Б. В. Орехов

БАШКИРСКИЙ СТИХ XX ВЕКА

Корпусное исследование

Санкт-Петербург
АЛЕТЕЙЯ
2019

УДК 821.512.141.08
ББК 83.3(2Рос=Баш)
О 654

Утверждено к печати
Федеральным государственным бюджетным учреждением науки
Институтом языкознания Российской академии наук

Рецензенты:
член-корреспондент РАН *А. В. Дыбо*
кандидат филологических наук *К. М. Корчагин*

Орехов Б. В.
О654 Башкирский стих XX века. Корпусное исследование / Б. В. Орехов. – СПб.: Алетейя, 2019. – 344 с.: ил.

ISBN 978-5-907189-29-4

Представлены результаты всестороннего количественного исследования башкирской системы версификации в XX веке. С использованием современных статистических инструментов подвергаются анализу все уровни организации стихотворного текста от фоники до лексики и грамматики, с особенным вниманием к метру и ритму. Количественные данные получены на корпусе текстов 103 башкирских поэтов общим объемом в 1.77 млн словоупотреблений. Анализ предварен подробным обзором науки о тюркском стихе, начиная с 1950-х годов. Утверждается, что основную роль в башкирском стихосложении XX века играют силлабические формы фольклорного происхождения узун-кюй и кыска-кюй, первая из которых специфична для поволжско-кыпчакского слогосчитающего стиха. Приводится подробное сопоставление башкирского стиха с киргизским.

Книга завершается примерами поэтических текстов на башкирском языке, сгенерированных с использованием искусственных нейронных сетей.

УДК 821.512.141.08
ББК 83.3(2Рос=Баш)

ISBN 978-5-907189-29-4
9 785907 189294

16. Лексика стиха и семантический ореол метра

16.1. Дистрибутивно-семантическая модель башкирской поэзии

Лексика с точки зрения стиховедения играет в тексте подчиненную роль. Семантика слов ближе всего стоит к смыслам более высокого уровня, которые обычно ищет читатель в поэтическом произведении, но при этом меньше других единиц работает на формальную организацию текста, которая является объектом науки о стихе. Но поскольку художественное целое произведения пронизывает множество связей, объединяющих различные уровни в систему, взаимодействие между стихом и значением всё же существует.

Терминологическому аппарату описания таких взаимодействий и конкретным наблюдениям в этой области на материале русского стиха посвящена книга [Гаспаров 2012]. Кажется, что именно М. Л. Гаспарову удалось дальше всего продвинуться в исследовании этого феномена, хотя уже довольно давно количественные исследования поэтических текстов предпринимались при помощи частотных словарей (см. [Шестакова 2011]).

Материалы к частотному словарю башкирской поэзии можно найти в Приложении 3. Список первых 300 наиболее употребимых слов составлен на основе результатов автоматической лемматизации (см. §7.2), снабжён указанием на часть речи, число вхождений на миллион словоупотреблений (IPM) и количество стихотворений, в которых эта лемма встретилась.

Но кажется, что традиционные подходы к изучению семантики в стихе можно расширить с использованием современных компьютерных инструментов обработки естественного языка.

Одна из ключевых на сегодняшний день компьютерных технологий, применяющихся для работы с лингвистическим значением, это векторное представление слов (word embeddings). Его теоретическая база — это дистрибутивная гипотеза, которая формулируется следующим образом: степень семантического сходства двух слов является функцией от схожести контекста, в котором эти слова могут появиться. Ещё в середине прошлого

века эта идея прозвучала в афористичной формулировке Дж. Р. Фёрса: «Вы узнаете слово по той компании, в которой оно ходит» [Firth 1957: 11]. Однако только в последние годы были созданы эффективные алгоритмы для машинного представления такого контекста. Разработчики сочли удобным хранить контекст слова в виде вектора. Координатами вектора слова является определенным образом нормализованная встречаемость слов, составляющих контекст данного. Существует несколько алгоритмов векторизации контекста, самым популярным из которых является word2vec.

Представив контекст слова в векторном пространстве, мы можем считать, что каждый вектор — это семантика слова, с которой можно совершать те же математические операции, какие в принципе допускают над собой вектора: складывание, вычитание, нахождение ближайшего вектора путем вычисления косинусного расстояния.

