М. Е. Васильев[†], *М. Н. Саенко*[‡]
[†] Институт языкознания РАН (Россия, Москва); mvhumanity@gmail.com
[‡] Институт славяноведения РАН (Россия, Москва); veraetatis@yandex.ru

Анализ топологии и оценка точности лексикостатистических классификаций (на примере славянских языков)*

Благодаря своей простоте и универсальности лексикостатистика остается одним из самых популярных методов для установления языкового родства и построения генеалогических классификаций. Среди российских компаративистов наибольшее распространение получило приложение Starling, использующее видоизменённую методику «присоединения соседей» при реконструкции филогенетических деревьев. Применение данной методики на материале разных языковых семей показывает хорошие или правдоподобные результаты в большинстве случаев. В то же время исследователи отмечают ряд недостатков в построенных классификациях, наиболее существенными из которых являются неустойчивость структуры древа даже к минимальным изменениям в составе идиомов, а также наличие в ней фиктивных таксонов и узлов, трудно объяснимых или противоречащих существующим представлениям.

В данной статье проводится детальное рассмотрение отмеченных проблем на примере лексической классификации 25 славянских идиомов. При этом показано, что главной причиной обоих явлений является несовершенство процедуры построения древа, используемой в Starling. По результатам исследования была предложена методика, позволяющая минимизировать влияние установленных недостатков путем выявления в топологии древа недостоверных узлов (на основе статистических расчетов) и их последующего исключения. Особенности предложенной методики делают ее применимой для анализа любых лексикостатистические классификаций, а также легко реализуемой в виде дополнительного компонента Starling или отдельного приложения.

Ключевые слова: лексикостатистика; метод присоединения соседей; генеалогическая классификация; среднее абсолютное отклонение.

I. Введение

Несмотря на то, что вопрос о надежности и эффективности лексического критерия при изучении языкового родства регулярно становится предметом обсуждения[1], лексикостатистические классификации по-прежнему остаются востребованными и одними из наиболее популярных среди исследователей. Главными причинами привлекательности лексикостатистики, очевидно, являются ее универсальность и относительная простота по сравнению с другими существующими методами. Так, например, если выявление общих инноваций, которое, как правило, считается более надежным классификационным критерием, требует диахронического исследования фонетических, морфологических и лек-

* Авторская работа М. Е. Васильева осуществлена при поддержке Российского научного фонда (проект № 20–18–00159); организация, осуществлявшая финансирование, — Институт языкознания Российской академии наук (ИЯз РАН). Авторская работа М. Н. Саенко выполнена в рамках работы по теме НИР «Славянская межкультурная и межъязыковая коммуникация в диахронии и синхронии».

[1] См., например, Бурлак, Старостин 2005: 148; Starostin 2010: 194; Поздняков 2014: 221–222; Грунтов, Мазо 2015: 211–216.

сических изменений в каждом из сравниваемых языков на протяжении всего рассматриваемого периода, то для лексикостатистического анализа достаточно сведений о составе базисного словаря этих языков всего на один произвольно выбранный момент времени, причем для каждого языка этот момент может быть разным. Это обстоятельство не только значительно упрощает получение генеалогической классификации, но также позволяет использовать лексикостатистику для установления родственных связей между малоизученными языками, в случае с которыми применение других методов часто оказывается затрудненным из-за недостатка данных.

Еще одним немаловажным фактором является формализованный характер лексикостатистических расчетов, благодаря которому процедуру построения древа можно выполнить автоматически с помощью компьютерной программы. Такая возможность была реализована в приложении Starling[2], разработанном С. А. Старостиным в 1985–2000 гг. и получившим распространение как среди российских, так и зарубежных компаративистов[3]. Для определения степени родства между языками программа рассчитывает процент этимологически совпадающих лексем в их 100-словных (или 110-словных) списках и по итогам расчетов формирует таблицу с долями совпадений между всеми идиомами попарно. Непосредственно построение генеалогического древа осуществляется на основе полученной таблицы с помощью методики[4], представляющей собой несколько видоизмененный и адаптированный для лингвистического материала метод «ближайших соседей»[5], широко применяемый в биологии.

За последнее десятилетие представителями Московской школы компаративистики был накоплен значительный опыт применения лексикостатистического метода для построения генеалогических классификаций различных языков мира, включающих как близкородственные малые группы, так и крупные языковые общности с большой временной глубиной. При этом построенные деревья повсеместно используются в ходе исследовательской работы, регулярно приводятся в научных публикациях, а также демонстрируются во время конференций.

