Структура слога в татарском языке: от данных к модели

А.М. Галиева

Аннотация—Квантитативный анализ слоговых структур — первый и необходимый шаг к построению модели слога и дальнейших теоретических обобщений. В статье исследуется слоговая структура татарских слов на текстовых данных. Исследование проводилось с выделением шумных и сонорных согласных. Анализ показал, что в татарском языке преобладают слоги простой структуры: открытые (типы cv и sv) и закрытые (типы cvs, svs, cvc, svc). Сложные интервокальные кластеры для татарского языка не характерны. Во многих случаях присоединение словоизменительных аффиксов приводит к тому, что группы согласных перераспределяются в слоги более простой структуры. Распределение типов слогов в односложных словах в тексте в значительной степени определяется частотностью повторяющихся служебных слов и местоимений.

Основное внимание в работе уделено строению начальных и финальных слогов, составу их инициалей и код. Мы установили, что начальные и финальные слоги в татарском языке имеют разную структуру, что можно объяснить неодинаковой фонологической организацией основ и аффиксальных цепочек. Мы определили, что распределение шумных и сонорных согласных в инициалях и кодах неодинаково и также зависит от позиции слога в слове.

***Ключевые слова*—квантитативные исследования, слог, структура слога, татарский язык.**

I. Введение

Принципы сегментации речевого потока определяют слогоделение в языке. Слоги составляют важнейшую нишу между незначимыми (фонемы) и значимыми (морфемы и слова) единицами языка. Большинство языков мира имеют фиксированную слоговую структуру, когда возможные сочетания фонем в слоге жестко детерминированы.

Несмотря на интуитивную ясность концепта «слог», место и роль слога в лингвистике продолжают оставаться дискуссионными, что приводит к отсутствию ясных и общепринятых определений слога. Ситуацию очень точно описал K. Kohler: “Слог очень часто рассматривается как сущностная универсалия в фонологии; но также можно доказать, что слог не является *необходимым* концептом, потому что деление речевой цепочки на такие единицы можно объяснить иными причинами; *невозможным* концептом, так как такое деление будет произвольным; *вредным* концептом, поскольку это приводит к столкновению с грамматическими формативами. Если слог имеет какой-либо реальный статус в фонологии, его границы должны определены (must be discernible)” [1] (курсив наш – А.Г.).

Концепт слога остается на периферии татарского языкознания, хотя современные компьютерные технологии позволяют достаточно легко разработать системы для автоматического выделения слоговых структур и их количественного исследования. В татарских грамматиках проблеме слогоделения уделяется крайне мало внимания [2, 3]. Квантитативный анализ слоговых структур — первый и необходимый шаг к построению модели слога на данных татарского языка. Данное исследование носит предварительный характер и нацелен на разработку методологии исследования структуры слога в татарском языке для создания в дальнейшем комплексной модели слога. Результаты исследования имеют также прикладное значение и могут быть использованы в речевых технологиях и в текстовых редакторах, например, для уточнения правила автопереноса слов.

В качестве материала для данной статьи выбран законченный в смысловом отношении фрагмент первой части романа классика татарской литературы А. Еники «Әйтелмәгән васыять» («Невысказанное завещание») длиной в 2169 слов.

II. Краткий обзор литературы по теме

В последние десятилетия появилось большое количество специальных исследований, посвященных проблеме слога [4-16].

В работе С.В. Кодзасова и И.А. Муравьевой предлагается выделить два разных способа организации звукового потока в языках, которые исследователи называют «квантовым» и «волновым» [4]. Квантовые языки имеют слогоделение со строго заданной структурой и четко заданными границами слогов. В волновых языках структуры консонантных сочетаний являются неопределенными, а границы слога нечеткими, в связи с чем носители языка могут по-разному маркировать границы слогов (например, русс. *рес-публика и ре-спублика*).

Статья получена 15 октября 2019.
Галиева А.М., к.филос.н., доцент, ведущий научный сотрудник Института прикладной семиотики Академии наук Республики Татарстан, Россия (e-mail: amgalieva@gmail.com).

В статье [5] представлен обзор основных подходов к изучению строения слога.

В структуре слога в современных исследованиях выделяют такие составляющие:

- инициаль (onset) — согласный или группа согласных, обязательные в одних и факультативные в других языках;
- ядро (nucleus) — гласный или слогообразующий согласный, обязательный элемент слога;
- кода (coda) — согласный или группа согласных, факультативные в одних языках и недопустимые в других [5].