Векторное представление слов уже применялось для анализа поэтического текста и в исследованиях русской [Орехов 2016], и в исследованиях башкирской поэзии [Гречачин 2018]. В них использовалось ручное качественное сравнение квази-синонимов некоторых слов. Далее мы предпримем количественное исследование векторной семантики с помощью анализа списков квази-синонимов. Квази-синонимы в векторном представлении — это ближайшие вектора, то есть, в соответствии с дистрибутивной гипотезой, слова с наиболее похожим значением. Префиксоид «квази» появляется здесь потому, что похожее значение не обязательно должно быть у синонима. По своим контекстам очень близки и антонимы (близкими соседями являются «холодный» и «горячий», «черный» и «белый»), и слова, связанные гипо-гиперонимичными отношениями («студент» и «учащийся»). Обычно мы не можем разграничить эти отношения в модели, поэтому все соседи называются «квази-синонимами».

Качество векторной модели заметно зависит от размера корпуса [Altszyler et al. 2016]. Это значит, что реальные языковые закономерности будут отражаться в векторном пространстве только в случае, если корпус контекстов, на котором такая модель будет построена, окажется достаточно большим. Башкирский поэтический корпус имеет средний размер. Хотя семантические отношения на большем корпусе проявлялись бы более отчетливо, все же более полутора миллионов словоупотреблений — достаточный для релевантных выводов объем.

В качестве контрастивных данных нами взят лемматизированный корпус статей из газеты «Йәшлек». Обе коллекции текстов послужили исходными данными для тренировки моделей с помощью архитектуры Continuous Bag-of-Words (CBOW). Газетные тексты были разбиты на предложения, а поэтические — на отрезки по 4 строки.

В дальнейшем мы воспользовались методикой сравнения моделей, предложенной в [Kutuzov, Kuzmenko, 2015]. Из-за недостаточности объема поэтического корпуса мы были вынуждены ограничиться 4000 частотными именами вместо 10 000, которыми оперировали названные авторы.

Для каждого из имен были получены 10 ближайших соседей. Ниже приведен пример списков квази-синонимов для слова «йорт» ‘дом’:

Модель башкирского поэтического корпуса (после самой лексемы указано значение косинусной близости между *йорт* и этой лексемой):
нигеҙ ‘фунамент’ 0,762615
өй ‘дом’ 0,756706
ауыл ‘деревня’ 0,690155
тупһа ‘крыльцо’ 0,625492
бина ‘здание’ 0,616340
яг ‘сторона’ 0,601330
өфө ‘Уфа’ 0,596297
гостиница ‘гостиница’ 0,594483
өҫтәл ‘стол’ 0,594145
мөйөш ‘угол’ 0,586830

Модель корпуса газеты «Йәшлек»:
бина ‘здание’ 0,664548
һарай ‘дворец’ 0,579985
бүлмә ‘комната’ 0,543696
өй ‘дом’ 0,536694
мәсет ‘мечеть’ 0,515221
интернат ‘интернат’ 0,505365
бакса ‘огород’ 0,504052
дауахана ‘больница’ 0,500521
ятаҡ ‘ночлег’ 0,4967596
арай ‘пойма’ 0,4957030

Хорошо видно, что списки квази-синонимов отражают специфику и стилистику корпусов. С одной стороны, в публицистической газетной прозе обнаруживается сходство контекстов слова «дом» и предельно прозаических «интернат», «больница». С другой стороны, стихотворный корпус осмысляет «дом» как этическую категорию, и в нем мы наблюдаем сходство контекстов этого слова с другими словами, несущими дополнительные поэтические смыслы.

Используя эти списки, для каждого слова с помощью коэффициента Жаккара мы подсчитали меру сходства значений в обоих корпусах. Эта мера представляет собой отношение числа пересечений списков к длине объединенного списка, и принимает значение от 0 (нет пересечений) до 1 (все квази-синонимы в обоих списках совпали, таких случаев для наших моделей не наблюдается).

Более 82 % выбранных для анализа слов не имеют пересечений среди квази-синонимов в моделях из газетного и поэтического корпусов. Это показывает большую дистанцию в значении слов, в которых они употребляются в каждом из жанрово-стилистических типов речи.