Анализируя полученные результаты[6], исследователи, как правило, подчеркивают полезность классификации в целом и акцентируют внимание на ее особенностях, значимых для целей исследования, но в то же время указывают на отдельные несоответствия в структуре древа, плохо поддающиеся объяснению или противоречащие известным данным. Среди таких странностей особо следует выделить два наиболее характерных недостатка, которые проявляются независимо от выбора рассматриваемых языков и, как можно предположить, обусловлены самой методикой формирования деревьев:

---

[2] StarLing for Windows, v. 2.6.10: computerized system for multilingual database processing, (c) 1985–2005 by S. A. Starostin, StarLing Software Inc. Текущая версия программы доступна на сайте проекта «Вавилонская башня» по адресу: https://starling.rinet.ru/downl.php?lan=ru#soft.

[3] В отличие от многочисленных существующих программ для построения филогенетических деревьев, предназначенных в первую очередь для классификации биологических видов, Starling изначально создавался как специализированное приложение для сбора, обработки и анализа именно лексических данных, благодаря чему и завоевал свою популярность у лингвистов.

[4] Суть данной методики подробно разбирается в учебнике (Бурлак, Старостин 2005: 163–167). Некоторые из ее особенностей мы более подробно рассмотрим далее.

[5] Метод «ближайших соседей» или «присоединения соседей» (Neighbor-Joining Method) — алгоритм построения филогенетических деревьев, в основу которого положен принцип последовательного попарного объединения «ближайших» (т.е. имеющих наибольшее сходство) таксонов. Первоначально метод предназначался для классификации нуклеотидных последовательностей (см. Saitou, Nei 1987), однако в дальнейшем стал широко применяться также за пределами генетики.

[6] См., например, обсуждение полученных классификаций в работах Kogan 2016: 235–238; Vydrin 2009: 112–114.

1. Конфигурация древа является неустойчивой и крайне чувствительна к изменению количества или состава идиомов. В частности, нередки случаи, когда исключение или добавление одного языка приводит к абсолютно неожиданным и радикальным изменениям в топологии, затрагивающим не только таксон с новым или исключенным элементом, но также ветви, максимально от него удаленные.
2. Древо содержит большое количество незначимых узлов, интерпретация которых крайне проблематична или невозможна на основании имеющихся сведений об истории развития языков и их взаимной дивергенции. Как правило, такие узлы располагаются в непосредственной близости или на незначительном временном расстоянии от других узлов, а иногда образуют непрерывные цепочки в виде характерной ступенчатой структуры.

Несмотря на то, что обе указанные особенности очень распространены (и хорошо знакомы всем пользователям Starling), они крайне редко удостаиваются отдельного обсуждения: в большинстве случаев авторы ограничиваются констатацией несовершенства методики, с которым неизбежно приходится мириться. При этом большинство исследователей признают, что наличие подобных «артефактов» существенно снижает практическую ценность построенных деревьев, а также ставит под сомнение их достоверность. Таким образом, мы сталкиваемся с необходимостью анализа выявленных методических погрешностей, а также поиска способов их оценки и минимизации.

В рамках нашей статьи мы рассмотрим возможный подход к решению данной задачи на примере лексикостатистической классификации 25 славянских языков и диалектов, уделяя особое внимание вышеупомянутым проблемам вариативности и незначимой кластеризации в строении деревьев.

II. Исходные данные

Дадим краткое описание идиомов, списки базисной лексики которых задействованы в исследовании[7]. В соответствии с принципами проекта «Глобальная лексикостатистическая база данных», в рамках которого были собраны списки, предпочтение отдавалась диалектным данным, поскольку базисная лексика литературных языков, предположительно, отличается бóльшим консерватизмом.

1. *Банатский болгарский*. Переселенческий говор, на котором говорят болгары-католики в румынском и сербском Банате. Переселение состоялось в двух волнах. Сначала в 1688 г. после неудачного восстания в Валихию сбежали жители города Чипровци и окрестностей. В 1720-е гг. к ним присоединились так называемые павликиане из-под Свиштова и Николпола. Обе группы встретились и смешались в Банате, где восточноболгарский говор более многочисленных свиштовцев и николполцев почти полностью вытеснил западноболгарский говор чипровцев (Стойков 2002: 193). Словарь составлен выдающимся болгарским диалектологом С. Стойковым на материале, собранном начиная с 1953 г., преимущественно в румынских селах Стар-Бешенов (рум. Dudeştii Vechi) и Винга (рум. Vinga) (Стойков 1968).

