УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 149, кн. 4 Гуманитарные науки *2007*

УДК 81:004.9

СТАТИСТИЧЕСКИЕ ПАРАМЕТРЫ ГРАММАТИК КЫПЧАКСКИХ И ОГУЗСКИХ ЯЗЫКОВ

В.Д. Соловьев, В.Р. Байрашева

Аннотация

В статье сопоставляются частоты встречаемости грамматических свойств в двух группах тюркских языков. Для анализа использована база данных «Языки мира». Показано, что для огузской группы языков обнаруживается значительно больше признаков, содержащихся ровно в одном языке данной группы. Результат основан на анализе большого числа данных и является статистически достоверным. Предлагается гипотеза, объясняющая данное явление.

Введение

Традиционно принято считать, что тюркская группа языков является весьма однородной по составу. В целом это действительно так, различия между тюркскими языкам невелики. С появлением базы данных «Языки мира» появились новые возможности для широкомасштабного сопоставления грамматических систем различных языковых групп. Одним из авторов данной статьи предложена оригинальная методика [1] качественной и количественной оценки структурных инноваций в языковой семье, включая заимствования. В данной статье эта методика применена к сопоставлению динамики развития грамматики кыпчакских и огузских языков.

База данных «Языки мира» (далее сокращенно – БД) создана в Институте языкознания РАН и является одной из двух крупнейших типологических баз данных в мире. Разработка концепции БД была начата в 1980-е годы практически одновременно с реализацией проекта по созданию энциклопедии «Языки мира», осуществляемого под руководством члена-корреспондента РАН В.Н. Ярцевой. В качестве основного средства формализации была разработана так называемая «модель реферата», представляющая собой универсальную схему описания языка, где языковые явления и категории представлены в жестко заданной форме. БД содержит описание следующих сфер языка: фонетика, морфология, синтаксис.

Основную работу по заполнению БД выполнила Е.И. Ярославцева. Были созданы программные оболочки сначала для DOS, а затем для Windows. В 2006 г. в рамках Российского научно-образовательного Центра по лингвистике имени И.А. Бодуэна де Куртенэ при Казанском государственном университете создан пилотный вариант Web-версии БД. Он размещен в интернете по адресу: www.dblang.ru и с 2006 г. доступен в тестовом режиме. По состоянию на конец 2006 г. БД характеризуется следующими основными параметрами: 315 языков

(в основном, Европы, Северной и Центральной Азии), 3821 признак, бинарный формат представления данных. Подробное описание БД может быть найдено в монографии [1].

1. Диаграммы распределения частотности признаков

Различные языковые свойства распределены по языкам мира крайне неравномерно. Некоторые, например свойство быть языком номинативного строя, высокочастотны, другие же – эргативность – встречаются редко. БД позволяет получить детальную статистику частотности всех свойств. Эти данные можно суммировать, подсчитав для каждого числа *n*, сколько признаков существует ровно в *n* языках. На диаграмме 1 в форме графика приведены соответствующие данные, рассчитанные для всей БД. Без обширной базы данных получить информацию такого рода не представляется возможным. В дальнейшем графики этого типа будем называть ЯП-диаграммами (‘язык-признак’-диаграмма). Признак, существующий ровно в *n* языках, будем называть *n*-языковым признаком.

Как оказалось, большинство признаков, представленных в БД, очень редки. Так, 1-признаков (т. е. присутствующих ровно в одном языке или уникалий) – 1488, т. е. более трети, 2-признаков – 466 и т. д. Начиная примерно с уровня $n = 25$ число *n*-признаков становится очень мало, меньше 10.

Внешний вид графика (близкого к гиперболе) на первый взгляд вызывает удивление. В [1] предложено объяснение формы графика, которая связана с особенностями эволюции языковой системы.

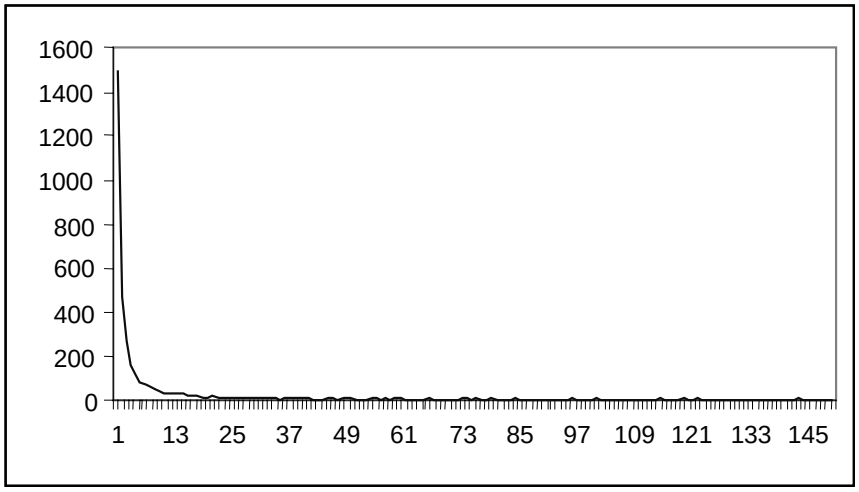

Диаграмма 1. ЯП-диаграмма всех языков базы данных (левая половина графика)

