Found only at the end:
йю рч уя ұы цы ще юы юі

Found only in the middle:
аа ае аө ау ач әг әқ бб бв бг бд бж бз бк бп бт бф бц бч бш бъ бэ вб вв вг вд вк вн вф вш вя гб гг гд гж гз гк гө гс гт гх гш ғб ғд ғз ғл ғн ғо ғі дб дг дғ дд дк дл дм дн дп дс дц дш дъ еа еә еғ ее еи ек еө еұ еү еч ещ еі еэ жб жғ жд жк жн жс жт жш зб зв зг зғ зж зз зк зқ зн зө зп зр зс зт зх зш зъ зю иә иғ ик иө иц ищ иэ йа йб йв йг йғ йж йз йқ йө йұ йх йц йш йэ кг кд кж кз кқ км кп кф кц кч кш қб қг қд қе қж қк ққ қл қм қн қп қр қс қт қф қх қш лб лв лг лғ лж лз лм лн лө лф лх лц лч лш лэ мв мг мғ мд мж мз мк мқ мл мф мх мш мь нб нв нғ нқ нл нм нп нф нх нц нъ нэ оа оә ое ои оө оу оұ оү оы оі пб пв пд пж пк пқ пм пф пх пц пч пш пю рғ рж рз рө рр ръ ря сб сг сд сж сз сқ ср сш ся тб тг тғ тд тж тз тк тқ тн тп тф тх тц тш тэ тя уғ уж уө уұ уү уц уэ ұб ұх ұщ ұю уа үә уе үф үх ұы фв фг фк фс фш фы фя хб хв хг хм хн хс хт хх хш хы хь цг цд цз цк цл цм цт цц чж чк чм чт чч шб шг шғ шд шж шқ шс шш шю ща щи ъе ъю ыа ыв ыг ые ыи ыо ыұ ых ыц ыч ыя іг ід іж іи іо іх ьв ьг ьл ьо ьп ьр ьх ьц ьч ьш ьэ ью юа юж юи юо юц юч юш юэ яб яв яғ яц

Found only at the beginning:
бь ву вы вь гә ғү дз дө дх дь дэ дю жә зә зұ зү кә кй кұ кэ кю кя мр нө өд пә пъ рэ съ сэ уй үқ хұ цх шв шү шх щұ эг эж эо эу эх эш юг

It is very important to know item distributions of vowel sounds in a word as in Kazakh language the accent falls on a final syllable. It considerably simplifies and reduces diphone base. For example, the sound **О** meets only at the beginning of a word and all diphones with sound participation **О** have no shock analogs.

Vowel sounds meeting at the beginning and in the middle of a word sound almost equally that also does diphone base of more compact.

Conclusion: Further development of the program involves automating the generation of other phonetic units (allophones, Trifonov, syllables). The criterion can serve not only phonetic (segment) text properties, but also communicative and syntactic component (for punctuation): for example, you can specify the proportion of interrogative sentences, or sentences containing nonfinal syntagma (by commas). Obviously, the efficiency of selection depends on the parameters of the reference body text: the more complete and varied it is, the more informative and compact it will be obtained on the basis of textual material.

It is also planned to automate the analysis of statistics of occurrence in the text of the speech units of different levels: phonemes, allophones, syllables, sound sequences of words, the development of automatic phonetic transkriptor Kazakh speech text to speech synthesizer Kazakh.