Часто ядро и коду объединяют в рифму (rime) — ветвление вправо, противостоящее инициали [5]. Данный термин имеет такое название, так как в стихах, как правило, рифмуются именно слогообразующие гласные и коды.

Многие исследователи обращаются к проблеме ранжированию ограничений, определяющих структуру слога в отдельном языке, создавая модели слогоделения в рамках так называемой теории оптимальности (Optimality Theory) [9, 10].

Разные исследователи применяют разные классификации фонем при анализе слоговых структур. Часть исследователей не выделяют подклассов согласных (например, [8, 13]), в отечественной лингвистике традиционно противопоставляют шумные и сонорные согласные [12, 14].

В последние годы появились статьи, в которых производится квантитативный анализ слоговых структур на материале отдельных языков.

В статье С. Андреева в структуре слога на данных русского языка противопоставляются гласные и согласные; количество одинаковых слогов позволило автору получить ранжированные частоты, которые затем сравнивались с теоретическими частотами, полученными с использованием специальных функций [13].

В в работе Г.А. Мороза предложена общая схема структуры адыгейского слога с выделением гласных и шумных и сонорных согласных. В связи с предположением о волновом характере адыгейского языка (когда носители затрудняются однозначно выделить границы слогов), в статье проанализированы следующие типы данных:

1) инициали слогов, с которых начинаются слова;
2) финали слогов, на которые заканчиваются слова;
3) консонантные кластеры в середине слов [14].

Имеются также исследования по теоретическому моделированию числа типов канонических слогов на материале различных языков [15] и создаются базы данных слогов [16].

III. Подготовка данных

Подготовительная работа по анализу структуры слога в татарском языке включала несколько основных этапов.

1. Отбор языкового материала.

Мы не используем материалы словарей, так как отбор и представление форм слова в словарях, например, фиксация глаголов в форме инфинитива или имени действия, большое количество заимствованных слов с нетипичной для татарского языка фонетической организацией и т.п., среди которых может оказаться очень много слов, редко используемых в реальных текстах, существенно искажает данные. Кроме того, в словарях лексемы даются в основной форме, поэтому словарные данные совершенно не отражают аффиксальных цепочек, которые связывают слова в предложениях, что для татарского языка с ее агглютинативной морфологией совершенно критично. В данном случае нас интересует распределение слогов в реальном употреблении, поэтому текстовый материал мы считаем оптимальным.

2. Приведение кириллического текста к форме, близкой к фонетической: 1 буква — 1 звук.

В целом татарское письмо основано на фонетическом принципе, согласно которому слова пишутся так, как произносятся. Тем не менее имеется достаточно много разных случаев нарушения принципа 1 буква — 1 звук. Поэтому нужно было найти в текстах единицы, в которых графическая форма не соответствует произношению, и преобразовать их.

Здесь нужно отметить 3 основных случая, существенных с точки зрения слогоделения и структуры слога:

- непроизносимые символы (1 буква — 0 звуков), влияющие на произношение соседних букв: *ь* и *ъ;*
- буквы *е, ю, я, ё* в определённых позициях (1 буква — 2 звука);
- буквы *у* и *ү* после некоторых гласных (буква обозначает гласный звук, но произносится как сонорный согласный (неслоговой гласный).

Кроме того, буква *в* может читаться как шумный согласный *в* (в словах, заимствованных из русского или европейских языков) или сонант *w* (в словах тюркского происхождения или в восточных заимствованиях) в зависимости от происхождения слов.

3. Подготовка правил, определяющих слогоделение татарских слов, и их реализация в программном коде. Деление исконно тюркских слов на слоги не вызвало затруднений. Для относительно небольшого числа заимствованных слов в выбранных для анализа текстах, в принципе допускающих альтернативное слогоделение (например, *ре-спу-бли-ка* и *рес-пуб-ли-ка*) вопрос решался следующим образом: эти слова произносились не изолированно (когда возможно двоякое выделение слогов), а в потоке татарской речи, и выбирался вариант, который получался при такой сегментации (в нашем случае это *рес-пуб-ли-ка*).

