DOI 10.37882/2223-2982.2021.12.17

ФИЛОГЕНЕТИЧЕСКИЕ ДЕРЕВЬЯ НА ОСНОВЕ ГРАММАТИКИ: ПЕРСПЕКТИВЫ И НЕДОСТАТКИ[1]

GRAMMAR-BASED PHYLOGENETIC TREES: PROSPECTS AND DISADVANTAGES

E. Makarova

Summary: The present work aims at proving that the information on the language structure can be used as a tool to classify it. Despite grammar being more prone to borrowings and changes than lexical and phonetic data, the history of linguistics provides a big number of examples when hypotheses on the kinship of languages were suggested basing on one or several grammar features. Two sets of languages from the typological database "Languages of the World" of IL RAS are used to calculate the Hamming distances and build two phylogenetic trees: one for eight different families (Uralic, Mongolic, Turkic, Northeast Caucasian, Mande, Dravidian, Semitic, and Indo-European), and the other for five branches of the Indo-European family (Slavic, Germanic, Romance, Baltic, and Anatolian). Further analysis of the results will allow us to answer the question: how accurately grammar-based trees represent the real state of affairs and, consequently, can they be used to classify languages.

Keywords: quantitative linguistics, phylogenetics, "Languages of the World" of IL RAD database, computer linguistics.

Макарова Елена Андреевна
М.н.с., Институт языкознания Российской Академии Наук, г. Москва
antaresselen@mail.ru

Аннотация: В данной статье предпринимается попытка доказать, что структурные данные о языке могут служить инструментом его классификации. Несмотря на то, что грамматика более склонна к заимствованию и изменению, чем лексикофонетические данные, известно большое количество примеров, когда гипотезы о родстве языков высказывались на основании одного или нескольких грамматических признака. На материале двух выборок языков из типологической базы данных «Языки мира» ИЯз РАН будет попарно для всех языков рассчитано расстояние Хэмминга и построены два филогенетических дерева: одно для восьми различных семей (уральская, монгольская, тюркская, нахско-дагестанская, манде, дравидийская, афроазиатская, семитская и индоевропейская), второе – для пяти ветвей индоевропейской семьи (славянские, германские, романские, балтийские и анатолийские). Дальнейший анализ полученных результатов позволит ответить на вопрос, насколько деревья, основанные исключительно на грамматических данных, соответствуют действительности и, следовательно, могут ли они использоваться для классификации языков.

Ключевые слова: Квантитативная лингвистика, филогенетика, база данных «Языки мира» ИЯз РАН, компьютерная лингвистика.

Введение

Задача установления родства языков на основании грамматических данных имеет свою историю. До начала XXI века в своих гипотезах о родстве языков многие лингвисты руководствовались одним или несколькими структурными признаками. Например, в 1939 году Н. Трубецкой сделал попытку определить набор грамматических признаков, свойственных только индоевропейским языкам, и выделил шесть структурных признаков [1]. Позднее, в 1953 году, эта гипотеза была опровергнута Э. Бенвенистом [2], который обнаружил североамериканский язык такелма, обладающий всеми шестью грамматическими признаками, выделенными Трубецким. При этом такелма оказался языком-изолятом и, по данным сравнительно-исторического языкознания, не принадлежал к индоевропейской семье.

В современной лингвистике доминирует точка зрения, что родство между языками может быть установлено только на основе лексикофонетических данных, в то время как грамматические характеристики языка считаются более склонными к заимствованию и, таким образом, не способны стать надежной основой для установления родственных связей между языками.

В 2009 году было проведено исследование [3], одной из задач которого было сравнение филогенетических деревьев, построенных на лексикофонетических данных проекта ASJP и на структурных данных двух типологических баз данных - World Atlas of Language Structures (WALS) [4] и базы данных «Языки мира» ИЯз РАН [5]. В работе было показано существенное преимущество лексикофонетических данных из ASJP перед грамматическими данными из WALS и базы данных «Языки Мира» ИЯз РАН в их способности описывать родство языков.

