Новые методы в генеалогической классификации языков и лингвистической реконструкции

Алексей Касьян, Георгий Старостин

В заметке описывается проект Лаборатории востоковедения и сравнительно-исторического языкознания Школы актуальных гуманитарных исследований РАНХиГС, связанный с формализацией генетической классификации языков.

Текущее положение дел в мировой науке

Одним из наиболее актуальных вопросов современного исторического языкознания как дисциплины, составляющей комплекс наук о предыстории человечества, является обоснование гипотез так называемого дальнего родства языковых семей, которые претендуют на реконструкцию языковой, а в связке с молекулярной биологией и генетикой — и этнополитической истории человечества на отрезках, превышающих пять-шесть тысяч лет. Пять-шесть тысяч лет до настоящего времени — это традиционно принимаемая глубина общепризнанных и хорошо изученных языковых семей, например, индоевропейской, уральской, сино-тибетской (видимо, тут мы имеем дело с временны́м порогом, после которого очевидность языкового родства начинает резко утрачиваться).

Традиционные методы сравнительно-исторического языкознания, разработанные для семей менее глубокого уровня, оказываются недостаточными для надежного обоснования таких гипотез и требуют серьезной доработки как на базе опыта, накопленного в ходе исторического изучения различных языковых семей планеты, так и с учетом новейших достижений в области филогенетического моделирования.

В последние десятилетия формальные методы филогенетической классификации, перенесенные в лингвистику из молекулярной биологии, переживают научный бум. См., например, такие обзоры применения современных филогенетических алгоритмов в сравнительно-исторической лингвистике (в основном речь идет именно о лексикостатистике и глоттохронологии): [McMahon, McMahon 2005; Nichols, Warnow 2008; Heggarty et al. 2010]. В частности, в связи с удешевлением и распространением мощных компьютерных станций все большую популярность приобретают признаковые методы филогении (вроде байесовской техники Монте-Карло с цепями Маркова и алгоритма максимальной парсимонии), а дистантные методы (вроде метода ближайших соседей или попарного внутригруппового невзвешенного среднего) отчасти отходят на второй план. Подробнее об этих методах см.: [Makarenkov et al. 2006]. Входным материалом при таком анализе служат многозначные или бинарные матрицы, т. е. двумерные таблицы, где каждый таксон (язык) охарактеризован по всему набору признаков. Бинарные матрицы содержат только бинарные признаки (с состояниями 0 или 1), а многозначные матрицы имеют хотя бы один многозначный признак. Признаки на практике используются самые разные: от лексических до культурно-антропологических, хотя предпочтение, конечно, отдается базисной лексики (так называемому списку Сводеша).

Принципиальные этапы исследования

Задача формализации генетической классификации языков может быть разделена на несколько принципиальных этапов.

1. Подготовка максимально качественного языкового материала, который будет подаваться на вход. Важность очистки входных данных ни в коем случае нельзя недооценивать, как бы ни хотелось сэкономить человеко-часы на данной процедуре. Дело в том, что компьютерная программа породит генетическую классификацию из любого подаваемого материала, но робастность получаемых дендрограмм и их историческая надежность зависят от адекватности лингвистических данных (как это правило традиционно формулируется для биологической филогении, «Garbage in, garbage out»).
2. Апробация биологических методов на конвенциональных группах и семьях языков, т. е. на языках, о факте родства которых и о внутренней классификации которых среди специалистов наблюдается научный консенсус. Это, к примеру, такие группы, как славянская, германская, лезгинская, с некоторыми оговорками — уральская семья. Индоевропейская семья в этот список уже не входит: ее состав учеными не оспаривается, но общепринятой классификации групп внутри индоевропейской семьи пока нет. Серия таких тестов должна указать на слабые и сильные стороны того или иного метода и выявить основные подводные камни при переносе биологических приемов на лингвистический материал.
3. Построение гипотез дальнего языкового родства, т. е. родства между языковыми семьями, относящегося к доисторической эпохе.

Практические проблемы

Несмотря на десятки регулярно появляющихся статей по формальной классификации тех или иных языковых групп, в мировой практике наблюдаются существенные лакуны.

Во-первых, многие, если не большинство авторов не вполне осознают важность тщательной подготовки входных данных (в основном лексических списков Сводеша). Например, классификации индоевропейской семьи, предложенные в [Gray, Atkinson 2003; Bouckaert et al. 2012], некритически базируется на 200-словных списках из [Dyen et al. 1997]. Однако база данных [Ibid.] содержит множество лексикографических ошибок (см. [Kushniarevich et al. 2015]). Как результат, в указанных классификациях мы видим явно неприемлемые узлы вроде белорусско-польского единства.