4. Преобразование словоформ из текста в структуры, состоящие из слогов соответствующего типа. После данного этапа производилась ручная проверка полученных данных и, при необходимости, их корректировка.

Таблица. 1. Примеры преобразования татарских словоформ в слоговые структуры

Словоформа в стандартной графической форме	**Словоформа после приведения к фонетической форме**	**Слоговая структура словоформы**
урман 'лес' *картлардан* *'у стариков'* *ямьле* " аулау 'охотиться' кияү"	урман картлардан йәмле awлaw кийәw	VS-SVS CVSC-SVS-CVS SVS-SV VS-SVS CV-SVS

5. Статистическая обработка данных и визуализация результатов.

Все этапы работы были реализованы на языке R [17], кроме базового R, были использованы пакеты tidyverse и stringr [18].

IV. Слоговые структуры в татарском языке

Современный татарский язык имеет богатую систему фонем: вокалическая система включает 9 исконных и 3 дополнительных гласных, используемых в заимствованных словах; консонантная система включает 25 исконных и 5 дополнительных согласных, которые встречаются в заимствованной лексике [2, 3]. Отличительная черта фонетики татарского языка — сингармонизм по ряду.

В аспекте нашего исследования важно разделение согласных на шумные и сонорные (сонанты) по соотношению основного тона и шума. К сонорным относятся плавные *р* и *л*, носовые *м, н, ң*, а также полугласные *й* и *w* [2, 3].

Вначале посмотрим, как распределяются слоги разной структуры в татарском тексте в целом. В таблице 2 представлено количественное распределение слоговых структур на материале фрагмента романа А.Еники. Материал таблицы показывает, что около половины (49%) всех слогов образуют слоги простой структуры, состоящие из начального согласного (сонорного или шумного) и слогообразующего гласного (ядра). Слоги, образованные инициалью, состоящей из одного согласного, ядра и коды, также состоящей из одного согласного, суммарно составляют 38%.

Таблица 2. Распределение слогов разной структуры в тексте А.Еники.

Ранг	**Слоги**	**Количество**	**Доля**
1	cv	1536	0.31
2	sv	881	0.18
3	cvs	787	0.16
4	svs	461	0.09
5	cvc	421	0.08
6	v	359	0.07
7	svc	240	0.05
8	vs	130	0.04
9	vc	105	0.02
10	cvsc	24	0.005
11	csvc	3	0.0006
	...		
	Всего	4963	1

Обратим особое внимание на самый частотный слог с достаточно тяжелой кодой, состоящей из сонорного и шумного согласного - тип cvsc, который в нашей выборке встречается 24 раза. Можно предположить, что такая структура может быть в составе заимствований, например, ориентализмов. Тем не менее, мы выяснили, что это не так: структура cvsc в анализируемом фрагменте романа А.Еники встречается в составе исконно тюркских основ (например, *карт* 'старый, старик', *дүрт* 'четыре', *кайт* 'вернись' и др.) и на стыке двух аффиксов понудительного залога (например, слог - *герт* в слове *йө-герт* 'заставь бежать'). В нашей выборке слог типа cvsc обнаружен только в одном случае в заимствовании — в составе слова *зәгыйфь* 'слабый', пришедшем из арабского языка.

Следует отметить, что сам по себе интервокальный кластер sc в татарском языке достаточно частотен, но его элементы при словоизменении часто распределяется по разным слогам, ср.:

кайтты (cvsc-cv) 'вернулся' — *кайта* (cvs-cv) 'возвращается';

карт (cvsc) 'старик' — *карты* (cvs-vc) 'его/ее старик'.

Особенности распределения единиц с разным рангом и частотностью удобно представить в виде графика (рис. 1). Для сравнения на график наложены также теоретические значения, вычисленные на основе закона Ципфа по формуле: $P_n = P_1 / n^{\alpha}$, где P_n— количество слов со слоговой структурой *n*-го ранга; P_1 — количество слов со слоговой структурой 1-го ранга. Значение α подбиралось эмпирически, в нашем случае наиболее подходящее значение было определено как α = 0,99.