References

1. Shelepov V.U. Leksii o raspoznavanii rechi. – Donetsk: IPSHI «Nauka i osvita», 2009. – 196 p.

УДК 81’33

МНОГОФУНКЦИОНАЛЬНАЯ МОДЕЛЬ ТЮРКСКОЙ МОРФЕМЫ: БАЗОВЫЕ ФОРМАЛИЗМЫ

Сулейманов Джавдет Шевкетович *д.т.н., директор, Институт прикладной семиотики Академии наук Республики Татарстан, Казанский федеральный университет.*
E-mail: dvdt.slt@gmail.com
Гатиатуллин Айрат Рафизович *к.т.н., зав.отделом, Институт прикладной семиотики Академии наук Республики Татарстан. E-mail: agat1972@mail.ru*

Статья является продолжением описания многофункциональной модели тюркской морфемы, которая включает структурно-функциональную модель морфем, базу данных

модели и информационно-программную оболочку, технологический инструментарий, для заполнения базы данных. В статье рассматриваются базовые формализмы, которые включают выражения для алломорфов, морфем и семантем тюркских языков. Многофункциональная модель может быть использована, как ресурсная база для программных продуктов, осуществляющих компьютерную обработку тюркских языков, информационно-справочная система и инструментарий для исследований ученых-тюркологов, в частности, для сравнительного анализа тюркских языковых единиц.

Ключевые слова: многофункциональная модель тюркской морфемы, морфема, алломорф.

MULTIFUNCTIONAL COMPUTER BASED LINGUISTIC MODEL: FORMALIZMES

Dzhavdet Suleymanov *Tatarstan Academy of Sciences Institute of Applied Semiotics, Kazan Federal University E-mail: dvdt.slt@gmail.com*
Ayrat Gatiatullin *Tatarstan Academy of Sciences Institute of Applied Semiotics, E-mail: agat1972@mail.ru*

This article is a continuation of the description of the multi-functional model of Turkic morpheme, which includes structural and functional model of morphemes, the database model and the information and the program shell, the technological tools to fill the databases. The article discusses the basic formalisms which include expressions for allomorphs, morphemes and semantemes of Turkic languages. The multifunctional model can be used as a resource base for the software which carries out the computer processing of the Turkic languages; information and reference system and tools for researches of Türkologists, in particular, for the comparative analysis of the Turkic language units.

Keywords: multifunctional model of Turkic morpheme, morpheme, allomorph.

Введение

В последние 25 лет наблюдается значительное увеличение количества разработок, а также публикаций в области компьютерной обработки тюркских языков [1, 6, 7]. Этому способствует как интерес к информации на тюркских языках в Интернет-пространстве, так и научно-исследовательская и прикладная активность носителей языка по компьютерной поддержке тюркских языков. Как показывает анализ имеющихся работ, наименее развитым и представленным в публикациях является отсутствие соответствующего формального инструментария, адекватной спецификации и семиотических (лексико-грамматических и семантических) моделей, отражающих структурно-функциональные особенности тюркских языков. Очевидно, целый ряд вопросов, возникающих у разработчиков программного инструментария для работы с тюркскими языками, связанных с неопределенностью понятий, многозначностью языковых единиц, могли бы разрешаться уже на этапе формального описания моделей.

Морфологическая и синтаксическая близость между тюркскими языками позволяет описывать соответствующие языковые явления схожими лингвистическими моделями и использовать для их обработки практически один и тот же программный инструментарий. Очевидно, в основе своей на 70-80 и более процентов они являются общими для всех тюркских языков, как при разработке структуры и функционала электронных корпусов языков, грамматических анализаторов, так и машин поиска и машинных переводчиков. А это означает, что возможно создать единую модель, наиболее полно отражающую лингвистические особенности тюркских языков.

Авторы статьи считают, что одним из таких объединяющих механизмов может стать многофункциональная модель тюркской морфемы. В данной статье предлагаются базовые формализмы этой модели.

1. Многофункциональная модель

Многофункциональная модель тюркской морфемы включает структурно-функциональную модель морфем [2], базу данных модели и информационно-программную оболочку, технологический инструментарий, для заполнения базы данных.

Способы применения многофункциональной модели:

- в качестве ресурсной базы для программных продуктов, осуществляющих компьютерную обработку тюркских языков.
- в качестве информационно-справочной системы, содержащей практически полную информацию о тюркских языковых единицах – морфемах.
– в качестве инструментария для исследований ученых-тюркологов, в частности, для сравнительного анализа тюркских языковых единиц.