В 2010 году в работе [6] на основании статистических расчетов показывалось, что база данных WALS не может рассматриваться как источник валидных данных для филогенетических расчетов. Авторы наглядно показали, что WALS недооописан, однако они не доказали того факта, что родство не может передаваться с помощью грамматических данных.

1 Исследование было поддержано грантом РФФИ №19-012-00476 А

В данной работе будет предпринята попытка доказать, что типологическая база данных может быть использована для филогенетических расчетов при условии высокой степени описания включенных в нее языков. Для этой задачи была выбрана база данных «Языки мира» ИЯз РАН. В 2020 году была выпущена ее новая версия [7], в которой кардинальным образом был пересмотрен способ представления данных, что сделало «Языки мира» более удобным инструментом для проведения квантитативных исследований.

Для исследования были сделаны две выборки языков. Первая состоит из 33 языков, относящихся к восьми различным семьям, вторая выборка включает в себя 51 индоевропейский язык, представляющий 5 групп: романскую, германскую, анатолийскую, балтийскую и славянскую.

Материалы и методы

В данной работе используется материал 4-ой версии базы данных «Языки мира» ИЯз РАН. Она была выпущена в качестве десктоп-версии в 2020 году и доступна для скачивания с официального сайта ИЯз РАН: https://iling-ran.ru/web/ru/news/201008_langworld. Главным отличием от всех предыдущих версий стал переход от иерархической системы представления данных к парадигматической. Иерархическая структура, представлявшая собой дерево бинарных признаков, часто подвергалась критике ввиду неоднозначности обозначений (например, «0» мог означать как отсутствие признака в языке, так и отсутствие информации). Кроме того, это значительно затрудняло проведение квантитативных исследований без использования вспомогательных программ. В текущей версии базы данных информация о каждом языке представлена в виде парадигм: признак – варианты значения признака. Всего признаковое пространство охватывает 14 разделов, начиная от фонемного состава и заканчивая особенностями построения сложного предложения. На данный момент в базу данных включено свыше 280 языков, продолжается активная работа по добавлению новых языков. К концу 2021 года планируется официальный выпуск онлайн-версии базы данных «Языки мира» ИЯз РАН.

В работе [7], целью которой ставилась классификации некоторых языков-изолятов с использованием структурной информации о языке, было проведено попарное сравнение 173 языков, представленных в базе данных «Языки мира» ИЯз РАН. Для всех этих языков был рассчитан процент совпадающих значений признаков. Для дальнейшего анализа было отобрано 464 пар языков: как правило, по три пары для каждого языка с наивысшим процентом совпадающих значений признаков. Из 464 пар языков только 20 оказались неродственными языками, что составило 4,3%. Это позволило предположить, что и филогенетические деревья, построенные на основе данных о грамматической структуре языков, в достаточной степени достоверно отразят эволюционные взаимосвязи между языками.

В настоящем исследовании планируется построить два филогенетических дерева. Одно дерево представит собой выборку наиболее полно описанных языков для следующих семей, включенных в базу данных «Языки мира» ИЯз РАН:

- Тюркская (башкирский, казахский, чувашский);
- Монгольская (монгольский, калмыцкий, дунсянский);
- Уральская (карельский, коми-пермяцкий, удмуртский);
- Нахско-дагестанская (лезгинский, агульский, рутульский);
- Манде (яурэ, какабе, мандинка);
- Дравидийская (телугу, каннада, малалаям);
- Семитская (древнееврейский, иудейско-палестинский арамейский, современный иврит);
- Индоевропейская (болгарский, белорусский, сербохорватский, испанский, сефардский, французский, английский, африкаанс, нидерландский, милийский, лувийский, ликийский).