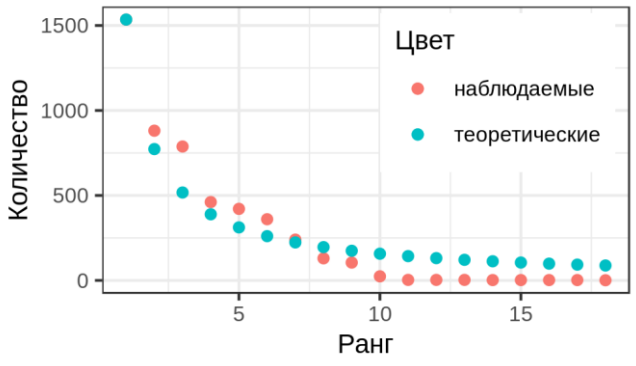

Рис. 1. Распределение слогов с разной структурой в тексте А. Еники

В анализируемом тексте выявлено 18 слоговых структур, 7 из которых достаточно частотны и

составляют не менее 5% каждый. Очевидно, что новые текстовые данные, особенно тексты с большим количеством заимствованной лексики, дадут новые типы слоговых структур с более сложными инициалями и кодами.

V. Структура односложных слов

Проведем анализ структуры слогов односложных слов. Односложные слова интересны тем, что они представляют собой естественным образом вычлененные слоговые структуры.

Таблица 3. Слоговая структура односложных слов

Ранг	**Слоги**	**Количество**	**Доля**
1	cv	106	0.279
2	cvs	94	0.247
3	cvc	60	0.158
4	vs	32	0.084
5	vc	23	0.060
6	svc	18	0.047
7	v	16	0.042
8	cvsc	11	0.029
9	sv	10	0.026
10	svs	6	0.016
11	csvc	2	0.005
12	cvss	1	0.003
13	svsc	1	0.003
	Всего	380	1

В односложных словах выявлено 13 слоговых структур (см. таб. 3). Основную массу односложных слов в анализируемом тексте составили служебные слова (частицы, союзы, послелоги) и местоимения (включая слова слова, функционирующие как местоимения). В частности, в нашей выборке односложных слов 89 случаев употребления (23% от общего количества единиц) составляют фонетические варианты частицы *да* 'тоже, также' (вариант *да* встречается 47 раз, *дә* — 33 раза, *тә* — 7 раз, *та* — 2 раза).

Если сравнить частотность слоговых структур в односложных словах с данными по слоговым структурам текста в целом, то можно заключить, что тип cv в обоих случаях является самым частотным. Второй по рангу в тексте в целом тип sv на данных односложных слов имеет лишь ранг 9. Тип cvs достаточно частотен в обоих случаях. Мы можем сделать вывод о том, что распределение слоговых структур односложных слов в тексте определяется количеством повторяющихся служебных и местоименных слов.

VI. Структура начальных и финальных слогов в слове

Далее обратимся к исследованию структуры начальных и финальных слогов, эти данные интересны тем, что позволяют определить, насколько отличаются фонетические структуры основ (начальные слоги представляют собой основы или их части) и аффиксальных цепочек (в многосложных словах последний слог — как правило, аффикс или фрагмент аффиксальной цепочки).

Выборка в данном случае включала данные по словам, состоящим из двух и более слогов, всего в анализируемом тексте оказалось 1785 таких слов. Соответственно, выборки начальных и финальных слогов содержали по 1785 единиц, средние слоги многосложных слов не рассматривались.

Из 18 типов слогов, обнаруженных в тексте А. Еники (с учетом односложных слов и слогов, стоящих в середине слова), в начальных и финальных слогах реализовано 14 типов.

Мы исходили из гипотезы о том, что в распределении слоговых структур, стоящих в начале и конце словоформ, должны быть статистически важные отличия; это должно быть связано с разным фонетическим устройством основ и цепочек аффиксов, которое было отмечено выше.

На рисунке 2 представлено распределение слогов разных типов в начале и конце словоформ. Данные свидетельствуют о том, что финальные слоги тяготеют к меньшему количеству типов: представлены преимущественно типы cv (474 раза), sv (391 раз), svs (296 раз), cvs (313 раз), svc (148 раз), cvc (148 раз). Кроме того, финальные слоги крайне редко начинаются на гласный (в выборке вообще не представлен тип v, всего 3 раза встречаются тип vc и 6 раз тип vs).

Начальные слоги характеризуются большим разнообразием, при этом тип cv доминирует, встречаясь 683 раза (26% всей выборки).