Использование в качестве подобной ресурсной базы именно модели морфем обусловлено исключительной значимостью морфологического языкового уровня при обработке естественно-языковых текстов. Это особенно актуально для языков агглютинативного типа с богатой морфологией, к которым относятся все языки тюркского семейства.

Зададим базовые формализмы модели.

Пусть L_i – языки тюркской группы.

Например:

L_1 – татарский язык, L_2 – казахский язык, L_3 – турецкий язык и т.д.

Обозначим Σ_i – алфавит языка L_i. $a_{(i,j)} \in \Sigma_i$ - символ языка L_i.

Например, алфавит языка L_1 (татарского языка):

Σ_1 = {‘аәбвгдеёжҗзийклмнңоөпрстуүфхһцчшъыьэюя’}

Расширим алфавит языка L_i пустым символом (пробелом) e = {‘ ’}.

$\Sigma_i' = \Sigma_i + \{e\}$ – (расширенный) аналитический алфавит языка L_i

Тогда словоформа языка L_i:

$w_{(i,k)} = (a_{(i,1)}\ a_{(i,2)}\ a_{(i,2)}\ ..\ a_{(i,n)})$ – k-е слово языка L_i, $a_{(i,j)} \in \Sigma_i$; $a_{(i,j)} \neq e$

Например:

$w_{(1,k)}$ = ‘баралар’ – k-е слово татарского языка

Множество всех возможных слов языка L_i обозначим Σ_i^*.

2. Алломорфы

Рассмотрим обозначения морфологических единиц языка. В.А. Плунгян в своей работе [4] определяет три модели морфологии:

1) элементарно комбинаторная;
2) элементарно процессная;
3) словесно парадигматическая.

Элементарно комбинаторная модель – это модель, основным инструментом которой является линейная сегментация. В языках с элементарно процессной моделью морфологии некоторые алломорфы рассматриваются как исходные, а другие — как производные, которые могут быть получены из первых путем применения различных операций типа «фонологических процессов». В словесно парадигматических моделях вообще происходит отказ от морфемного членения при описании словоизменения. Именно словоформа и оказывается минимальной единицей грамматического описания в парадигматической модели.

Согласно этой классификации тюркские языки относятся к языкам элементарно-комбинаторного типа, комбинируемыми элементами являются морфы и алломорфы. Мельчук [3] дает им следующее определение:

Морфа (морф) – элементарный сегментный знак – минимальная сущность с одним и тем же морфологическим поведением, которое достаточно случайным образом связано с ее формой.

Алломорф - совокупность морфов одной морфемы, имеющих одинаковый фонемный состав.

В нашей модели будем использовать в качестве элементов комбинаторики – алломорфы:

$m_{(i,j)} = (a_{(i,1)}\ a_{(i,2)}\ a_{(i,2)}\ ..\ a_{(i,k)})$ – алломорф с номером j языка L_i, где $a_{(i,j)} \in \Sigma_i$

Плунгян [4] определяет словоформы как морфемный комплекс, между составными частями которого существуют особенно тесные связи — на порядок более тесные, чем между самими этими комплексами.

$w_{(i,k)} = (m_{(i,1)}\ m_{(i,2)}\ m_{(i,3)}\ ..\ m_{(i,p)})$ - слово языка L_i из p – алломорфов, где $p \geq 1$.

Если p=1, то словоформа состоит из одного алломорфа.

Каждый алломорф обладает рядом определенных свойств. Введем обозначения для этого набора свойств:

$C(m_{(i,j)}) = \{c^1(m_{(i,j)}), c^2(m_{(i,j)}), c^3(m_{(i,j)}), \ldots c^t(m_{(i,j)})\}$ – свойства алломорфа $m_{(i,j)}$

Одними из основных свойств алломорфов являются свойства, описывающие контекст этого алломорфа. Эти свойства делятся на два типа контекст алломорфа в пределах словоформы и контекст алломорфа за пределами словоформы.

