УДК 519.688

DOI: 10.15827/0236-235X.121.172-176

Дата подачи статьи: 13.06.17

2018. Т. 31. № 1. С. 172–176

ИНСТРУМЕНТАРИЙ ДЛЯ ОПРЕДЕЛЕНИЯ ЛИНГВО-СТАТИСТИЧЕСКОЙ БЛИЗОСТИ ЯЗЫКОВ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ ТЮРКСКОЙ МОРФЕМЫ

А.Б. Альменова¹, аспирант, младший научный сотрудник, *almen_akmaral-baijan@mail.ru*

¹ Институт прикладной семиотики Академии наук Республики Татарстан,
ул. Баумана, 20, г. Казань, 420111, Россия

В статье предлагаются новый подход и инструментарий для лингво-статистического сравнения близости тюркских языков с использованием модели тюркской морфемы. Модель тюркской морфемы является структурированным описанием тюркских морфем, которое состоит из свойств морфем и ситуаций их проявления на всех языковых уровнях (фонологическом, морфологическом, синтаксическом, семантическом). Инструментарий для определения лингво-статистической близости тюркских языков реализован в виде веб-ресурса, который обрабатывает информацию о языке, представленную в БД модели тюркской морфемы.

Сравнение близости языков лингво-статистическими методами с использованием модели тюркской морфемы можно осуществлять для разных языковых уровней: морфологического, синтаксического, лексического. В данной работе описаны четыре метода для сравнения на морфологическом уровне: сравнение количества аффиксальных морфем, соответствующих грамматической категории, количества алломорфов, входящих в состав одной морфемы, слитности и раздельности написания морфемы, порядка следования аффиксальных морфем в словоформе.

В настоящее время БД модели заполнена для татарского, казахского, крымскотатарского, турецкого, узбекского языков, что позволяет проводить попарное сравнение их лингво-статистической близости.

Предложенный метод может быть использован при сравнении морфологий языков для получения информации о близости диалектов к литературному языку, а также о близости диалектов одного языка диалектам и литературному языку других языков. Эти методы и инструментарий применимы не только для тюркских, но и для любых близкородственных языков агглютинативного типа.

Ключевые слова: лингво-статистическое сравнение близости языков, тюркские языки, модель, тюркская морфема, БД.

Одним из направлений информационных технологий, активно используемых в лингвистических (в частности, типологических) исследованиях, являются лингво-статистические исследования. К задачам типологических исследований относятся исследования фонологического и грамматического строя языков, а также сопоставление фонологических, грамматических и семантических систем языков. Применение статистических методов в типологических исследованиях позволяет устанавливать количественные характеристики и изменения, вызывающие качественные преобразования языковых явлений. В последние годы наблюдается появление большого количества лингвистических БД, многие из них являются многоязычными. Это позволило проводить лингво-статистические исследования многоязычных лингвистических БД с информацией о разных языковых подсистемах.

Сравнительный анализ работ по лингво-статистическим исследованиям с многоязычными БД

Метод сравнения в лингвистике используется для решения как теоретических, так и практических задач. Основателем сопоставительного исследования языков был профессор Казанского университета Бодуэн де Куртанэ [1]. В его понимании сопоставительная лингвистика нацелена на выяв-

ление различий и сходств между двумя сравниваемыми языками, в том числе и родственными.

Проведем сравнительный анализ как лингвистических многоязычных БД, так и статистических исследований, осуществляемых с использованием многоязычных БД: Вавилонская башня [2], Upsid [3], WALS [4], APiCS [5], AfBO [6] и др.

К наиболее известным БД относится лингвистическая онлайн-база Atlas of Pidgin and Creole Language Structures (APiCS) [5]. В ней описано 76 языков, каждый из которых охарактеризован по 130 параметрам: фонетическим, лексическим и грамматическим.