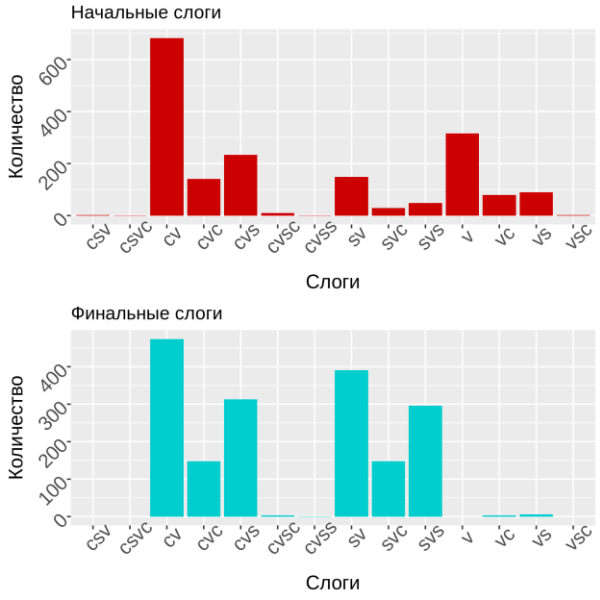

Рисунок 2. Начальные и финальные слоги.

Анализ состава инициалей и код позволяет получить более полные данные об устройстве начальных и финальных слогов татарских слов. На рисунке 3 представлено распределение слогов, у которых слогообразующий гласный прикрыт инициалью.

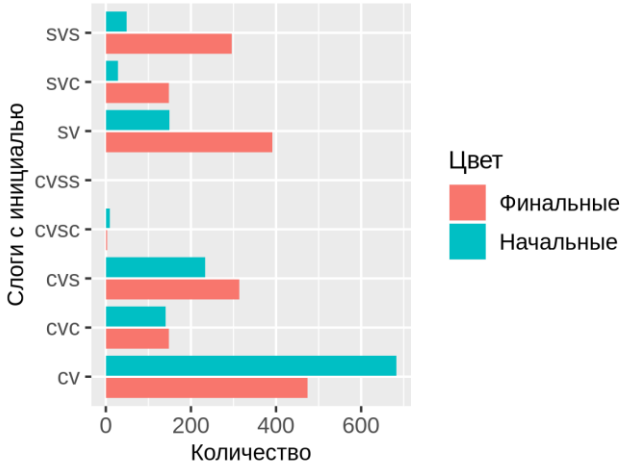

Рисунок 3. Прикрытые слоги.

Рисунок 4 представляет распределение слоговых структур, обладающих кодой.

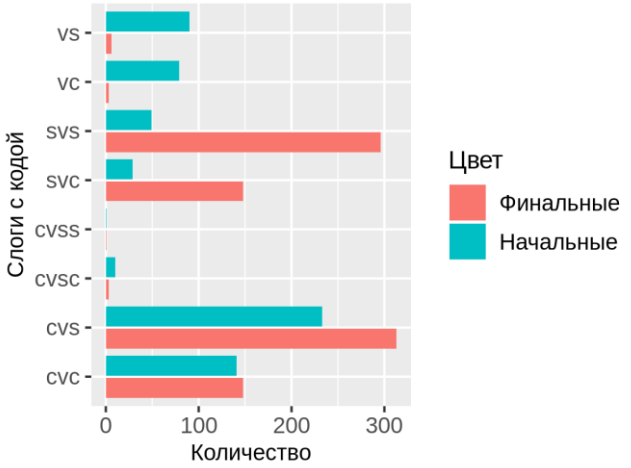

Рисунок 4. Закрытые слоги.

Таким образом, мы получили данные о составе инициалей и код в начальных и финальных слогах татарских слов. Теперь зададимся вопросом, случайны ли распределения шумных и сонорных согласных в начальных и финальных слогах? Ответить на этот вопрос поможет тест хи-квадрат, который позволяет оценить статистическую значимость различий двух или нескольких относительных показателей в таблице сопряженности.

Критерий χ^2 Пирсона– это непараметрический метод, позволяющий оценить значимость различий между наблюдаемым количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы. Нулевая гипотеза теста хи-квадрат заключается в том, что нет зависимости между столбцами и строками в таблице сопряженности: событие «наблюдение в строке i» не зависит от события, когда то же самое наблюдение оказывается в столбце j [19].