Наивысшее значение коэффициента Жаккара: 0,33, его достигли слова *һыйыр* ‘корова’, *үр* ‘высота’, *йәмле* ‘красивый’, *кәзә* ‘коза’, *бейек* ‘высокий’. В эту группу вошли прецедентные имена сельского быта и имена, выражающие качественные признаки. Контексты для этих слов и в газетной, и в поэтической речи отчасти сходны. Это может объясняться устойчивостью речевых ситуаций, в которых упоминаются термины деревенского быта, важного как для региональной газеты, так и для поэтического творчества. Последнее не специальное свойство именно башкирской поэтической культуры. Ориентация на сельского читателя называется одной из стержневых тенденций карачаевской поэзии 1930-х годов [Хапаева 1999: 19].

Группу слов с коэффициентом Жаккара 0,25 образует ряд имен, которые можно отнести к разряду поэтизмов: *йыр* ‘песня’, *моң* ‘напев’, *йөрәк* ‘сердце’ *диңгеҙ* ‘море’, а также термины родства әсә ‘мать’, бала ‘ребенок’. Можно предположить, что их сходство вызвано переключением стилистического регистра в газетной статье, то есть попыткой автора создать некоторый поэтический эффект в публицистическом тексте. Иными словами, сходство в значениях возникает благодаря воздействию стилистики стихотворных произведений на прозаические, а не наоборот.

Количественный анализ представительного лексического материала с помощью дистрибутивных векторных моделей показывает, что семантика слов в башкирской поэзии и башкирской газетно-деловой прозе сильно отличается. Даже частотные существительные фигурируют в текстах в резко различных контекстах, что позволяет говорить о стихотворной речи как о сформировавшейся за прошедшие десятилетия XX века устойчивой языковой подсистеме. Несмотря на отмечавшиеся в истории башкирской литературы периоды, когда художественная и публицистическая речь имели не вполне определенную границу («публицистическая лирика» 1940-х гг., см. [История 2014: 314]), в целом

башкирская поэзия выработала для себя четко обозначенные формы, что говорит о ее стилистической зрелости, а также о стилистической зрелости всего литературного языка.

16.2. Тематическое моделирование и семантический ореол метра

Одним из замечательных достижений стиховедения XX века стала формулировка гипотезы о связи метрической формы и плана содержания поэтического произведения. В упрощенном виде ее можно пересказать так: выбирая размер для своего текста, поэт редко руководствуется отрефлектированными рациональными соображениями. Однако неподконтрольно самому автору на него воздействует культурная память — припоминание о том, какие важные для традиции тексты уже были написаны этим размером ранее. Создавая произведение, поэт вступает в сложные отношения наследования-противоречия с предшественниками, что отражается в том числе на тематической составляющей его стихотворения. В результате мы видим ситуацию, в которой некоторые заметные стихотворения, написанные одним размером, тяготеют к воспроизводству похожего круга тем. Это вовсе не означает, что все стихотворения, написанные одним размером, должны с обязательностью вращаться около одного тематического набора. Кроме того, темы из этого набора свободно фигурируют в стихотворениях, написанных другими размерами. Но некоторая тенденция, связывающая темы и размеры, в культуре всё же присутствует. Эта тенденция, не создающая жесткой зависимости между темой и ее версификационным оформлением, была названа «семантический ореол метра». Слово «ореол» в этом сочетании как раз указывает на нестрогость тех закономерностей, которые этот термин описывает.

В науке уже были попытки указать на компьютерные разработки, способные автоматизировать выделение в стихах тем, сопоставимых с понятием семантического ореола. Так, А. Ч. Пиперски предлагает использовать для этого предложенный А. Килгарриффом метод извлечения ключевых слов [Piperski 2017]. Но, как показали наши опыты[71], для этих целей успешно может применяться тематическое моделирование.

Тематическое моделирование — это набор алгоритмов, которые, основываясь на данных о частотности и совместной встречаемости слов

[71] Сошлемся на совместный с Р. Г. Лейбовым устный доклад «Семантический ореол метра и тематическое моделирование» в Тартуском университете 28 февраля 2018 г.