Другая открытая лингвистическая база – AfBo (A world-wide survey of affix borrowing) [6] посвящена заимствованиям словообразовательных и грамматических показателей – аффиксов. На данный момент в этой базе описано 657 заимствованных аффиксов. В ней содержится информация о том, какие аффиксы заимствованы из другого языка, указывается язык заимствования.

Среди российских работ следует выделить публикацию [7]. В ней описывается сравнение близости языков с использованием современных информационных технологий, в число которых входят и нейронные сети. Авторы использовали метод сопоставления, основанный на наборах признаков с вычислением меры сходства или различия. Меры сходства ориентированы на двоичную логику, со-

гласно которой язык либо обладает данным признаком, либо нет. Для выявления и представления структуры объектов в пространстве признаков разработаны и используются различные методы ординации, то есть представления объектов на плоскости. В данной работе для исследований отобрано 48 языков, рассчитаны расстояния, выполнена ординация и рассчитаны ее ошибки. Все вычисления осуществлялись с помощью профессионального статистического пакета R [8], предназначенного для лингво-статистических исследований.

Для получения результата с использованием 48 выбранных языков авторами был применен метод нейронных сетей Кохонена с применением всех признаков и евклидова расстояния. Согласно этому подходу, языки, относящиеся к одной близкородственной группе, должны располагаться на ординационной плоскости компактно. В результате работы нейронной сети программа сгруппировала близкородственные языки на ординационной плоскости в соответствии с классификацией типологов.

По мнению авторов этой работы, БД с описанием модели тюркской морфемы также является эффективным инструментом, позволяющим проводить сравнительный анализ разных языковых подсистем тюркских языков. Модель тюркской морфемы представляет собой комплекс подмоделей, где модель каждого из тюркских языков является составной частью общей модели и содержит подмодели как корневых, так и аффиксальных морфем. Подробно структура модели тюркской морфемы описана в работах [9, 10].

В данной работе предлагается подход к проведению лингво-статистических исследований с использованием компьютерной модели тюркской морфемы.

Технология лингво-статистического анализа

Предлагаемые методы сравнения морфологической близости языков используют модель тюркской морфемы, поэтому для анализа были применены описания аффиксальных морфем.

Сравнение морфологической близости языков с использованием модели тюркской морфемы включает в себя четыре метода сравнения:

- по количеству аффиксальных морфем, соответствующих грамматической категории;
- по количеству алломорфов, входящих в состав одной морфемы;
- по слитности и раздельности написания морфемы;
- по порядку следования аффиксальных морфем в словоформе.

Для реализации этих методов написан программный инструмент, который выдает информацию о степени близости, используя данные из заполненной БД тюркской морфемы. В БД на

момент проведения вычислений представлены описания морфем для пяти языков тюркской группы: татарского, казахского, крымскотатарского, турецкого, узбекского, в совокупности для этих языков было представлено описание 398 аффиксальных морфем.

Рассмотрим методы сравнения. **Метод сравнения количества аффиксальных морфем, соответствующих грамматической категории**, исходит из того, что в разных тюркских языках для выражения одной и той же грамматической категории может использоваться разное количество морфем. Так, в татарском языке для выражения категории будущего времени глаголов используются две аффиксальные морфемы – -ЫР и -АчАК, а в казахском только одна морфема -ЫР. И наоборот, в казахском языке для выражения категории прошедшего времени глаголов используются три морфемы – -ДЫ, -ГАН и -Ып, а в татарском языке только две – -ДЫ и -ГАН.

В модели тюркской морфемы существует общая для всех тюркских языков таблица грамматических категорий и каждой аффиксальной морфеме присвоена своя грамматическая категория. Данная информация представляется в идентификационном аспекте модели тюркской морфемы (рис. 1). В процессе вычисления система получает из БД для каждого языка список морфем, после чего определяет их грамматические категории. По полученным данным вычисляется отношение пересечения грамматических категорий к общему числу грамматических категорий: $Similarity = \frac{C_c \times 2}{C_1 + C_2}$, где C_c – число

общих для обоих языков грамматических категорий; C_1 и C_2 – количество грамматических категорий каждого из языков.