Второе дерево будет построено для следующих пяти ветвей индоевропейских языков: германская, романская, славянская, балтийская, анатолийская. Результаты позволят ответить на вопрос, насколько качественно такие деревья отражают эволюционные взаимосвязи между представителями разных семей и ветвей (групп), и, следовательно, может ли информация о структуре языка использоваться как инструмент классификации языков.

Основной метод, использованный для проведения данного исследования, - лингвистическая филогения [8]. Филогения нашла широкое применение сначала в генетике [9], а впоследствии начала активно использоваться и в лингвистике как метод установления родства языков. К одним из ранних использований достижений филогении в сравнительном языкознании можно отнести работы [10], [11]

Лингвистическая филогения – это универсальный и общепризнанный метод сравнения и классификации языков. При условии использования множества хорошо изученных и описанных языков, построенные филогенетические деревья отражают качество данных. Именно по этой причине лингвистическая филогения была выбрана как метод для тестирования выборки языков из базы данных «Языки мира».

Для построения дерева было рассчитано расстояние Хэмминга [12], т.е., процент несовпадающих значений признаков и построена матрица расстояний. Далее, с помощью системы MEGA11 [13] были построены филогенетические деревья для: 1) 33 наиболее полно описанных языков-представителей 8 различных семей и 2) для 51 индоевропейского языка – представителя

5 различных групп. Построенные деревья и их анализ представлены далее.

Результаты

Для построения первого дерева был выбран 21 язык из следующих 7 семей: тюркская, монгольская, уральская, нахско-дагестанская, манде, дравидийская, семитская, и 12 языков – представителей 4 групп индоевропейских языков: анатолийские, германские, романские и славянские. Для второго дерева были проанализированы данные о 51 индоевропейском языке. Для языков было рассчитано расстояние Хэмминга и с помощью программы MEGA11 методом присоединения соседей [14] построено филогенетическое дерево. Результаты представлены на рис. 1 и рис. 2.

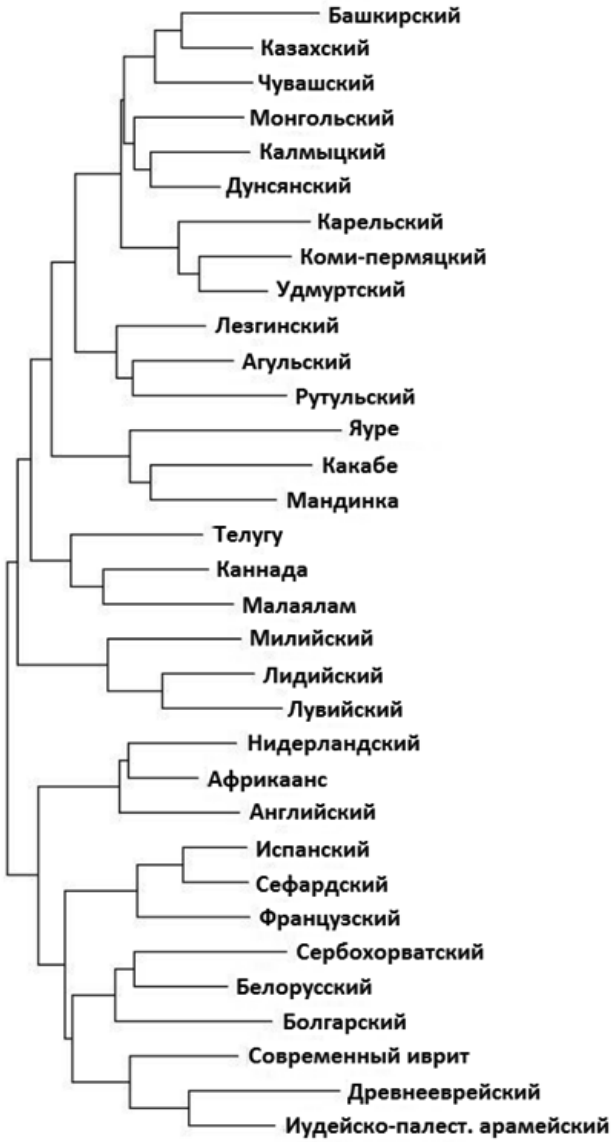

Рис. 1. Филогенетическое дерево для 33 языков из 8 семей, представленных в базе данных «Языки мира» ИЯз РАН

