DOI: 10.30842/ielcp230690152469

Е. В. Коровина
(Институт языкознания РАН, Москва)

МЕТОД СКЛАДНОГО НОЖА: ОЦЕНКА УСТОЙЧИВОСТИ ЯЗЫКОВОЙ КЛАССИФИКАЦИИ

В данной статье рассматривается влияние на языковую классификацию полноты имеющегося в распоряжении материала, то есть вероятности изменения языкового дерева при добавлении/ исключении из рассмотрения тех или иных языков. Показывается, что все дистантные методы в этом отношении довольно сильно подвержены такого рода изменениям, при этом при сравнении результатов разных методов на материале разных языковых групп какой-то конкретный метод, дающий везде лучший результат выявить не удалось, однако метод программы Starling (который является комбинацией метода полной связи и метода средней связи) и метод полной связи во всех, взятых случаях показали не лучший результат.

Ключевые слова: сравнительно-историческое языкознание, статистика, лексикостатистика, классификация.

E. V. Korovina
(Institute of Linguistics, RAS, Moscow)

Jackknife resampling: some remarks about the stability of the language classification

This article discusses the impact on the language classification of the completeness of the available material, the probability of a change in the language tree when adding or excluding from the consideration of certain languages. As it is shown that all distant methods in this regard are quite susceptible to such changes, while comparing the results of different methods on the material of different language groups, it was not possible to identify a specific method that gives the best result everywhere, however, the method of the Starling program (which is a combination complete linkage method and pair-group method using arithmetic mean) and the complete linkage method in all examined cases showed not the best result.

Keywords: comparative linguistics, statistics, lexicostatistics, classification

Одним из известных феноменов классической лексикостатистики является то, что при увеличении количества идиомов, например, при включении неизвестных ранее диалектных данных, получаемое на выходе лексикостатистическое дерево может оказаться в той или иной степени перестроенным. Более того, из-за особенностей работы используемых алгоритмов в некоторых случаях дерево может быть перестроено даже при

изменении порядка следования языков в базе данных. В данном случае под классической лексикостатистикой понимается такая, которая основана на построении дерева на основе работы только с матрицей расстояний между идиомами без учета распределения значений признаков т. е. собственно строящая дерево при помощи дистантных методов (distance-based methods).

В качестве примера рассмотрим фрагмент дерева славянских языков построенного на основании списков из работы (Kushniarevich et al. 2015) при помощи стандартного метода программы Starling: сверху — исходное дерево, без удаления кашубского, снизу — дерево, полученное после данной трансформации (см. схему далее).

Видно, что после удаления кашубского языка к чехословацкой подгруппе оказывается ближе польский, а не лужицкие языки, как в исходном дереве. Исходная матрица приведена ниже:

	Polish	Kashubian	Czech	Slovak	Upper Sorbian	Lower Sorbian
Polish	0.0	**0.91**	0.86	0.85	0.84	0.86
Kashubian	99/109	0.0	0.81	0.81	0.81	0.82
Czech	95/110	88/109	0.0	**0.92**	0.88	0.87
Slovak	93/110	88/109	101/110	0.0	0.86	0.83
Upper Sorbian	92/109	88/108	96/109	94/109	0.0	**0.97**
Lower Sorbian	94/109	89/108	95/109	91/109	105/108	0.0