Система выводит результат схожести в процентах, а также указывает, какое количество категорий из общего числа совпало.

Следующий метод заключается в **сравнении количества алломорфов, входящих в аналогичные морфемы** в разных тюркских языках. Алломорф – лингвистический термин, обозначающий вариант морфемы, которая может иметь разное произношение, но при этом не изменяет свое значение.

Например, татарская морфема -нЫкЫ состоит из алломорфов -ныкы и -неке, а казахская морфема -Нікі из трех алломорфов: -нікі, -дікі, -тікі. Соответственно, разница по количеству алломорфов для этой морфемы будет равна 1. Подобным образом суммируется разница для всех морфем в каждой из пар языков. Для подсчета этой разницы используется информация, представленная в таблице 1.

При данном методе сравнения система определяет количество алломорфов для каждой соответствующей морфемы в разных языках и сравнивает их. Результатом является отношение количества различий к общему числу алломорфов: $Similarity =$

$$= 1 - \frac{\sum_{i=0}^n (|a_{i1} - a_{i2}|)}{\sum_{i=0}^n (a_{i1} + a_{i2})}, \text{ где } a_{i1} \text{ и } a_{i2} - \text{количество алло-}$$

морфов i -й морфемы каждого из языков; n – число общих морфем сравниваемых языков.

Таблица 1

Сравнение алломорфов модели тюркской морфемы

Table 1

Comparison of allomorphs of the Turkic morpheme model

Татарский язык			Казахский язык		
Номер	Мор-фема	Алло-морф	Номер	Мор-фема	Алло-морф
01.2.024	-лЫI	-лы	02.2.024	-лЫI	-лы
01.2.024	-лЫI	-ле	02.2.024	-лЫI	-лі
			02.2.024	-лЫI	-ды
			02.2.024	-лЫI	-ді
			02.2.024	-лЫI	-ты
			02.2.024	-лЫI	-ті

Следует обратить внимание, что необходимо сравнивать алломорфы для каждой из морфем. При сравнении только общего количества алломорфов получится иной результат. Так, например, при сравнении числа алломорфов по морфемам в татарском и казахском языках получаем 116 различий, тогда как разница между общим числом алломорфов в указанных языках составляет всего 14.

Система выводит результат схожести в процентах, указывает количество различающихся алломорфов, общее число алломорфов в обоих языках, а также разницу между общим количеством алломорфов в сравниваемых языках (рис. 2, 3).

Третий метод заключается в подсчете разницы в слитности/раздельности написания морфем, выражающих одну и ту же грамматическую кате-

горию, в разных тюркских языках. Если в обоих языках морфемы пишутся одинаково слитно или одинаково раздельно, значение разницы написания будет равно 0. Например, частица -мЫI в татарском языке пишется слитно, а mI в турецком и МА в казахском языках пишутся раздельно. В итоге разница между татарским и казахским будет 1, а между казахским и турецким 0.

Информация о слитности/раздельности написания морфем также представлена в идентификационном аспекте модели морфем (рис. 1.).

Сравнение правил написания татарского и казахского языков показывает, что разница в слитности/раздельности написания аффиксальных морфем между татарским и казахским языками равна 2. Так, морфемы [-мЫI, -мЫни] в татарском языке пишутся слитно, а морфемы [МА, МАни] в казахском языке пишутся раздельно.

При данном методе сравнения система определяет общие грамматические категории в разных языках и сравнивает для каждой из них различие в слитном/раздельном написании соответствующих морфем. Результатом является отношение числа несовпадений слитности написания к общему числу соответствующих морфем языков: $Similarity = \frac{C_{diff}}{C_c}$, где C_{diff} – число несовпадений слитно-

сти написания; C_c – число общих для обоих языков грамматических категорий.

Система выводит результат совпадения в процентах, а также указывает, какое количество морфем из общего числа не совпадает по слитности написания.

