МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12–06–06061–г),
гранта РГНФ (проект № 12–04–14154–г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:
В. А. Баранов, д-р филол. наук, проф.
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

И741 **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

ФИЛОГЕНЕТИКА В ЛИНГВИСТИКЕ: ЯЗЫКИ, ДЕРЕВЬЯ, ЭВОЛЮЦИЯ

С. В. Русаков, Д. М. Нурбакова

Пермский государственный национальный исследовательский университет, Пермь

Languages evolve through time. But the past of languages could be untangled like the species are put in the tree of life. Just as DNA is used in biology to uncover the relationships between species, the most stable linguistic features (e.g. the basic vocabulary, etc.) seem to identify the phylogeny of languages. Here, we evaluate the phylogenetic relationships between Slavic languages using a group of methods. Our results correlate with the traditional classification and prove the stability of the chosen methods.

> *It might be that some very ancient language had altered little, and had given rise to few new languages, whilst others […] had altered much, and had given rise to many new languages and dialects.*
> Ch. Darwin ‘The Origin of Species’, 1859

Язык — один из важнейших феноменов развития человека, который меняется вместе с ним на протяжении всей истории человека. То, насколько далеко в прошлое языков мы можем заглянуть, во многом зависит от скорости изменения языковых единиц. Некоторые элементы языковой структуры являются достаточно устойчивыми и могут служить своего рода генами, последовательностями ДНК, на основе сравнения которых можно построить дерево эволюции — филогенетическое древо.

Филогенетическое древо — граф, отражающий эволюцию различных видов или других сущностей, имеющих общего предка. Различают три класса вершин филогенетического дерева: листья, узлы и корень. Листья — это конечные (концевые) вершины, то есть те, которым инцидентно только одно ребро. Каждый лист отображает некоторый вид или сущность, например, современный язык. Каждый узел представляет эволюционное событие: разделение предкового вида на два или более, которые в дальнейшем эволюционировали независимо. Корень — выделенная вершина дерева, отображающая общего предка всех рассматриваемых сущностей. Рёбра филогенетического дерева принято называть «ветвями». Взаимное расположение ветвей называется топологией.

Выявление филогенетических отношений языков — одна из сторон предмета исследования сравнительно-исторического языкознания. Однако в последнее время, помимо традиционных лингвистических методов, стали применяться математические методы, разработанные изначально для решения задач эволюционной и молекулярной биологии. Однако следует отметить, что приложение данных методов к лингвистике обладает своими особенностями, связанными с предметной областью. Стоит также помнить о том, что любое дерево представляет собой лишь одну из гипотез взаимоотношений между таксонами, поскольку все модели — лишь упрощённое представление реальных процессов, имеющих весьма сложную структуру и природу.

Существуют различные методы филогенетической реконструкции, которые делятся на две группы: *дистанционно-матричные методы* и *статистические методы*. Первая группа методов построена на расчете матрицы расстояний, при этом расстояние понимается как мера различия. Примерами дистанционно-матричных методов могут служить: метод невзвешенного попарного среднего (UPGMA) [Michener and Sokal, 1958], метод связывания ближайших соседей (NJ) [Nei and Saitou, 1987], метод построения филогенетической сети NeighbourNet [Moulton and Bryant, 2004]. Статистические (или дискретные) методы работают непосредственно с последовательностями данных, а не с коэффициентами их сходства, и решают задачу оптимизации. Среди статистических методов можно назвать метод максимальной экономии (MP) [Tassy and Darlu, 1993], метод максимального правдоподобия, метод Байеса [Holden et al., 2005].

Задача данного исследования заключалась в сравнении филогенетических деревьев славянских языков, построенных с использованием следующих методов филогенетической реконструкции:

– иерархическая кластеризация методом взвешенного среднего (WGMA) на основе корреляционного расстояния, реализованная нами в пакете Wolfram Mathematica 7.0;

– построение филогенетической сети методом NeighbourNet, реализованным в программе SplitsTree [Huson and Bryant, 2006].

– иерархическая кластеризация на основе расстояния Левенштейна [Левенштейн, 1965] и Дамерау–Левенштейна [Chakrabarti, 2003], реализованная в пакете Wolfram Mathematica 7.0.

– построение филогенетического дерева методом Байеса с использованием программного продукта MrBayes [MrBayes].

В данном исследовании рассматриваются двенадцать современных славянских языков: словенский, нижне- и верхнелужицкий, чешский, словацкий, украинский, белорусский, польский, русский, македонский, болгарский и сербохорватский. Согласно традиционной классификации славянских языков, выделяются следующие группы:

1. *восточно-славянская*: русский, белорусский, украинский, русинский;

2. *южно-славянская*, которая в свою очередь делится на:

– *западную подгруппу*: сербохорватский и словенский;

– *восточную подгруппу*: македонский, болгарский, церковнославянский;

3. *западно-славянская*, которая подразделяется на:

– *лехитские языки:* польский, кашубский;

– *лужицкие языки:* нижнелужицкий и верхнелужицкий;

– *чехословацкие языки*: чешский, словацкий.