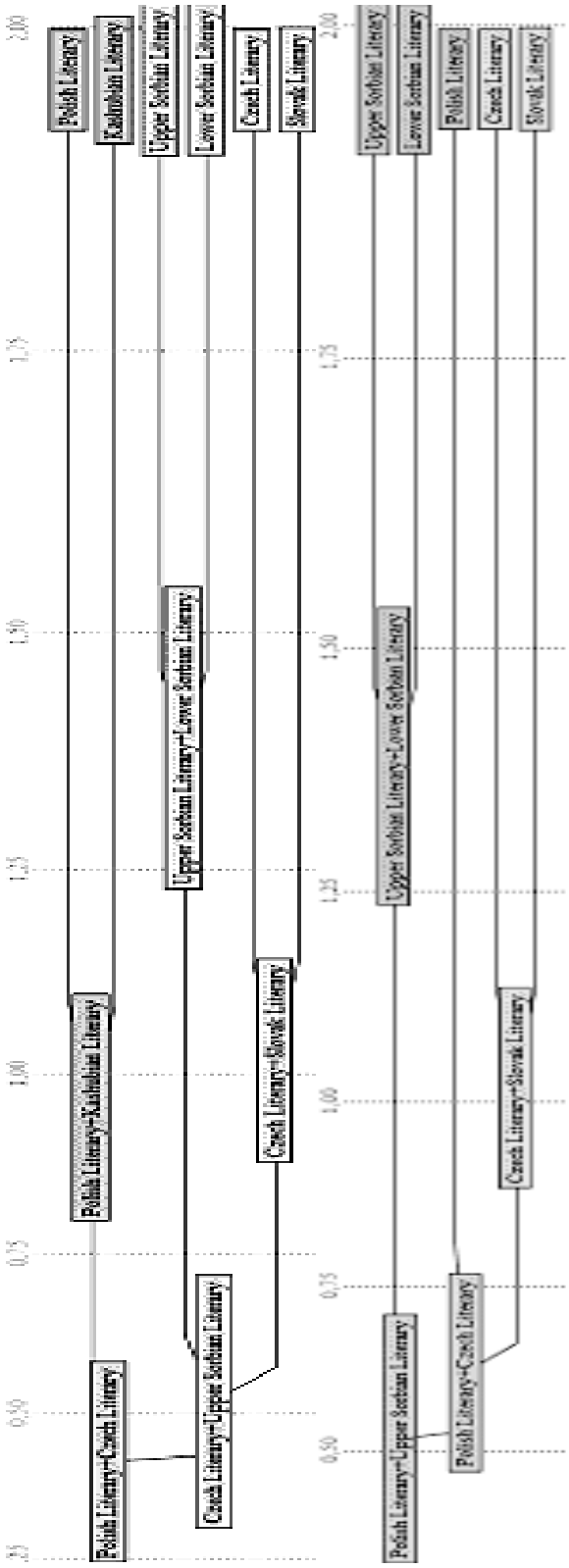
Таблица 1. Проценты сходства между западнославянскими языками. (Percent resemblance between West Slavic languages)

После объединения чешского и словацкого, а также лужицких языков она приобретает следующий вид:

	Kashubian	Czech + Slovak	Upper Sorbian + Lower Sorbian
Polish	**0.91**	0.85	0.84
Kashubian		0.81	0.81
Czech + Slovak			0.83

Таблица 2. Проценты сходства между 2 западнославянскими подгруппами и польским и кашубским (Percent resemblance between 2 West Slavic subgroups and Polish and Kashubian)

Фрагмент дерева славянских языков



При наличии кашубского польский объединяется с ним и объединенная группа получает по 81% сходства с чехословацкой и лужицкой и соответственно две последние на следующем шаге и должны объединиться между собой. В отсутствии кашубского же наибольший % (85%) оказывается у польско-чехословацкого.

Очевидно, что данный класс явлений безусловно известен исследователям, поскольку, в частности, позволяет до некоторой степени манипулировать полученным деревом. Например, в уже упомянутой работе (Kushniarevich et al. 2015), посвященной балто-славянской классификации для получения дерева более привычного вида из приводимого дерева был, в частности, убран словенский язык, хотя во вспомогательных материалах к статье его данные имеются. Однако теоретическое рассмотрение подобного рода явлений в лингвистической литературе, как представляется, практически отсутствует. Исключением в этом плане служит разве что работа (Rama, Wichmann 2018), где обсуждается наличие в австронезийской семье языков, которые согласно методологии, применяемой в проекте ASJP, не вписываются в традиционную классификацию, и формальные методы выявления таких языков.

Целью данной статьи является заполнение этой, как кажется, довольно существенной лакуны. Важность этого связана с тем, что для внешнего по отношению к специалистам по сравнительно-историческому языкознанию сообщества, языковые деревья, получаемые компаративистом, часто представляются некой стабильной данностью, хотя это и не соответствует действительности. В частности, как показано в Kassian 2015, разные методы, используемые при построении деревьев, дают разные результаты. Зависимость же результата классификации от набора взятых языков, особенно учитывая то, что набор идиомов, используемых для классификации далеко не всегда является полной и далеко не всегда эта недостача может быть заполнена (например, язык или диалект мог просто вымереть без минимальной документации), нуждается в отдельном осмыслении.