Определим свойства 1 и 2 как свойства сочетания алломорфов в словоформе. Эти свойства определим множествами алломорфов, которые могут встретиться в словоформе слева и справа от искомого алломорфа:

$c^1(m_{(i,j)}) = \{ m_{(i,1)}, m\ m_{(i,2)}, \ldots m\ m_{(i,k)}\}$ – множество алломорфов, которые могут следовать в словоформе слева

$c^2(m_{(i,j)}) = \{ m_{(i,1)}, m\ m_{(i,2)}, \ldots m\ m_{(i,k)}\}$ – множество алломорфов, которые могут следовать в словоформе справа

Например:

c^2(‘га’) = {‘мы’, ‘мыни’, ‘дыр’, ‘рак’, e}

c^2(‘ка’) = {‘мы’, ‘мыни’, ‘дыр’, ‘рак’, e}

Наличие в свойстве символа e показывает, что алломорф может быть последним в словоформе. Если символ e отсутствует в свойстве, то этот алломорф не может быть последним в словоформе.

Например:

c^2(‘китап’) = {‘ка’, ‘тан’, ‘та’, …, e}

c^2(‘китаб’) = {‘ым’, ‘ың’, ‘ы’, ‘ын’, ‘ыбыз’, ‘ыгыз’}

Если $c^1_i(m_{(i,j)}) = \{e\}$, то этот алломорф может быть в словоформе только первым.

Если $c^2_i(m_{(i,j)}) = \{e\}$, то этот алломорф может быть в словоформе только последним.

Например: c^2(‘мы’) = {e}

Обозначим $Mi^* = \{m_{(i,j)}\}$ – множество всех алломорфов языка L_i.

$M_i^* = M_i' + M_i''$

$M_i' = \{m_{(i,j)}: c^1(m_{(i,j)}) = \{e\}\}$ – множество всех алломорфов языка L_i, которые могут стоять в словоформе только первыми.

$M_i'' = \{m_{(i,j)}: e\ !\in c^1(m_{(i,j)})\}$ – множество всех алломорфов языка L_i, которые не могут стоять первыми в словоформе.

3. Семантемы

У Мельчука языковая единица определяется тройкой: означающее, означаемое и синтактика. Соответственно, каждому алломорфу соответствует определённое значение, называемое семантемой: $m_{(i,j)} \rightarrow s_k$ или $m_{(i,j)}(s_k)$.

Введем обозначение $S^* = \{s_k\}$ – множество всех семантем языка, которые можно выразить отдельными алломорфами.

Множество семантем делится на подмножества, в соответствии с обозначаемыми концептами: $S^* = E + R$

E – множество всех сущностей языка; R – множество отношений между сущностями

$E = O + V + A$

O – множество объектов; A – множество действий; V – множество значений

Алломорфы, используемые для выражения одной и той же семантемы являются синонимами.

Например:

L_1: $m_{(1,j1)}(s_k)$ = ‘сандугач’; $m_{(1,j2)}(s_k)$ = ‘былбыл’

L_2: $m_{(2,j3)}(s_k)$ = ‘бұлбұл’

L3: $m_{(3,j4)}(s_k)$ = ‘bülbül’

4. Морфемы

Морфема в лингвистике определяется, как минимальная значащая часть слова, совокупность морфов (алломорфов), имеющих одинаковое значение и ряд других общих признаков:

$M_{(i,j)}(s_k) = \{m_{(i,j1)}(s_k)..\ m_{(i,jp)}(s_k): c^k(m_{(i,j1)}(s_k) = c^k(m_{(i,j2)}(s_k)\}$.

Этим другим свойством, согласно работам московской фонологической школы, является свойство регулярного чередования [3].

В каждом из тюркских языков эти множества чередований свои:

$G(L_i) = G_i = \{G_{(i,1)}, G_{(i,2)}, \ldots G_{(i,t)}\}$

Например:

G_1 = {{‘г’,’к’}; {‘д’,’т’}; {‘л’,’н’}; {‘а’,’ә’}; {‘у’,’ү’}; {‘б’,’п’}; … } – чередования в татарском языке.