Свяжем частотность признаков с их местом в языковой эволюции. К настоящему времени нет общепринятой теории изменения языковых признаков. Наиболее распространенной является гипотеза о циклическом изменении по типу: язык изолирующего строя → язык агглютинативного строя → язык флективного строя → язык изолирующего строя (об этом см. монографию В. Крофта [2]).

Однако недавно была предложена альтернативная версия языковых изменений. О. Даль [3] считает, что в языках постоянно генерируются новые признаки, которые ведут борьбу за существование. Многие из них по прошествии небольшого промежутка времени исчезают, некоторым удается распространиться на несколько языков и закрепиться в пуле признаков на более или менее длительный срок. Сохранившиеся признаки называются зрелыми, они могут существовать в языках в неизменном виде неограниченно долго. ЯП-диаграммы дают новые данные в поддержку гипотезы О. Даля.

Рассматривая совокупность языковых признаков и классифицируя их по частоте, естественно выделить центр и периферию – структуры, четко распознаваемые в других разделах лингвистики. Центр состоит из зрелых, имеющих достаточно широкое распространение признаков, изучением которых, собственно, и занимается типология. На ЯП-диаграмме они занимают среднюю и правую часть графика. Периферию же составляют молодые, относительно недавно появившиеся признаки, еще не успевшие широко распространиться и занимающие левую часть графика. Как и остальные признаки, 1-языковые признаки могут подвергнуться вытеснению другими признаком, и в этом случае они исключаются из пула признаков окончательно. Для *n*-языковых признаков при небольших *n* также относительно велика вероятность исчезновения в результате последовательного вытеснения из всех *n* языков.

Эти признаки занимают левую часть ЯП-диаграммы и образуют как раз динамичную периферию, необходимую для эволюции языка. Естественно, что периферия должна быть достаточно большой по размерам. Если каждый язык для обеспечения возможностей развития генерирует хотя бы по несколько новых признаков, то все 315 языков базы данных в совокупности будут иметь более тысячи 1-языковых признаков, что и наблюдается в действительности.

Разумеется, 1-языковые признаки могут и не быть абсолютно новыми, а могут быть заимствованы из каких-либо языков, не включенных в базу данных, в том числе из языков, располагающихся вне области Европы, Северной и Центральной Азии. Но это не меняет картину принципиально, так как даже если бы мы располагали исчерпывающей базой данных, содержащей все языки мира, все равно она должна включать определенное количество новых признаков, появление которых обеспечивает языковое разнообразие и развитие.

Посмотрим, как обстоят дела в этом смысле в двух основных подгруппах группы тюркских языков. Сопоставим огузские и кыпчакские языки, представленные в БД почти одинаковым числом языков – 16 и 15 соответственно.

Несмотря на близость огузских и кыпчакских языки, их ЯП-диаграммы имеют заметные отличия. На диаграмме 2 хорошо видно, что огузские языки имеют значительно большее количество 1-языковых признаков, а кыпчакские – больше высокочастотных признаков (присутствующих во всех или почти всех языках данной группы).

Таким образом, обнаруженное большее число 1-языковых признаков является свидетельством большей активности огузских языков в плане генерации новых признаков либо заимствования их из языков вне рассматриваемой группы (внешних заимствований).