Четвертый метод заключается в сравнении разницы в порядке следования аффиксальных морфем. Это связано с тем, что в разных тюркских языках правила следования аффиксальных морфем

Идентификационный аспект

Морфологический аспект

Морфонологический аспект

Синтаксический аспект

-ЛАр	01.2.001
-[Ы]м	01.2.002
-[Ы]ң	01.2.003
[-с][Ы][н]	01.2.004
-[Ы]быз	01.2.005
-[Ы]гыз	01.2.006
-мЫн	01.2.008
-сЫң	01.2.009
-быз	01.2.011

Обозначение морфемы

Цифровой идентификатор

Идентификатор для разметки корпуса

Название морфологической категории

Типологическое: английское

Типологическое: русское

Русское

Национальное

Написание

Сохранить

Добавить и сохранить

Удалить

Рис. 1. Идентификационный аспект

Рис. 1. Identification aspect

Таблица 2

Порядок следования морфем

Table 2

The order of morphemes

Татарский язык			Казахский язык		
Номер	Морфема	Порядок следования	Номер	Морфема	Порядок следования
01.1.030	-ЛЫК	-ЛАр	02.2.030	-ЛЫК	-ЛАр
01.1.030	-ЛЫК	-[Ы]м	02.2.030	-ЛЫК	-[Ы]м
01.1.030	-ЛЫК	-[Ы]н	02.2.030	-ЛЫК	-[Ы]н
01.1.030	-ЛЫК	-[с]Ы[н]	02.2.030	-ЛЫК	-[с]Ы[н]
01.1.030	-ЛЫК	-[Ы]ЫЫз	02.2.030	-ЛЫК	-[Ы]ЫЫз
01.1.030	-ЛЫК	-[Ы]ГЫз	02.2.030	-ЛЫК	-[Ы]НЫз
01.1.030	-ЛЫК	-[Г]А	02.2.030	-ЛЫК	-[Г]А
01.1.030	-ЛЫК	-[Г]АчА			-
01.1.030	-ЛЫК	-Дан	02.2.030	-ЛЫК	-ДАН
01.1.030	-ЛЫК	-ДА	02.2.030	-ЛЫК	-ДА
01.1.030	-ЛЫК	-н[Ы]	02.2.030	-ЛЫК	-Н[Ы]
01.1.030	-ЛЫК	-НЫң	02.2.030	-ЛЫК	-НЫң
		-	02.2.030	-ЛЫК	-Бен[ен]
01.1.030	-ЛЫК	-ДАгы	02.2.030	-ЛЫК	-ДАҒЫ
01.1.030	-ЛЫК	-НЫкЫ[н]	02.2.030	-ЛЫК	-Нікі[н]
01.1.030	-ЛЫК	-ДАЙ	02.2.030	-ЛЫК	-ДАЙ
01.1.030	-ЛЫК	-мы	02.2.030	-ЛЫК	МА
01.1.030	-ЛЫК	-мыни	02.2.030	-ЛЫК	МАНИ
01.1.030	-ЛЫК	-Дыр			-

могут отличаться. Например, в татарском и казахском языках морфема модальности в словоформе следует после морфемы предикативности, а в турецком наоборот.

Информация о порядке следования представлена в морфологическом аспекте модели тюркской морфемы (табл. 2).

В этом методе сравнения система определяет общие грамматические категории в разных языках, после чего из БД получает соответствующие морфемы для каждого из языков. Для каждой морфемы каждого из языков определяются морфемы, которые в словоформе следуют справа. Затем сравнива-

ется количество общих для обоих языков грамматических категорий полученных наборов последующих морфем. Результатом является отношение числа пересечений общих последовательностей к общему числу последовательностей: $Similarity = \frac{S_c \times 2}{S_1 + S_2}$, где S_c – число общих для обоих языков

последовательностей грамматических категорий; S_1 и S_2 – количество последовательностей грамматических категорий каждого из языков.