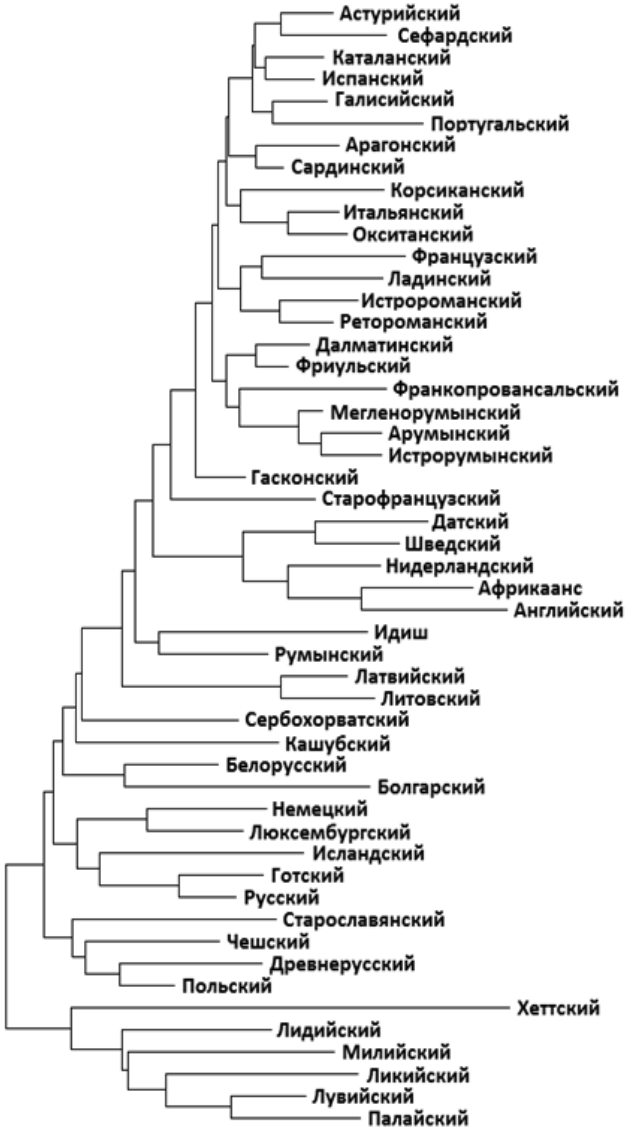

Рис. 2. Филогенетическое дерево для 51 индоевропейского языка, представленного в базе данных «Языки мира» ИЯз РАН

Обсуждение

Как видно из рис. 1, в целом, филогенетическое дерево для 33 языков верно распределяет эти языки по семьям. Однако рассмотрим это дерево более подробно. Тюркские языки (башкирский, казахский, чувашский) отражают общепринятую точку зрения о том, что распад пратюркского языка начался с отщепления чувашского языка от других языков [15]. То же самое касается уральских языков: филогенетическое дерево показывает общего предка (пермский праязык) коми-пермяцкого и удмуртского языка [16].

Кроме того, близость тюркских и монгольских языков на дереве не противоречит достаточно популярной гипотезе об алтайской семье, приверженцы которой счи-

тают сходство между двумя вышеуказанными семьями следствием происхождения от общего праалтайского языка [17]. Также стоит отметить соседство тюркских и монгольских языков с уральскими, с которыми эти семьи имеют ряд общих черт, в свое время даже послуживших основанием для выдвижения урало-алтайской гипотезы.