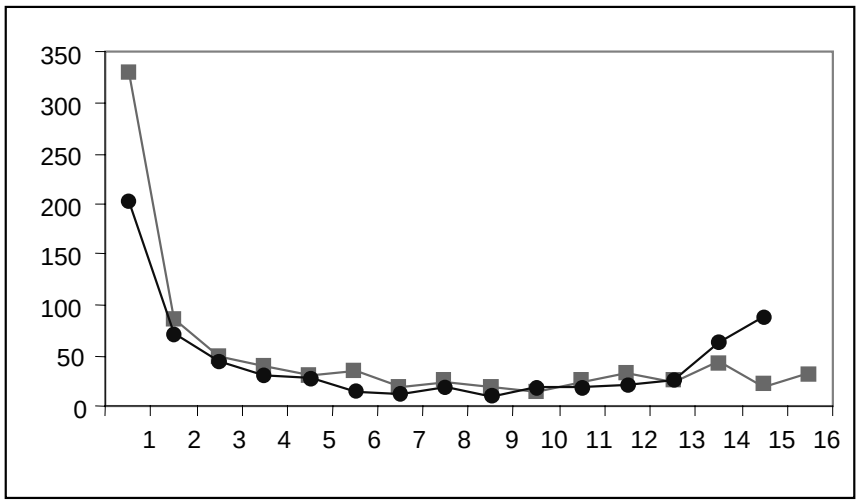

Диаграмма 2. Огузские (линия с квадратиками) и кыпчакские языки

2. Количественные оценки эволюционного процесса

Интенсивность процессов новообразования в той или иной группе языков можно охарактеризовать количественными параметрами. В монографии [1] предложена соответствующая математическая модель. Пусть C – количество новых признаков, появляющихся в единицу времени, α – вероятность заимствования (в пределах рассматриваемой группы) или потери признака языком. Тогда параметр $S = C/\alpha$ характеризует интенсивность процессов изменения грамматической структуры группы языков. Чем S больше, тем более интенсивно идет процесс изменения группы языков. Согласно математической модели эволюции, предложенной в [1], S может быть вычислено по формуле: $S = 2(Y(1) - Y(2))$, где $Y(1)$ и $Y(2)$ – число 1- и 2-языковых признаков соответственно.

Для огузских и кыпчакских языков параметр S принимает следующие значения: $S = 488$ для огузских языков, $S = 262$ – для кыпчакских. Столь большая разница в значениях S означает, что C больше у огузских языков, а α – у кыпчакских.

Большее значение параметра C хорошо коррелирует с большим числом 1-языковых признаков, отмеченным выше. Итак, имеющиеся данные позволяют предположить, что для огузских языков, по сравнению с кыпчакскими, более характерны новообразования, в том числе внешние заимствования. С другой стороны, большее значение α для кыпчакских означает большую интенсивность внутренних заимствований (в пределах своей языковой группы), что заставляет предположить, что кыпчакские языки формировались в меньшем по размеру ареале.

Данные выводы подтверждаются результатами других исследований. В книге [4, с. 22] утверждается: «Факты показывают, что степень смешения огузизмов и кыпчакизмов в отстаивающейся кыпчакской общности проявляется меньше (в смысле контрастности черт), чем в огузской. Последнее объясняется от-

части географическими условиями: отстаивание кыпчакских языков происходило на относительно узкой территории».

Заключение

В статье новые методы компьютерного анализа типологических данных демонстрируются на примере тюркских языков. Установлено, что в языках огузской ветви значительно больше уникальных признаков, существующих лишь в одном языке данной группы. Согласно развитой в монографии [1] модели эволюции языковых признаков это означает большую динамичность огузских языков по сравнению с кыпчакскими, их более активное участие в процессах заимствования признаков у контактирующих с ними языков из иных языковых групп. Полученные результаты подкреплены математическими выкладками, а также находятся в хорошем согласии с фактами, ранее выявленными при анализе лексического состава языков. Предложенный новый метод компьютерного анализа с применением типологических баз данных может быть использован как дополнительный к традиционным методам компаративистики.

Работа выполнена при поддержке РФФИ (проект № 07-06-00229а).

Summary

V.D. Solovyev, V.R. Bajrasheva. Statistical parameters for kypchak and oguz languages.

In the paper frequencies of grammatical features emergence are compared for two groups of Turkic languages. The “Languages of the World” database is used for analysis. It is shown that for oguz languages group there are much more features, which occur exactly in one language of the group. The result is based on the analysis of a great amount of data and is statistically reliable. A hypothesis explaining this phenomenon is suggested.

Литература

1. *Поляков В.Н., Соловьев В.Д.* Компьютерные модели и методы в типологии и компаративистике. – Казань: Казан. гос. ун-т, 2006. – 207 с.
2. *Croft W.* Typology and Universals. – Cambridge: Cambridge Univ. Press, 1990. – 310 p.
3. *Dahl Ö.* The grouth and maintenance of linguistic complexity. – Amsterdam: John Benjamins, 2004. – 287 p.
4. Языки мира: Тюркские языки / Отв. ред. Э.Р. Тенищев. – М.: Индрик, 1997. – 544 с.

Поступила в редакцию
03.10.07

Соловьев Валерий Дмитриевич – доктор физико-математических наук, профессор, проректор по информатизации Казанского государственного университета.
E-mail: *maki.solovyev@mail.ru*

Байрашева Венера Рустамовна – кандидат физико-математических наук, доцент кафедры теоретической кибернетики Казанского государственного университета.