Система выводит результат близости в процентах, а также указывает, какое количество последо-

Сравнение близости морфологии языков

Язык №1
Крымскотатарский

Язык №2
Узбекский (latin)

Результаты сравнения

Метод	Результат	Доп.
сравнение количества аффиксальных морфем, соответствующих грамматической категории	81.97%	Совпало 100 морфем из общего числа 122
сравнение количества алломорфов, входящих в состав одной морфемы	50.97%	Различается 152 алломорф(ов) из 310 в обоих языках (227, 83), разница между кол-вом алломорфов равна 144
сравнение слитности и раздельности написания морфемы	100.00%	Различается написание 0 морфем из 50
сравнение порядка следования аффиксальных морфем в словоформе	0.00%	Совпало 0 из общего числа 0 последовательностей

Рис. 2. Окно с результатами сравнения близости крымскотатарского и узбекского языков

Fig. 2. The window with comparison results of the Crimean Tatar and Uzbek languages affinity

Сравнение близости морфологии языков

Язык №1
Татарский

Язык №2
Казахский

Результаты сравнения

Метод	Результат	Доп.
сравнение количества аффиксальных морфем, соответствующих грамматической категории	75.86%	Совпало 88 морфем из общего числа 116
сравнение количества алломорфов, входящих в состав одной морфемы	40.15%	Различается 158 алломорф(ов) из 264 в обоих языках (193, 71), разница между кол-вом алломорфов равна 122
сравнение слитности и раздельности написания морфемы	97.73%	Различается написание 1 морфем из 44
сравнение порядка следования аффиксальных морфем в словоформе	0.00%	Совпало 0 из общего числа 0 последовательностей

Рис. 3. Окно с результатами сравнения близости татарского и казахского языков

Fig. 3. The window with comparison results of the Tatar and Kazakh languages affinity

вательностей морфем соответствующих грамматических категорий из общего числа сравниваемых последовательностей сравниваемых языков совпадает (рис. 2, 3).

Заключение

В статье представлены лингво-статистические способы сравнения морфологической близости тюркских языков с использованием БД модели тюркской морфемы. Модель тюркской морфемы описывает морфологический уровень языка в связи с другими уровнями, соответственно, в модели представлена также синтаксическая и семантическая информация, которая может быть использована для сравнения иных типов близости тюркских языков, отличных от рассмотренных в данной статье.

Предложенный метод может быть использован для сравнения морфологий языков для получения информации о близости диалектов к литературному языку, а также о близости диалектов одного языка диалектам и литературному языку других языков. Например, восточный диалект татарского языка имеет ряд аффиксов/грамматических категорий, имеющих в соседних родственных язы-

ках: казахском (-Ып, -Гала), узбекском (-Ып), но отсутствующих в татарском литературном языке.

Литература

1. Бодуэн де Куртене И.А. О смешанном характере всех языков // Избран. тр. по общему языкознанию. М.: Изд-во АН СССР, 1963. Т. 1. С. 362–372.
2. The tower of babel. URL: <http://starling.rinet.ru> (дата обращения: 11.06.2017).
3. Database UPSID. URL: <http://menzerath.phonetik.uni-frankfurt.de/upsid> (дата обращения: 11.06.2017).
4. The World atlas of language structures online. URL: <http://wals.info/> (дата обращения: 11.06.2017).
5. The Atlas of Pidgin and Creole Language Structures Online. URL: <http://apics-online.info/> (дата обращения: 12.06.2017).
6. AfBo: a World-Wide Survey of Affix Borrowing. URL: <http://afbo.info/> (дата обращения: 12.06.2017).
7. Поляков В.Н., Савельев В.Д., Соловьев В.Д. Опыт применения методов интеллектуального анализа данных в компаративистских и типологических исследованиях (на материале созданной в ИЯ РАН БД «Языки мира») // КИИ-2006: тр. конф. М.: Физматлит, 2006. Т. 1. С. 217–224.
8. The R Project Foundation for Statistical Computing. 2006. URL: <http://www.R-project.org> (дата обращения: 12.06.2017).
9. Сулейманов Д.Ш., Гатиатуллин А.Р., Альменова А.Б., Баширов А.М. Многофункциональная модель тюркской морфемы // Филология и культура (Philology and Culture). 2016. № 2. С. 143–151.
10. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. Казань: Фэн, 2003. 345 с.