Распределение языков манде на дереве соответствует принятой классификации, согласно которой языки мандинка и какабе относятся к западной ветви, а язык яурэ – к восточной. Результаты для нахско-дагестанских языков противоречат современной классификации [18], которая относит лезгинский и агульский языки к восточнолезгинской группе, а рутульский – к рутульско-цахурской.

Достоверность данных, полученных для нескольких ветвей индоевропейских языков, будет рассмотрена далее, в сравнении с филогенетическим деревом, построенным для выборки из 51 романского, славянского, германского, балтийского и анатолийского языка

В отличие от первого дерева, где все языки были сгруппированы по семьям, филогенетическое дерево для индоевропейских языков демонстрирует значительное расхождение с общеизвестными фактами. Например, несмотря на то что большинство романских языков объединено в одну ветвь дерева, распределение внутри этой ветви чаще всего некорректно. Например, французский, франкопровансальский и старофранцузский языки (галло-романская подгруппа) находятся на дереве на значительном удалении друг от друга. При этом французский язык объединен с ладинским, франкопровансальский – с балкано-романской подгруппой, а старофранцузский образует отдельную ветвь. Более того, румынский язык был объединен в одну ветвь с идишем.

Германские языки, в отличие от романских, не образуют на дереве отдельной ветви, объединяясь со славянскими и балтийскими языками. Например, готский и русский языки образуют ветвь с общим узлом

Анатолийские языки также демонстрируют некоторые расхождение с фактами. Как следует из дерева, первым из этой ветви отделился хеттский язык, после него – лидийский. Это соответствует действительности. Однако, по данным работы [19], далее произошло отщепление палайского языка, и, на последнем этапе – лувийской подгруппы (лувийского, милийского и ликийского языков). Построенное в данной статье дерево показывает, что палайский язык отделился последним.

На основании полученных результатов можно сделать вывод, что филогенетические деревья, построенные исключено на основании структурных данных о языках, не отражают реальную картину мира. В определенной степени они справляются с задачей общей классификации языков по семьям, однако ответить на вопрос об эволюции языков, истории и времени их отщепления от общего праязыка они не могут.

Заключение

В работе была сделана попытка ответить на вопрос, может ли грамматическая информация о языке служить источником данных для классификации языков. Материалом для исследования послужила 4-ая версия базы данных «Языки мира» ИЯз РАН. Данные о языках представлены в виде 124 признаков, охватывающих 14 разделов. В настоящий момент в базу данных включено свыше 280 языков.

Для проведения исследования было сделано две выборки. В первую вошел 21 язык из следующих 7 семей: тюркской, монгольской, уральской, нахско-дагестанской, манде, дравидийской и семитской, и 12 индоевропейских языков из 4 ветвей: германской, славянской, романской и анатолийской. Вторая выборка включила в себя 51 индоевропейский язык. Главной задачей работы было построить филогенетические деревья для выбранных языков и проверить, насколько качественно они отражают родство и порядок отщепления языков. Для всех языков в двух выборках было рассчитано расстояние Хэмминга и методом присоединения соседей построены два филогенетических дерева.

При анализе полученных деревьев было обнаружено следующее. Языки из первой выборки верно распределились по семьям. Расположение тюркских, уральских, монгольских языков и языков манде отражает общепринятую точку зрения о порядке расщепления этих языков внутри семей. Однако для нахско-дагестанских информация, представленная на дереве, противоречит современной классификации и мнению о последовательности расхождения языков.

Дерево, построенное для второй выборки языков, которая включает только представителей индоевропейской семьи, значительно уступает по качеству первому. Несмотря на некоторые соответствия известным фактам, можно сказать, что результат получился неудовлетворительным. Романские языки объединены в общую ветвь, однако родственные связи и порядок отщепления внутри этой ветви нарушены. Германские языки объединены и перемешаны со славянскими и балтийскими языками.