Связано это с разницей узусов биологии и лингвистики. В биологии опубликованные данные, скажем, по морфологии того или иного вида или по секвенированию генома, считаются надежными, их можно непосредственно использовать в филогении. Совершенно иначе обстоит дело в лингвистике, где, например, категорически не рекомендуется использовать лексические списки, механически извлеченные из обратных словарей. Напротив, качественная подготовка стословного списка одного языка под стандарт конкретного исследования может занять несколько недель работы квалифицированного лингвиста.

Во-вторых, довольно плохо обстоит дело с тестированием различных методов на консенсусном материале. Например, в работе [Nakhleh et al. 2005] основные филогенетические методы применены к индоевропейской семье. Они дают различающиеся деревья, но мы не можем сказать, какой из методов лучше других справился с реконструкцией филогении, так как общепринятой классификации индоевропейской семьи не существует. Пока полноценным тестированием можно считать такие публикации, как [Barbançon et al. 2013] (на вход подавались искусственно смоделированные лингвистические данные) и [Kassian 2015] (110-словники лезгинских языков).

Цель Лаборатории

Исходя из необходимости закрыть вышеописанные лакуны, основную цель исследования нашей Лаборатории мы можем сформулировать так: разработать и апробировать усовершенствованную методику построения оптимального сценария генетического родства языковых семей на средних и глубоких хронологических уровнях, сочетающую элементы традиционного сравнительно-исторического метода с новейшими достижениями исторической типологии, лексикостатистики и формальных алгоритмов.

Исследования Лаборатории базируются на лексических данных нашего активно развивающегося онлайн-проекта «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database» (сокращенно ГЛБД/GLD; см. [Starostin et al. 2011]). Идеологическую основу проекта составляют следующие положения.

1. Сравнение лексики — это надежный инструмент для генеалогической классификации языков. Иногда утверждается, что приоритет в подобных задачах должны иметь грамматические (фонетические, морфологические, синтаксические) признаки. Однако мы предполагаем, что грамматические данные следует использовать с осторожностью, так как, во-первых, эти признаки не универсальны; во-вторых, они легко могут образовывать вторичные ареальные изоглоссы (особенно если речь идет о языках, родство которых еще ощущается носителями), причем выявить источник инновации часто оказывается затруднительно; в-третьих, грамматические признаки образуют систему, т. е. изменение одного признака с высокой вероятностью влечет за собой изменение других признаков. Для лексических же признаков эти недостатки характерны в значительно меньшей степени.
2. Точность реконструкции филогенетического дерева зависит в первую очередь не от математического метода, а от степени очистки входных данных, иными словами, не от труда компьютера, а от труда лингвиста, кропотливо анкетирующего индивидуальные диалекты по принятому списку признаков.

Проект «Глобальная лексикостатистическая база данных»

По своей форме ГЛБД не представляет собой одну, единую базу данных — это иерархическая система, включающая списки слов разных уровней, от высшего до низших. Такая структура не только облегчает работу с огромнейшим объемом информации, но и находится в строгом соответствии с концепцией генеалогического древа, согласно которой из языков-предков произрастают многочисленные языки-потомки, на основе которых методами исторической лингвистики можно реконструировать их общий язык-предок.

Первый уровень составляют сравнительно небольшие базы данных, каждая из которых содержит списки слов языков, разделившихся, по предположительным оценкам, не более трех тысяч лет назад, близкое родство которых не вызывает сомнений, а также список слов праязыка, являющегося их общим предком. Типичные примеры таких баз — германская, тюркская, полинезийская, северно-койсанская и т. п. За генетическими общностями такого уровня закреплено традиционное название языковой *группы*.

Второй уровень — базы, содержащие списки только реконструированных слов праязыков, которые достоверно или хотя бы предположительно родственны между собой. Реальность существования таких праязыков обычно не подвергается сомнению в лингвистическом сообществе, а время их выделения из общего языка-предка — не более шести тысяч лет назад. Базы второго уровня включают также список слов праязыка, являющегося общим предком представленных в данной базе праязыков. К числу типичных примеров относятся индоевропейские, уральские, австронезийские, северно-кавказские и др. общности. Такие генетические общности мы, опять-таки традиционно, называем языковыми *семьями*.

Третий уровень составляют базы, в которых сопоставляется лексика нескольких праязыков разных *семей* — в случае, если существует предположение, что между этими семьями имеется очень глубокое генетическое родство. Поскольку такие сверхглубокие генетические связи часто подвергаются серьезному сомнению (особенно специалистами, убежденными в том, что ни сравнительно-исторический метод, ни какие-либо альтернативные подходы не позволяют получить убедительных результатов, когда речь идет о хронологической глубине, превышающей шесть-восемь тысяч лет), создание и анализ гипотетических прасписков для столь глубоких таксонов является непременным условием подтверждения их исторической реальности. Типичные примеры — ностратические, сино-кавказские, афроазиатские, нигер-конголезские и т. п. языки; такого рода общности мы называем *макросемьями*.