Подставив требование чередования к определению морфемы, получаем следующую формулу:

$M_{(i,j)}(s_k) = \{m_{(i,j1)}(s_k)..\ m_{(i,jp)}(s_k): (a_{(i,j1)} \in m_{(i,j2)}) \,\&\, (a_{(i,j1)} \in m_{(i,j2)}): (a_{(i,j1)} = a_{(i,j2)}) \wedge$
$((a_{(i,j1)} \in G_{(i,p)}) \,\&\, (a_{(i,j2)} \in G_{(i,p)})\}$

Согласно этой формуле, одной морфеме принадлежат только алломорфы с чередованием фонем. Таким образом, алломорфы -к и -быз оба являются показателями предикативности 1 лица множественного числа, однако они не будут принадлежать одной морфеме, так как у них нет свойства чередования.

При p = 0, получается пустая морфема, а при p=1 морфема, состоящая из одного алломорфа.

Если следовать определению морфемы, приведенному выше, то морфема –[Г]А в следующем примере - это разные морфемы –[Г]А с одинаковым обозначением (знаком).

Мин урманга гөмбәгә ике сәгатькә барам. Я в лес за грибами на два часа пойду.

В нашей модели морфем предлагается считать одной морфемой те варианты, которые совпадают по правилам чередования и синтактике, но различаются по выражаемым семантемам (т.е. многозначная морфема).

Тогда согласно нашей модели –[Г]А в предыдущем примере, это одна и та же морфема –[Г]А, используемая для выражения значения разных семантем. А в словоформах барды-лар и алма-лар это алломорфы разных морфем $–ЛАр_1$ и $–ЛАр_2$, так как у них разные синтактики, то есть правила следования в словоформе.

Обозначим M_i^{**} - множество всех морфем языка L_i

5. Корневые и аффиксальные морфемы

Как отмечает Плунгян [4], различие между корневыми и аффиксальными морфемами представляется интуитивно очевидным, но в действительности оно с трудом поддается формализации. Одним из основных признаков, отличающих корень и аффикс, является то, что корень может образовывать словоформу, а аффикс - нет. Другая формулировка такова: словоформа может состоять только из одного корня, но не может состоять только из аффиксов.

Согласно морфологической классификации языков А.П. Володина [5] все языки делятся на две большие группы:

1. Языки, в которых представлена алтайская линейная модель словоформы – слово всегда начинается с корня. Алтайская модель имеет два вида:
1. алтайская 1: R+(m) - запрещена префиксация, композиция и инкорпорация (тюркские, эскимосский);
2. алтайская 2: (r)+R+(m) - запрещена префиксация, в некоторых языках (финно-угорские, корейский, японский) разрешена инкорпорация существительного.
2. Языки, в которых представлена американская линейная модель словоформы - слово не всегда начинается с корня. Американская модель также имеет три вида:
1. американская 1: (m)+(r)+R+(M) - айну, нивхский, чукотский, кетский, баскский, шумерский, многие языки американских индейцев;
2. американская 2: (m)+(r)+R+F+(m) - индоевропейские языки, F - флексия;
3. американская 3: (m)+R+(M) - ительменский, некоторые из языков американских индейцев.

Здесь R, r - корневые морфемы, M, m - аффиксы, Скобки означают, что данный элемент может быть представлен в конкретной словоформе более одного раза или вообще отсутствует.

Таким образом для тюркских языков $M_i' = \{m_{(i,j)}: c^1(m_{(i,j)}) = \{e\}\}$, определенное выше, совпадает с множеством корневых алломорфов.

Множество $M_i'' = \{m_{(i,j)}: e \notin c^1(m_{(i,j)})\}$ – совпадает с множеством аффиксальных алломорфов.