Тест хи-квадрат проведем дважды - отдельно для инициалей и код.

В таблице 4 приводятся наблюдаемые значения для слогов, прикрытых инициалью, их ожидаемые значения и результаты теста хи-квадрат.

Таблица 4. Результаты теста хи-квадрат для слогов с инициалью (с поправкой Ейтса)

	Начальные слоги		**Финальные слоги**	
	Наблюдаемые значения	**Ожидаемые значения**	**Наблюдаемые значения**	**Ожидаемые значения**
Шумный	1068	846.88	939	1160.12
Сонорный	227	448.12	835	613.88
Результаты теста χ^2	$\chi^2 = 287.34$, df = 1, p-value < 2.2e-16			

Основываясь на тесте хи-квадрат, мы можем делать вывод о том, что зависимость употребления шумных или сонорных согласных в инициали от позиции слога в слове является статистически значимой (мы получили очень маленькое значение p-value = 2.2e-16).

Далее проведем аналогичный анализ для код. Поскольку слоги со сложными кодами единичны (например, тип cvsc встречается 10 раз в начальных и 3 раза, cvss 1 раз в начальных и 1 раз в финальных слогах, мы будем учитывать только более многочисленные слоги — с простой кодой, состоящей из одного согласного — шумного или сонорного).

В таблице 5 приводятся наблюдаемые значения для слогов с кодой, их ожидаемые значения и результаты теста хи-квадрат.

Таблица 5. Результаты теста хи-квадрат для слогов с кодой (с поправкой Ейтса)

	Начальные слоги		**Финальные слоги**	
	Наблюдаемые значения	**Ожидаемые значения**	**Наблюдаемые значения**	**Ожидаемые значения**
Шумный	249	221.70	299	326.30
Сонорный	372	399.30	615	587.70
Результаты теста	$\chi^2 = 8.4623$, df = 1, p-value = 0.0036			

χ^2	

Тест хи-квадрат показал, что наблюдаемые значения статистически значимо отличаются от ожидаемых значений. Соответственно, мы можем делать вывод о том, что зависимость употребления шумных или сонорных согласных в коде от позиции слога в слове (в начальном или финальном слоге) статистически значимо (p-value = 0.0036).

Имеющиеся данные нам позволяют сделать более детальный анализ слоговых структур, например, используя различные классификации согласных. Так, таблица 6 представляет распределение сонантов разных классов в инициалях начальных слогов многосложных слов.

Таблица 6. Распределение классов сонантов в инициалях первых слогов.

Классы сонантов	**Количество**	**Доля**
Полугласный й	130	0.573
Носовые	71	0.313
Плавные	15	0.066
Полугласный w	11	0.048
Всего	227	1

Мы видим, что сонанты, относящиеся к разным классам, распределены очень неравномерно. Первые слоги в слове, а соответственно, основы, имеют тенденцию начинаться на *й* и относительно нечасто начинаются плавными *р* и *л* и полугласным *w.* Можно сделать вывод о том, что квантитативный анализ слоговых структур татарского языка, основанный на выделении согласных разных типов, имеет перспективы.

VII. Заключение

В качестве эмпирической базы для исследования был выбран фрагмент романа классика татарской литературы А.Еники, состоящий из 2169 слов (4963 слогов). Отказ от данных словарей в качестве материала исследования был обусловлен тем, что словари содержат большой объем заимствований с нетипичными для татарского языка слоговыми структурами, фиксируют слова в начальной форме и не отражают аффиксальных цепочек словоформ, все это может очень сильно исказить результаты исследования.

В проанализированном фрагменте романа А.Еники выявлено 18 разных слоговых структур, характеризующихся разной частотностью. Мы выяснили, что в татарском языке преобладают слоги простой структуры. 49% всех слогов образуют структуры, состоящие из начального согласного и слогообразующего гласного (типы cv и sv). Слоги, образованные инициалью, состоящей из одного согласного, ядра и коды, также состоящей из одного согласного, суммарно составляют 38% (типы cvs, svs, cvc, svc). Сложные интервокальные кластеры для татарского языка не характерны.