Software & Systems

DOI: 10.15827/0236-235X.121.172-176

Received 13.06.17

2018, vol. 31, no. 1, pp. 172–176

TOOLS FOR DETERMINING A LINGUISTIC-STATISTICAL AFFINITY OF LANGUAGES USING THE TURKIC MORPHEME MODEL

A.B. Almenova¹, Postgraduate Student, Junior Researcher, almen_akmaral-baijan@mail.ru

¹ Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan, Bauman St. 20, Kazan, 420111, Russian Federation

Abstract. The article proposes a new approach and tools for linguistic-statistical comparing of Turkic languages affinity using the Turkic morpheme model. This model is a structured description of Turkic morphemes, which consists of morpheme properties and their manifestation contexts at all language levels (phonological, morphological, syntactic and semantic) [3, 5]. The tools determining the linguistic-statistical affinity of Turkic languages is implemented as a web resource that processes information about the language represented in the morpheme model database.

Comparing the affinity of languages by linguistic-statistical methods using the Turkic morpheme model can be made at different language levels: morphological, syntactic, lexical. The paper describes four following methods: comparing the number of affixational morphemes corresponding to a grammatical category; comparing the number of allomorphs in one morpheme; comparing the joined-up and separate writing of morphemes; comparing the order of affixational morphemes within the word form.

Nowadays the model database is filled up for Tatar, Kazakh, Crimean Tatar, Turkish and Uzbek languages. It allows comparing linguistic-statistical affinity of the languages in pairs.

The proposed method can be used to compare language morphologies to obtain information on the affinity of dialects to a literary language, as well as to detect the affinity of one language dialects to dialects and a literary language of other languages. Our methodology and tools can be used not only for Turkic languages, but for any closely related languages of the agglutinative type.

Keywords: linguistic-statistical comparison of languages, Turkic languages, model, Turkic morpheme, database.

References

1. Boduen de Kurtene I.A. On the mixed nature of all languages. *Izbrannye trudy po obshchemu yazykoznaniyu* [Selected Works on General Linguistics]. Moscow, AN SSSR Publ., 1963, vol. 1, pp. 362–372 (in Russ.).
2. The tower of babel. Available at: <http://starling.rinet.ru> (accessed June 11, 2017).
3. Database UPSID. Available at: <http://menzerath.phonetik.uni-frankfurt.de/upsid> (accessed June 11, 2017).
4. The World atlas of language structures online. Available at: <http://wals.info/> (accessed June 11, 2017).
5. The Atlas of Pidgin and Creole Language Structures Online. Available at: <http://apics-online.info/> (accessed June 12, 2017).
6. AfBo: A World-Wide Survey of Affix Borrowing. Available at: <http://afbo.info/> (accessed June 12, 2017).
7. Polyakov V.N., Savelev V.D., Solovlev V.D. Experience in applying methods of data mining in comparative and typological studies. *KII-2006: tr. konf. [Proc. Conf. KII-2006]*. Moscow, Fizmatlit Publ., 2006, vol. 1, pp. 217–224 (in Russ.).
8. R Development Core Team. 2006. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. Available at: <http://www.R-project.org> (accessed June 12, 2017).
9. Suleymanov D.Sh., Gatiatullin A.R., Almenova A.B., Bashirov A.M. Multifunctional model of the Turkic morpheme. *Filologiya i kultura* [Philology and Culture]. 2016, no. 2, pp. 143–151 (in Russ.).
10. Suleymanov D.Sh., Gatiatullin A.R. *Strukturno-funktsionalnaya kompyuternaya model tatarskikh morfem* [Structurally Functional Computer Model of Tatar Morphemes]. Kazan, Fen Publ., 2003, 345 p.