем один. Анализатор ищет другие суффиксы, но уже не так неосведомленно, как ранее:

Nominal stem + [comp][gen][poss][plur][case-g1][p3][~~enc~~][~~red~~]

Nominal stem + [comp][gen][poss][plur][case-g3][p3][~~enc~~][~~red~~]

Nominal stem + [comp][gen][plur][poss][case-g1][p3][~~enc~~][~~red~~]

Nominal stem + [comp][gen][plur][poss][case-g3][p3][~~enc~~][~~red~~]

Nominal stem + [comp][gen][plur][case-g2][poss][p3][~~enc~~][~~red~~]

Nominal stem + [comp][gen][plur][shILL][p3][~~enc~~][~~red~~]

Nominal stem + [comp][gen][plur][case-g3][poss][p3][~~enc~~][~~red~~]

Verbal stem + {tmp}[comp][p3][~~enc~~][~~red~~]

Verbal stem + {infger}[comp][poss][p3][~~enc~~][~~red~~]

Postposition stem + [poss][p3][~~enc~~][~~red~~]

Any stem + [p3][~~enc~~][~~red~~]

В соответствии с нашей гипотезой последний суффикс – энклитика. Это значит, что далее не надо искать энклитики. Наши схемы позволяют иметь в слове только одну энклитику.

Появляется новая гипотеза: -лан может быть суффиксом дательного падежа (датива). Это значит, что теперь мы знаем, что имеем дело с основой слова.

~~Nominal stem + [comp][gen][poss][plur][case-g1][p3][enc][red]~~

Nominal stem + [comp][gen][poss][plur]~~[case-g3][p3][enc][red]~~

~~Nominal stem + [comp][gen][plur][poss][case-g1][p3][enc][red]~~

Nominal stem + [comp][gen][plur][poss]~~[case-g3][p3][enc][red]~~

~~Nominal stem + [comp][gen][plur][case-g2][poss][p3][enc][red]~~

~~Nominal stem + [comp][gen][plur][shILL][p3][enc][red]~~

Nominal stem + [comp][gen][plur]~~[case-g3][poss][p3][enc][red]~~

~~Verbal stem + [tmp][comp][p3][enc][red]~~

~~Verbal stem + [infger][comp][poss][p3][enc][red]~~

~~Postposition stem + [poss][p3][enc][red]~~

~~Any stem + [p3][enc][red]~~

Уже с новой гипотезой анализатор вновь обращается к словарю: есть ли в словаре основа слова (номинал) "пӧртем-влак?". Нет. Надо продолжать поиски. Но теперь вариантов меньше: только категории, которые появляются слева от самой левой группы суффиксов, которые должны быть учтены в рамках нашей гипотезы. Наш результат должен совпасть по крайней мере с одной схемой

Анализатор работает и наконец находит следующую гипотезу:

пӧрт n [poss][plur][case-g3][enc]

Nominal stem + [comp][gen][poss][plur][case-g3][p3][enc][red]

Пӧрт – основа слова (номинал). Первый суффикс, "ем" – притяжательный суффикс первого лица единственного числа. Второй суффикс, "-влак" – обозначает множественное число. Третий суффикс, "-лан", суффикс дательного падежа. И четвёртый суффикс, "-ат" – энклитика. Это совпадает с одной из наших схем. Анализатор обращается к словарю: есть ли в словаре основа слова (номинал) пӧрт? Есть! Обозначает "дом". Наша гипотеза верна.

Здесь был найден только один результат. Как я сказал, их может быть больше.

Подробную информацию о нашей деятельности и контакты участников проекта вы можете найти на сайте www.mari-language.com.