Был проведен анализ встречаемости шумных и сонорных согласных в составе инициалей и код в зависимости от положения слога в слове, тест проводился отдельно для инициалей и код. Тест хи-квадрат показал, что наблюдаемые значения в обоих случаях статистически значимо отличаются от ожидаемых значений. Полученные результаты говорят о том, что нулевую гипотезу о независимости появления каждого типа согласных в начальных и финальных слогах можно отвергнуть.

Предполагается, что дальнейшие исследования будут выполняться с учетом шкалы сонорности: с выделением подклассов сонантов (полугласные j и w, плавные, носовые) и шумных согласных (фрикативные, смычные), что должно дать более детальные сведения об устройстве интервокальных кластеров в татарском языке. Кроме того, планируются специальные исследования, нацеленные на анализ сложных взаимосвязей между структурой слога и морфемной структурой слова и словоизменением.

Библиография

[1] Kohler K. J. “Is the syllable a phonological universal?”, in *Journal of Linguistics, 1966,* 2, pp. 207–208.

[2] Закиев М.З. (ред.). *Татарская грамматика*. Казань: Татар. кн. Изд-во, 1993. - Т.1. - 584 с.

[3] Хисамова Ф.М. (ред.). *Татар грамматикасы.* Казан: ТӘhСИ, 2015. - Т. I. - 512 б.

[4] Кодзасов С. В., Муравьева И. А. “*Слог и ритмика слова в алюторском языке”* // Публикации отделения структурной и прикладной лингвистики МГУ. Филологический факультет. Вып. 9. М.: МГУ, 1980, с. 103–127.

[5] Hulst van der H., Ritter N. A. “Theories of the syllable”, in *The syllable: Views and facts.* Berlin: Mouton de Gruyter, 1999, pp. 13 — 52.

[6] Hulst van der H., Ritter N. A. *The syllable: Views and facts.* Berlin: Mouton de Gruyter, 1999. 777 p.

[7] Russo D. (ed.) *The Notion of Syllable across History, Theories and Analysis*. Cambridge: Cambridge Scholars Publishing, 2015. 624 p.

[8] Zörnig P. et al. *Quantitative Insights into Syllabic Structures* / P. Zörnig, K. Stachowski, A. Rácová, Y, Qu, l Místecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann. — Lüdenscheid: RAM-Verlag, 2019. 134 p.

[9] Prince A., Smolensky P. *Optimality Theory: Constraint interaction in generative grammar.* Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, 1993. Available: URL: http://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF

[10] Féry C., Vijver van de R. (eds.) *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press, 2003. 415 p.

[11] Aşliyan R., Günel K. “Design and Implementation for Extracting Turkish Syllables and Analyzing Turkish Syllables”, in *INISTA 2005, International Symposium on INnovations in Intelligent SysTems and Applications, 2005,* pp. 170-173.

[12] Князев С. В. *Структура фонетического слова в русском языке: синхрония и диахрония.* М.: Макс-Пресс, 2006. - 225 с.

[13] Andreev S. “Distribution of Syllables in Russian Sonnets”, in *Glottometrics*. 2018, vol. 41, pp.13-23.

[14] Мороз Г. А. “Слоговая структура адыгейского языка: от данных к обобщениям” // Вопросы языкознания, 2019, № 2, с. 82-95.

[15] Kelih E., Mačutek J. “Number of canonical syllable types: A continuous bivariate model”, in *Journal of Quantitative Linguistics,* 2013, 20, pp. 241-251.

[16] Dinu, L., Dinu A. "On the data base of Romanian syllables and some of its quantitative and cryptographic aspects.", in *LREC,* 2006, pp. 1795-1798.
[17] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2018. Available: URL: https://www.R-project.org/.
[18] Grolemund G., Wickham H. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O'Reilly, 2017. 494 p.
[19] Conover W.J. (1999). *Practical Nonparametric Statistics.* New York: Wiley, 1999. 584 p.

Syllable structure in Tatar: from data to modeling

Alfiya Galieva

Abstract—A quantitative analysis of syllable structures is the first and necessary stage to build the syllable model of a language and further theoretical generalizations. The paper studies the syllable structure of Tatar words on text data. The author abandoned the idea of using dictionaries as a data source, because dictionaries contain a large amount of loanwords with uncommon syllabic structures; besides dictionaries fix words in their initial form and do not reflect affixal chains of words.