Аналогично алломорфам все морфемы имеют конечный набор свойств. Свойства морфем формируются из свойств алломорфов, из которых состоит морфема:

$C(M_{(i,j)}) = \{c^1(m_{(i,j)}), c^2(m_{(i,j)}), c^3(m_{(i,j)}), \ldots c^t(m_{(i,j)})\}$ – свойства морфемы $M_{(i,j)}$ языка L_i.

Совокупность всех тюркских морфем, их свойств, семантем и соответствий между морфемами и семантемами образует многофункциональную модель тюркской морфемы:

$MM = \{ M_{(i,j)}, C(M_{(i,j)}), s_k\}$

Заключение

В статье дается описание базовых элементов многофункциональной лингвистической модели тюркских морфем. Весьма конструктивным и продуктивным представляется использование данной многофункциональной и многоязычной модели тюркских морфем в качестве одного из центральных, ядерных, модулей в едином веб-портале для тюркских языков. Авторы статьи выражают также надежду, что данный проект послужит интеграции усилий ученых-тюркологов для расширения базы данных описаниями различных тюркских языков, что обеспечит эффективное использование многофункциональной модели в качестве технологического инструментария и межязыкового модуля в системах компьютерной обработки тюркских языков.

Список литературы

1. Труды Первой международной конференции «Компьютерная обработка тюркских языков». – Астана: ЕНУ им. Л.Н. Гумилева, 2013. – 345с.

2. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань: Фэн, 2003. – 220с.

3. Мельчук И.А. Курс общей морфологии. Т. IV. / Пер. с фр. Е.Н. Саввиной под общ.ред. Н.В. Перцова. – М.: Вена: Языки славянской культуры: Венский славистический альманах, 2001. – 584с.

4. Плунгян В.А. Общая морфология: Введение в проблематику: Учебное пособие. М.: Эдиториал УРСС, 2000. – 384с.

5. Володин А.П. Палеоазиатские языки // Языки мира М., 1996.

6. Proceedings of the International Conference on Turkic Language Processing (TURKLANG-2014). (Istanbul, November 6-7, 2014). - Istanbul: Özkaracan Matbaacılık-Bağcılarr, 2014. – 135 pp.

7. Proceedings of the International Conference "Turkic Languages Processing: TurkLang-2015". – Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. – 488 pp.

UDC 004.8+81.322.2

SENTIMENT ANALYSIS OF KAZAKH PHRASES BASED ON MORPHOLOGICAL RULES

Yergesh Banu, *L.N.Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: b.yergesh@gmail.com*
Sharipbay Altynbek, *L.N.Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: sharalt@mail.ru*
Bekmanova Gulmira, *L.N.Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: gulmira-r@yandex.ru*
Lipnitskii Stanislav, *United Institute of Informatics Problems NAS Belarus, Minsk, Belarus E-mail: lipn@newman.bas-net.by*

Abstract. Sentiment analysis of texts in natural languages is one of the fastest growing technologies of natural language processing. There is no any study has been carried out on the sentiment analysis Kazakh texts, to date. We examined the texts in the Kazakh language and identified the parts of speech, which determines the text sentiments. In this work we described the linguistic approach for sentiment analysis of phrases in the Kazakh language, based on the morphological rules.

Keywords: sentiment, tonality, sentiment analysis, Kazakh language, classification, production rules, morphological rules

АНАЛИЗ ТОНАЛЬНОСТИ КАЗАХСКИХ ФРАЗ НА ОСНОВЕ МОРФОЛОГИЧЕСКИХ ПРАВИЛ

Ергеш Бану, *Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: b.yergesh@gmail.com*
Шарипбай Алтынбек, *Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: sharalt@mail.ru*
Бекманова Гульмира, *Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, E-mail: gulmira-r@yandex.ru*
Липницкий Станислав, *Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь, E-mail: lipn@newman.bas-net.by*

Ключевые слова: сентимент, сентимент анализ, казахский язык, анализ тональности, классификация по тональности, морфологические правила.