2. *Македонский д. Горно-Каленик*. Говор деревни Горно-Каленик, которая находится в Греции (греч. Ἀνω Καλλινίκη), недалеко от города Лерин (греч. Φλώρινα). Материал был собран П. Хиллом преимущественно в Австралии у македонцев, сбежавших из Греции

[7] Большая часть списков с описанием и аннотацией доступна по ссылке https://starling.rinet.ru/cgi-bin/response.cgi?root=new100&basename=new100\ier\slv.

Рисунок 1. География идиомов, списки которых используются в исследовании

во время гражданской войны (Hill 1991). Относится к леринскому говору юго-западного диалекта македонского языка.

3. *Штокавский сербохорватский племени Загарач*. Говор черногорского племени Загарач (местное произношение – *Загàрач*), населяющего территорию вокруг горы Гарач (Ћупић, Ћупић 1997). Относится к зетско-южносанджакскому диалекту штокавского наречия, согласно классификации П. Ивича.

4. *Чакавский сербохорватский о. Врґада*. Говор острова Врґáда. Словарь составлен хорватским лингвистом, носителем говора, Б. Юришичем на основе записей 1908–1960 гг. (Jurišić 1973). По классификации Брозовича и Ивича, говор относится к южночакавскому диалекту.

5. *Чакавский сербохорватский д. Орлец*. Говор деревни О́рлец (местное произношение – *Õrlec*), расположенной на острове Црес. Словарь составлен Х.П. Хоутзагерсом на основе полевых записей 1980–1982 гг. (Houtzagers 1985). По классификации Брозовича и Ивича, говор относится к северночакавскому диалекту.

6. *Чакавский сербохорватский д. Орбаничи*. Говор деревни Орбáничи (местное произношение — *Orbânići*), находящейся в двух километрах от города Жминь в центральной Истрии. Словарь составлен нидерландской исследовательницей Я. Калсбек на основе материала, собранного в 1980–1984 гг. (Kalsbeek 1998). Согласно классификации Брозовича и Ивича, говор относится к юго-западному истрскому диалекту чакавского наречия.

7. *Чакавский сербохорватский д. Девинска-Нова-Вес*. Переселенческий говор деревни Девинска-Нова-Вес в Словакии (по-словацки *Devínska Nová Ves*, в говоре — *Nụ̑ovo sèlo*; в настоящее время — район Братиславы). Носители говора — градищанские хорваты, поселившиеся на этой территории в XVI в. Словарь составлен чешским исследователем В. Важным на основе полевых записей 1923–1926 гг. Помимо основного материала из деревни Девинска-Нова-Вес часть была записана в соседних деревнях Дубравка (*Dúbravka*; *Dubráva*) и Ламач (*Lamač*; *Làmụ̑oč*) (Vážný 1927).

8. *Градищанский кайкавский сербохорватский*. Переселенческий говор, на котором говорят в двух деревнях в Венгрии — Хидегшег (венг. *Hidegség*, произношение в говоре — *Hęḍęšîn* / *Hęḍęšîn*) и Фертёхомок (*Fertőhomok*; *Hòmok*). Предки носителей говора переселились в начале XVI века из Славонии, предположительно из населенных пунктов Кралева-Велика (*Kraljeva Velika*) и Меджурич (*Međurić*), которые находятся значительно восточнее современной границы кайкавского наречия. Словарь составлен Х. П. Хоутзагерсом на основе полевых записей 1985–1994 гг. (Houtzagers 1999).

9. *Чабарский словенский*. Говор окрестностей города Ча́бар (схр. *Čàbar*) в Хорватии. Словарь составлен С. Малнаром (Malnar 2008). Говор относится к костельскому диалекту доленьского наречия словенского языка.

10. *Костельский словенский*. Говор деревни Де́лач (произношение в говоре — *ˈDėlaːč*) и окрестностей составляет южную часть костельского диалекта доленьского наречия. Словарь был составлен Й. Грегоричем, уроженцем Делача (Gregorič 2014).

11. *Чрновршский словенский*. Диалект плато Чрни-Врх (слвн. *Črni Vrh*) относится ровтарскому наречию. Словарь был составлен И. Томинцем, носителем диалекта, преимущественно на основе говора его родной деревни Ломе (слвн. *Lome*) (Tominec 1964).