Для демонстрации этого явления прибегнем к методам математической статистики. Одним из стандартных статистических методов генерации повторной выборки (resampling) является наряду со статистическим бутстрепом (bootstrap) метод складного ножа (jackknife), предложенный М. Кенуем (Quenouille 1949) и расширенным Д. Тьюки (Tukey 1958). Этот

метод до некоторой степени обратен бутстрепу, при котором для моделирования новой, производной ситуации используются тем или иным способом выбранные значения данных при неизменных рядах данных, при использование же метода складного ножа изменяются наоборот ряды данных при том, что параметры этих рядов (значения признаков) остаются неизменными. Таким образом, применяя этот метод, можно определить, насколько стабилен исходный набор данных. Как представляется этот метод можно применить и к построению генеалогических деревьев, что в частности видно из славянского примера.

Можно предположить, что в части случаев перестроения обусловлены крайней близостью идиомов друг к другу и значительной неполнотой входных данных, однако перестроение дерева можно наблюдать и в более проработанных случаях на более удаленных друг от друга идиомах.

Возьмем еще несколько языковых групп и посмотрим, как часто встречается подобное перестроение деревьев при использовании различных дистантных методов классификации. Результаты тестирования для 6 групп представлены в таблице (значение — число языков, в той или иной мере перестаивающих дерево, в скобках, общее число ветвей, которые были перестроены):

	Starling	Single Linkage	Complete Linkage	WPGMA	UPGMA
Обско-угорская группа [1] (21)	7 (19)	7 (12)	7 (19)	10 (16)	**6 (12)**
Цезская (9)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Славянская (18)	2 (3)	6 (10)	4 (4)	**1 (1)**	3 (3)
Лезгинская (20)	1 (1)	3 (7)	1 (1)	**0 (0)**	1 (1)
Атапасскская (18)	9 (14)	**5 (8)**	6 (21)	6 (14)	8 (16)
Хмонг (17)	9 (10)	**2 (4)**	4 (12)	5 (6)	9 (10)

Таблица 3. Количество изменений в дереве после применения метода складного ножа
(The number of changes in the tree after applying the folding knife method)

Видно, что хотя число идиомов «удаление» которых меняет структуру дерева может быть довольно значительным, как

[1] Все лексические списки, использованные в данной работе за исключением славянских, взяты с сайта проекта GLD http://starling.rinet.ru/cgi-bin/main.cgi?root=new100&encoding=utf-eng; славянские списки взяты из публикации Kushniarevich et al. 2015.

например, в случае хмонг — в случае методов UPGMA и Starling это половина всех языков группы, влияние на дерево они оказывают довольно локальное, часто в пределах узкой ветви и не затрагивают верхних узлов результирующего дерева. В случае же славянских языков перестроение затрагивает лишь вышеупомянутые западнославянские языки при условии удаления лужицкого, а также диалектную структуру русского языка, в отношении остальных идиомов славянское дерево представляется стабильным.

Важно отметить, при сравнении результатов разных методов на материале разных языковых групп выделить какой-то конкретный метод, дающий везде лучший результат не удалось, однако метод программы Starling (который по сути является комбинацией метода полной связи и метода средней связи) и метод полной связи во всех, взятых случаях показали не лучший результат.

Литература

Kassian, A. 2015: Towards a Formal Genealogical Classification of the Lezgian Languages (North Caucasus): Testing Various Phylogenetic Methods on Lexical Data. *PLoS ONE*. doi:10.1371/journal.pone.0116950

Kushniarevich A, Utevska O, Chuhryaeva M, Agdzhoyan A, Dibirova K, Uktveryte I, et al. 2015: Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. *PLoS ONE* 10 (9): e0135820. doi:10.1371/journal.pone.0135820

Quenouille, M. H. 1949: Problems in Plane Sampling. *The Annals of Mathematical Statistics*. 20 (3), 355–375. doi:10.1214/aoms/1177729989

Wichmann, S., Rama, T. 2018: Jackknifing the Black Sheep: ASJP Classification Performance and Austronesian. *Senri Ethnological Studies* 98, 39–58.

Tukey, J. W. 1958: Bias and confidence in not quite large samples (abstract). *The Annals of Mathematical Statistics*. 29 (2), 614. doi:10.1214/aoms/1177706647.