Материалом для исследования послужили лексические единицы, содержащиеся в списках Сводеша современных славянских языков, и некото-рые грамматические признаки (сохранение двойственного числа, наличие трех родов, наличие определенного артикля, суффиксация как наиболее час-тотная форма словообразования, использование супплетивных форм для вы-ражения видовременных значений глагола).

Источником лексических дан-ных для 12 славянских языков послужила база данных, собранная Краскалом и др. [Dyen et al., 1992], где словоформы классифицируются на когнаты (од-нокоренные слова, имеющие общее происхождение и похожее звучание в двух и более самостоятельных языках), сомнительные когнаты и “не когна-ты”. В случае применения метода NeighbourNet и Байеса словоформы каждо-го из анализируемых языков были закодированы и преобразованы в бинар-ные цепочки, каждый элемент которых соответствовал признаку нали-чия/отсутствия соответствующей когнаты (элемент кодировался “1” или “0” соответственно) в конкретном языке. Таким образом, для списка Сводеша из 200 значений число когнат составило 476. Аналогичным образом были закодированы грамматические признаки.

Полученные результаты во многом схожи. На всех полученных деревьях можно выделить три основные ветви. Первая соответствует южным славянским языкам и включает в себя болгарский, македонский, сербохорватский и словенский языки. При этом наблюдается раннее ответвление словенского языка и, наоборот, более позднее разделение македонского и болгарского. При добавлении к современным славянским языкам староцерковнославянского языка можно увидеть, что староцерковнославянский стоит у основы южной ветви. Вторая ветвь сформирована из западных языков: чешского, словацкого, а также пары тесно связанных лужицких языков. И наконец, третья ветвь состоит из русского, рано отделившегося от остальных, польского, украинского и белорусского. В отличие от традиционной классификации, большинство методов группируют польский язык с белорусским, украинским и русским, а не западными языками. Можно также отметить, что западнославянские языки продемонстрировали в некоторой степени разрозненность: лужицкая и чехословацкая группы отстают друг от друга, что является возможным следствием как исторического развития народов, говорящих на этих языках, так и влияния других языков. Что касается восточнославянских языков, можно заметить, что их разделение произошло достаточно рано, что отражается в отделении русского языка, подвергшегося влиянию финно-угорских языков и новгородского диалекта. Отметим, что добавление грамматических признаков качественно не влияет на топологию филогенетического дерева, построенного методом Байеса, и филогенетической сети, построенной методом NeighbourNet.

В ходе исследования была использована целая группа методов филогенетической реконструкции, однако были получены качественно непротиворечивые результаты. Данный факт свидетельствует об устойчивости и адекватности применяемых методов для решения поставленной задачи.

В ходе дальнейших исследований планируется сравнительное изучение славянских языков с другими группами языков, распространенных на территории России, в частности финно-угорской и тюркской группами. Расширение перечня анализируемых языков связано с тем, что данные группы языков распространены на территории одной страны и развивались в условиях взаимного влияния.

Кроме того, при анализе лексического материала представляется целесообразным учитывать звуковые соответствия и фонетические изменения в славянских языках. Также возможно пополнение материала исследования грамматическими признаками, подверженными наименьшим изменениям с течением времени.

Список литературы

Chakrabarti, 2003 — Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, 2003. Pp. 71–72.

Dyen et al., 1992 — Dyen I., Kruskal J.B., Black P. An Indo-European classification: a lexicostatistical experiment // Transactions of the American Philosophical Society. 1992. Vol. 82. Pp. 1–132.

Holden et al., 2005 — Holden C.J., Meade A., Pagel M. Comparison of maximum parsimony and Bayesian Bantu language trees // The Evolution of Cultural Diversity: a phylogenetic approach. London, 2005. Pp. 53–66.

Huson and Bryant, 2006 — Huson D.H., Bryant D. Application of Phylogenetic Networks in Evolutionary Studies // Molecular Biology and Evolution. 2006. Vol. 23. № 2. Pp. 254–267.

Michener and Sokal, 1958 — Michener C., Sokal R. A statistical method for evaluating systematic relationships // University of Kansas Science Bulletin. 1958. № 38. Pp. 1409–1438.

Moulton and Bryant, 2004 — Moulton V., Bryant D. An Agglomerative Method for the Reconstruction of Phylogenetic Network // Molecular Biology and Evolution. 2004. Vol. 21. № 2. Pp. 225–265.

MrBayes — MrBayes: Bayesian Inference of Phylogeny. Веб-сайт: http://mrbayes.csit.fsu.edu/index.php.

Nei and Saitou, 1987 — Nei M., Saitou N. The neighbor-joining method: a new method for reconstructing phylogenetic trees // Molecular Biology and Evolution. 1987. Vol. 4. № 4. Pp. 406–425.

Tassy and Darlu, 1993 — Tassy P., Darlu P. La Reconstruction Phylogénétique. Concepts et Méthodes. Paris, 1993.

Левенштейн, 1965 — Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады АН СССР. 1965. Т. 163. № 4. С. 845–848.