12. *Словенский д. Затолмин*. Говор деревни Затолми́н (слвн. лит. *Zatolmin*, в говоре — *Zatˈmin*), лежащей в 1 км от города Толмин в западной Словении, недалеко от границы с Италией. Материал собирался носительницей говора Х. Чуец-Стрес свыше десяти лет, начиная с 1996 г. (Čujec Stres 2011, 2014). Говор относится к толминскому диалекту ровтарского наречия.

13. *Словенский д. Била*. Говор деревни Би́ла (ит. *San Giorgio*, в говоре — *Bíla*) в Резьянской долине в Италии. Материал записан в 1987–1991 гг. Х. Стенвейком (Steenwijk 1992). Говор относится к резьянскому диалекту приморского наречия.

14. *Словенский д. Брдо*. Говор Зильской долины в Австрии. Материал был собран Т. Пронком в 2001–2006 гг. преимущественно у одной информантки, которая родилась и выросла в деревне Эгг (нем. *Egg bei Hermagor*, в говоре — *Bŕdo*), а после замужества проживала в Почахе (нем. *Potschach*, в говоре — *Pətǫ́čani*) (Pronk 2009). Говор относится к зильскому диалекту каринтийского наречия.

15. *Прлекийский словенский*. Говор деревень Бре́нгова (слвн. лит. *Brengova*, в говоре — *Bˈŕę̣ŋgọva*) и Це́нкова (слвн. лит. *Cenkova*, в говоре — *ˈCę̣ŋkọva*), входящий в северо-западную часть прлекийского диалекта паннонского наречия. Словарь составлен Б. Райхом (Rajh 2010).

16. *Подкрконошский чешский*. Диалект чешского Подкрконошья (чеш. *Podkrkonoší*), территории к югу от Крконошских гор (Bachmannová 2016).

17. *Моравский чешский д. Мистршице*. Говор деревни Мистршице (чеш. *Mistřice*), находящейся в 7 км от города Угерске-Градиште в Моравии. В словарь, составленный И. Малиной, включены также некоторые лексемы, записанные в соседних населенных пунктах (Malina 1946).

18. *Словацкий деревни Пилишсанто*. Переселенческий говор деревни Пилишсанто (венг. *Pilisszántó*), расположенной недалеко от Будапешта. Словаки поселились там предположительно в начале XVIII века. Большая часть пришла с территории Малых Карпат и говорила на западнословацком диалекте. Материал был собран Ф. Грегором в 1950-е гг. (Gregor 1975).

19. *Малопольский д. Венцюрка*. Говор деревни Венцюрка (пол. *Więciórka*), расположенной в мысленицком повяте Малопольского воеводства. Словарь, составленный уроженцем Венцюрки М. Куцалой, включает лексику, собранную в трех деревнях: Венцюрка (основной материал словаря), Сидзина-Гурна (*Sidzina Górna*) и Фацимех (*Facimiech*) (Kucała 1957).

20. *Коцевский великопольский*. Говоры региона Коцеве (*Kociewie*), относящиеся к великопольскому диалекту. Материал был записан Б. Сыхтой в 1930–1970-е гг. (Sychta 1–3).

21. *Белорусские говоры Гродненской области*. Говоры Гродненской области Белоруссии, входящей в ареалы юго-западного и центральнобелорусского диалектов. Материал собран Т. Ф. Стешкович в 1948–1960 гг. (Сцяшковіч 1972; Сцяшковіч 1983).

22. *Белорусские говоры Турова и окрестностей*. Говоры города Турова и 33 деревень в его окрестностях. Относятся к юго-западному диалекту белорусского языка. Материал записан коллективом исследователей в экспедициях 1967–1981 гг. (ТС 1–5).

23. *Украинский д. Торунь*. Говор села Торунь (местное произношение — *Tǫ́run*) в Закарпатской области. Относится к восточнобойковской группе юго-западных украинских говоров. Материал собран экспедицией под руководством С. *Л*. Николаева в 1990 г. (Николаев, Толстая 2001).

24. *Русский д. Деулино*. Говор деревни *Д*еулино (Рязанская область), относящийся к рязанской группе говоров южнорусского наречия. Материал записан в 1960–1963 гг. (ССРНГ 1969).