На данный момент в онлайн-компоненте ГЛБД представлены почти исключительно базы первого уровня, но со временем, по мере увеличения числа обработанного материала и формально верифицированных гипотез языкового родства, планируется последовательная интеграция их сначала в базы второго, а затем и третьего уровня. Конечная цель — сведение всех языков планеты к абсолютному минимуму таксонов, которые могут быть обоснованы с помощью лексикостатистической методологии и, тем самым, тестирование хронологических пределов действия лексикостатистического метода как такового.

Задачи, решаемые в Лаборатории

Основные фундаментальные и прикладные задачи, решаемые в рамках исследования в нашей Лаборатории, можно сформулировать так:

1. интеграция данных историко-фонетической и историко-семантической типологии в процедуру доказательства глубинного языкового родства;
2. совершенствование используемых в компаративистской практике алгоритмов статистического анализа сравнительных данных базисной лексики;
3. внедрение полученных результатов в программную оболочку компьютерной лингвистической среды STARLING и их апробация на базах данных по крупным языковым семьям Евразии, Африки и Америки.

В результате исследования планируется значительно усовершенствовать формальную методологию языковой классификации, что позволит предлагать достоверные сценарии исторического развития современной языковой ситуации на протяжении последних 10–12 тысяч лет. Разрабатываемая методология, интегрирующая достижения классического сравнительно-исторического языкознания, данные лингвистической типологии и современные статистические алгоритмы, не имеет реальных прецедентов в мировом языкознании.

Первая публикация текста: *Шаги: Журнал Школы актуальных гуманитарных исследований* 1(1), 2015, с. 206-212.

Литература:

Barbançon, F., Evans, S. N., Nakhleh, L., Ringe, D., Warnow, T. (2013). An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica, 30*(2), 143–170.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science, 337*, 957—960. [With corrections and revised supplementary materials in: *Science, 342. 2013, December 20*, 1446].

Dyen, I., Kruskal, J., Black, P. (1997). Comparative Indo-European Database. Last modified on Feb 5, 1997. http://www.wordgumbo.com/ie/cmp [accessed 15.04.2015].

Gray, R. D., Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature, 426*, 435–439.

Heggarty, P., Maguire, W., McMahon, Al. (2010). Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B, 365*, 3829–3843.

Kassian, A. (2015). Towards a formal genealogical classification of the Lezgian languages (North Caucasus): testing various phylogenetic methods on lexical data. *PLoS ONE* 10(2): e0116950, 2015. doi:10.1371/journal.pone.0116950.

Kushniarevich, A., Utevska, O., Dibirova, K., Uktverite, I., Agdzhoyan, A., Chuhryaeva, M., Möls, M., Kovačević, L., Pshenichnov, A., Frolova, S., Shanko, A., Metspalu, E., Reidla, M., Tambets, K., Tamm, E., Koshel, S., Atramentova, L., Churnosov, M., Kucinckas, V., Evseeva, I., Davydenko, O., Tegako, L., Yunusbaev, B., Khusnutdinova, E., Marjanović, D., Rudan, P., Rootsi, S., Zaporozhchenko, V., Yankovsky, N., Kassian, A., Dybo, A., The Genographic Consortium, Tyler-Smith, Ch., Balanovska, E., Metspalu, M., Kivisild, T., Villems, R., Balanovsky, O. (2015). Genetic heritage of the Balto-Slavic speaking populations: a synthesis of autosomal, mitochondrial and Y-chromosomal data. *PLoS ONE* 10(9): e0135820, 2015. doi:10.1371/journal.pone.0135820.

Makarenkov, V., Kevorkov, D., Legendre, P. (2006). Phylogenetic network construction approaches. In: D. K. Arora, R. M. Berka, G. B. Singh (eds.). *Applied Mycology and Biotechnology, 6: Bioinformatics*, 61–98. Amsterdam; Boston: Elsevier.

McMahon, A., McMahon, R. (2005). *Language classification by numbers.* Oxford: Oxford Univ. Press. xviii + 265 p.

Nakhleh, L., Warnow, T., Ringe, D., Evans, S. N. (2005). A comparison of phylogenetic reconstruction methods on an IE dataset. *The Transactions of the Philological Society, 103*, 171–192.

Nichols, J., Warnow, T. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass, 2(5)*, 760–820.

Starostin, G. et al. (2011). *The Global Lexicostatistical Database*. http://starling.rinet.ru/new100 [accessed 15.04.2015].