Полученные результаты позволили сделать вывод, что филогенетические деревья, для построения которых используются только грамматические данные, справляются только с задачей общей классификации языков по семьям, однако дать точную информацию об общем праязыке, эволюции и времени отщепления такие деревья не в состоянии.

ЛИТЕРАТУРА

1. Trubetzkoy N.S. Gedenken ihrer das Indogermanen Problem // Acta Linguistica. Copenhague. 1939, - 1. Pp. 81-89.
2. Benveniste E. La classification des langues // Conferences de l'Institut de Linguistique de l'Universite de Paris, XI, Annees 1952-1953; 1954: pp. 33-50.
3. Polyakov V., Solovyev, V., Wichmann, S., et al. Using WALS and Jazyki mira. Linguistic Typology, 2010 - 13(1), pp. 137-167.
4. Dryer Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2013. (Доступно по адресу: http://wals.info, дата обращения: 13.10.2021)
5. Anisimov I, Polyakov V., Solovyev V. 2013. Database "Languages of the World". New Version. New Research Horizons // Proceedings of the First International Forum on Cognitive Modeling 14-21 September, 2013, Italy, Milano Marittima". 2013: v.1, pp. 27-34.
6. Wichmann S., Holman, E. Pairwise comparisons of typological profiles. // In Rethinking Universals: How Rarities Affect Linguistic Theory (edited by Jan Wohlgemuth, Michael Cysouw). Walter de Gruyter Publ., 2010. Pp. 241-254/
7. Макарова Е.А. База данных «Языки мира» как инструмент классификации языков-изолятов // Мир науки, культуры, образования. 2020, - 6 (85). С. 524-528.
8. Nichols J., Warnow T.J. Tutorial on computational linguistic phylogeny. // Language and linguistics compass. 2008 - № 5 (2), pp. 760–820.
9. Edwards AWF, Cavalli-Sforza LL. Reconstruction of evolutionary trees. Heywood, Vernon Hilton; McNeill,J. Phenetic and Phylogenetic Classification. 1964 - pp. 67–76.
10. Gray R.D., Atkinson Q.D. Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin // Nature. 2003 - 426, 6965, pp. 435-439.
11. Ringe D., Warnow T., & Taylor, A. Indo-European and computational cladistics // Transactions of the Philological Society, 2002 - 100(1), pp. 59-129.
12. Hamming Richard W. Error detecting and error correcting codes // Bell System Technical Journal, 1950 - 29 (2): pp. 147–160.
13. Koichiro Tamura, Glen Stecher, Sudhir Kumar. MEGA11: Molecular Evolutionary Genetics Analysis Version 11 // Molecular Biology and Evolution, 2021 – v.38, Issue 7, pp. 3022–3027.
14. Saitou N., Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees // Molecular Biology and Evolution journal. — Oxford University Press, 1987. — Vol. 4, no. 4. — P. 406—425.
15. Дыбо А.В. Хронология тюркских языков и лингвистические контакты ранних тюрков. — М.: Академия, 2004. — С. 766
16. Напольских В.В. Очерки по этнической истории. – Казань: «Издательский дом «Казанская недвижимость», 2018. – 648 с.
17. Georg S., Michalove P., Ramer A., Sidwell P. Telling general linguists about Altaic // Journal of Linguistics, 1999 - 35 (1): pp. 65–98.
18. Лезгинские языки / Алексеев М.Е. // Лас-Тунас — Ломонос. — М. : Большая российская энциклопедия, 2010. — (Большая российская энциклопедия : [в 35 т.] / гл. ред. Ю.С. Осипов ; 2004—2017, т. 17).
19. Касьян А.С. , Якубович И.С. . Анатолийские языки // Языки мира. Реликтовые индоевропейские языки Передней и Центральной Азии. М. 2013, с. 15-25.

Журнал «Современная наука: актуальные проблемы теории и практики»