25. *Русский д. Островцы*. Говор деревни Островцы (Псковская область), относящийся к гдовской группе говоров среднерусского наречия. Материал собран в 1995–1998 гг. З. Хонселааром и опубликован в виде монографии, включающей словарь (Хонселаар 2001).

Как мы видим, имеющийся материал несколько неоднороден: какие-то словари предоставляют в наше распоряжение материал говора лишь одного населенного пункта, какие-то — нескольких, какие-то — целого большого региона. Некоторые словари являются дифференциальными, то есть дают лишь ту лексику, которая отличается от лексики литературного языка (и тогда собирать списки базисной лексики приходится по большей части из примеров, имеющихся в словаре), другие же — недифференциальными, то есть описывают словарный состав говора во всей его полноте. Кроме того, данные были записаны были в разное время и исследователями с отличающимися подходами. Неоднородность исходных данных значительно усложняет задачу для исследователя, желающего построить лексикостатистическое древо, и оказывает непосредственное влияние на качество полученной в итоге классификации.

Определенные искажения в структуру древа вносят случаи заимствований (в базе данных Starling им присваивает значение «-1») и синонимов или супплетивизма, когда одной строке в базе соответствует два или более корней. Также, к сожалению, не для всех списков удалось собрать полные 110-словные списки, иногда искомый материал отсутствует в словаре. В таблице 1 мы приводим краткие сведения о подобных изъянах в материале.

Оговорим сразу, что на основании собранных нами 25 списков нельзя построить репрезентативную классификацию славянских языков, поскольку они покрывают славянский мир неравномерно, и для создания качественного лексикостатистического древа требуется значительно больший объем данных. Однако в рамках данной работы мы и не ставим перед собой такой задачи. В нашем случае мы планируем использовать имеющийся славянский материал для проверки и уточнения некоторых аспектов современной лексикостатистической теории.

III. Анализ лексикостатистической классификации

Для расчета долей совпадений между 110-словными списками славянских идиомов и построения их генеалогической классификации использовалось приложение Starling («стандартный» метод). В результате проведённых вычислений была получена исходная лексикостатистическая матрица — Таблица 3 (см. Приложение), а также генетическое древо, представленное на рис. 2 ниже. Рассмотрим его более подробно.

№	Название	Лакуны	Заимствования	Синонимы или супплетивизм
1	Банатский	0	2 (belly, rain)	4 (bird, I, smoke, go)
2	Горно-Каленик	3 (bark, fat, swim)	1 (liver)	4 (come, I, person, say)
3	Загарач	1 (cloud)	0	7 (ashes, dog, I, leaf, person, say, we)
4	Вргада	0	2 (liver, sand)	4 (belly, I, person, we)
5	Орлец	0	1 (round)	3 (I, person, we)
6	Орбаничи	0	3 (dog, man, sand)	6 (fat, I, many, worm, person, we)
7	Девинска-Нова-Вес	0	3 (road, tree, snake)	3 (I, person, we)
8	Градищанский кайкавский	2 (moon, snake)	0	5 (belly, I, person, say, we)
9	Чабарский	0	0	4 (burn, person, road, we)
10	Костельский		1 (belly)	4 (I, louse, person, we)
11	Чрновршский	0	0	4 (I, many, person, we)
12	Затолмин	1 (worm)	0	3 (I, person, we)
13	Била	1 (warm)	5 (bark, fat, feather, mouth, tree)	4 (I, kill, person, we)
14	Брдо	0	2 (fly, neck)	3 (I, person, we)
15	Прлекийский	2 (bark, round)	1 (belly)	3 (I, person, we)
16	Подкрконошский	0	0	4 (I, many, person, we)
17	Мистршице	4 (cloud, fat, mountain, sand)	0	3 (I, person, we)
18	Пилишсанто	2 (bark, root)	1 (sand)	4 (cloud, I, person, we)
19	Венцюрка	0	5 (bark, feather, heart, red, skin)	4 (big, I, person, we)
20	Коцевский	1 (yellow)	2 (heart, red)	4 (I, many, person, we)
21	Гродненский	0	6 (heart, red, see, seed, skin, worm)	3 (hair, person, we)
22	Туровский	0	3 (dog, red, see)	5 (cloud, I, liver, person, we)
23	Торунь	1 (tooth)	3 (one, seed, short)	3 (I, many, we)
24	Деулинский	0	2 (cloud, dog)	3 (I, person, we)
25	Островцы	1 (horn)	3 (cloud, say, what)	5 (ashes, I, long, person, we)