The work is based on distinguishing sonorant and obstruent consonants. The analysis disclosed that in Tatar the syllables of simple structure prevail, the open (of cv and sv types) and closed ones (of cvs, svs, cvc, svc types). Complex intervocal clusters are not frequent; joining inflection affixes in many cases leads to redistributing consonant groups into syllables of a simpler structure. The distribution of syllables types in monosyllables in the text is largely determined by the number of frequently used functional words and pronouns.

The main focus of the work is on the patterns of the initial and final syllables of words. The author ascertained the dissimilar structure of the onsets and codas in the initial and final syllables, caused by different phonological organization of stems and affixal chains.

Key words **- syllable, syllable structure, the Tatar language, quantitative analysis.**

REFERENCES

[1] Kohler K. J. “Is the syllable a phonological universal?”, in *Journal of Linguistics, 1966,* 2, pp. 207–208.

[2] Zakiev M.Z. (ed.) The Tatar grammar [Tatarskaya grammatika]. Kazan: Tatar publishing House, 1993, vol. 1, 584 p.

[3] Khisamova F.M. Tatar grammar [Tatar grammatikası]. Kazan: Institute of Language, Literature and Art, 2015. Vol. 1. 512 p.

[4] Kodzasov S. V., Murav’eva I. A. “Syllable and word rhythmics in Alutor” [Slog i ritmika slova v alyutorskom yazyke] in *Publikatsii otdeleniya strukturnoi i prikladnoi lingvistiki MGU*. Filologicheskii fakul’tet. No. 9. Moscow: Lomonosov Moscow State Univ., 1980, pp.103–127.

[5] Hulst van der H., Ritter N. A. “Theories of the syllable”, in *The syllable: Views and facts.* Berlin: Mouton de Gruyter, 1999, pp. 13 — 52.

[6] Hulst van der H., Ritter N. A. *The syllable: Views and facts.* Berlin: Mouton de Gruyter, 1999. 777 p.

[7] Russo D. (ed.) *The Notion of Syllable across History, Theories and Analysis*. Cambridge: Cambridge Scholars Publishing, 2015. 624 p.

[8] Zörnig P. et al. *Quantitative Insights into Syllabic Structures* / P. Zörnig, K. Stachowski, A. Rácová, Y, Qu, l Místecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann. — Lüdenscheid: RAM-Verlag, 2019. 134 p.

[9] Prince A., Smolensky P. *Optimality Theory: Constraint interaction in generative grammar.* Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, 1993. Available: URL: http://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF

[10] Féry C., Vijver van de R. (eds.) *The Syllable in Optimality Theory*. Cambridge: Cambridge University Press, 2003. 415 p.

[11] Aşliyan R., Günel K. “Design and Implementation for Extracting Turkish Syllables and Analyzing Turkish Syllables”, in *INISTA 2005, International Symposium on INnovations in Intelligent SysTems and Applications, 2005,* pp. 170-173.

[12] Knyazev S.V. The structure of the phonetic word in Russian: synchrony and diachrony [Struktura foneticheskogo slova v russkom jazyke: sinhronija i diahronija]. Moscow: Max-Press, 2006/ - 224 p.

[13] Andreev S. “Distribution of Syllables in Russian Sonnets”, in *Glottometrics*. 2018, vol. 41, pp.13-23.

[14] Moroz G. A. “Adyghe syllable structure: From empirical data to generalizations” [Slogovaja struktura adygejskogo jazyka: ot dannyh k obobshhenijam]. Voprosy Jazykoznanija, 2019, 2, pp. 82–95.

[15] Kelih E., Mačutek J. “Number of canonical syllable types: A continuous bivariate model”, in *Journal of Quantitative Linguistics,* 2013, 20, pp. 241-251.

[16] Dinu, L., Dinu A. "On the data base of Romanian syllables and some of its quantitative and cryptographic aspects.", in *LREC,* 2006, pp. 1795-1798.

[17] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2018. Available: URL: https://www.R-project.org/.

[18] Grolemund G., Wickham H. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O’Reilly, 2017. 494 p.

[19] Conover W.J. (1999). *Practical Nonparametric Statistics.* New York: Wiley, 1999. 584 p.

Статья получена 15 октября 2019.

Alfiya Galieva, PhD in Philosophy, senior researcher of the Institute of Applied Semiotics, Tatarstan Academy of Sciences, Kazan, Russia (e-mail: amgalieva@gmail.com).