Таблица 1. Лакуны, заимствования и синонимы в используемых списках

В целом полученную классификацию можно охарактеризовать как удовлетворительную. На древе отчетливо выделяются болгаро-македонский, восточно-славянский, словенско-сербохорватский и западнославянский таксоны. Отсутствие объединения болгаро-македонского и сербохорватско-словенского таксонов в южнославянскую подгруппу само по себе не является критическим: ряд исследователей не поддерживает выделение такой подгруппы в составе славянских языков (см. обзор в Blažek 2017). Намного важнее то, что на самом нижнем уровне древо выглядит неверно: болгаро-македонский таксон объединен с восточнославянским, а словенско-сербохорватский — с западнославянским. Сербохорватские и словенские идиомы, представленные наибольшим числом списков, в рамках своих таксонов ведут себя не вполне корректно: словенские говоры выстроились «лесенкой» без какого-либо выраженного диалектного деления. Из шести сербохорватских списков четыре являются чакавскими, однако они не выделились в особую подгруппу, а объединились попарно и разбились штокавским и кайкавским списками.

Таким образом, положительно можно оценить «среднюю» часть древа (то есть объединение идиомов в четыре подгруппы), и отрицательно – «нижнюю» и «верхнюю» части, в которых мы наблюдаем фантомные корневые узлы и ступенчатое членение словенских и сербохорватских говоров, не соответствующее лингвистической действительности.

Перейдем теперь к анализу внутренних свойств классификации и выясним, насколько ее структура зависит от изменений в составе рассматриваемых языков. Для этого сформируем из них 25 дополнительных выборок, поочередно исключая из полного списка по одному идиому, а затем сравним деревья, построенные для каждого нового набора, с исходным. В результате сопоставления[8] были найдены три идиома, исключение которых из классификации привело к значимым изменениям в топологии древа:

1. Македонский говор д. Горно-Каленик;
2. Чакавский сербохорватский говор д. Орлец;
3. Градищанский кайкавский сербохорватский.

Начнем наше рассмотрение с первого случая. На рис. 3 приведено генеалогическое древо 24-х славянских языков с исключенным македонским говором. Сравнивая его с исходной классификацией (рис. 2), мы обнаруживаем, что все основные таксоны, соответствующие восточнославянской, западнославянской, сербохорватско-словенской подгруппам, полностью сохранили свою целостность и внутреннее строение. В то же время сокращение выборки привело к неожиданным и весьма существенным изменениям в корневой части древа: после исключения македонского восточнославянские идиомы образовали единую общность с сербохорватскими, словенскими и западнославянскими языками, тогда как банатский болгарский оказался обособленным от всех остальных идиомов[9].

Последнее отличие выглядит особенно проблематично, так как раннее отделение болгарского от основного массива славянских говоров не подтверждается никакими лингвистическими данными.

На первый взгляд, примеры такой вариативности в конфигурации древа при минимальных изменениях в составе языков заставляют усомниться в практической ценности лексикостатистических классификаций и перспективах использования методики в целом. Однако, прежде чем делать подобные неутешительные выводы, следует принять во внимание, что доли совпадений, определяющие последовательность объединения таксонов и взаимное расположение узлов древа, в действительности являются не детерминированными[10], а *статистическими* величинами[11], которые обусловлены случайным характером процесса лексических замен и могут отклоняться от расчетных значений в большую или меньшую сторону. Это означает, что проценты совпадений, соответствующие узлам древа, имеют некоторый разброс — погрешность, которую необходимо учитывать как при построении, так и последующем анализе найденной топологии. Для количественной оценки данной погрешности мы воспользуемся величиной

[8] Все полученные деревья, а также исходные лексикостатистические данные представлены в сопровождающих материалах, которые доступны онлайн на сайте ВЯР.

[9] При этом разница между первым (корневым) и вторым узлами древа, соответствующим отделению болгарского и началу разделения остальных групп, достигает 2% (что эквивалентно разнице в 2 слова при сравнении 100-словных списков).

[10] То есть точно заданными.

[11] К сожалению, это важное обстоятельство, указание на которое содержится в самом названии лексикостатистики, в большинстве случаев попросту игнорируется при анализе лексикостатистических расчетов, что неизбежно приводит к абсурдным результатам и в конечном итоге дискредитирует весь метод в целом.

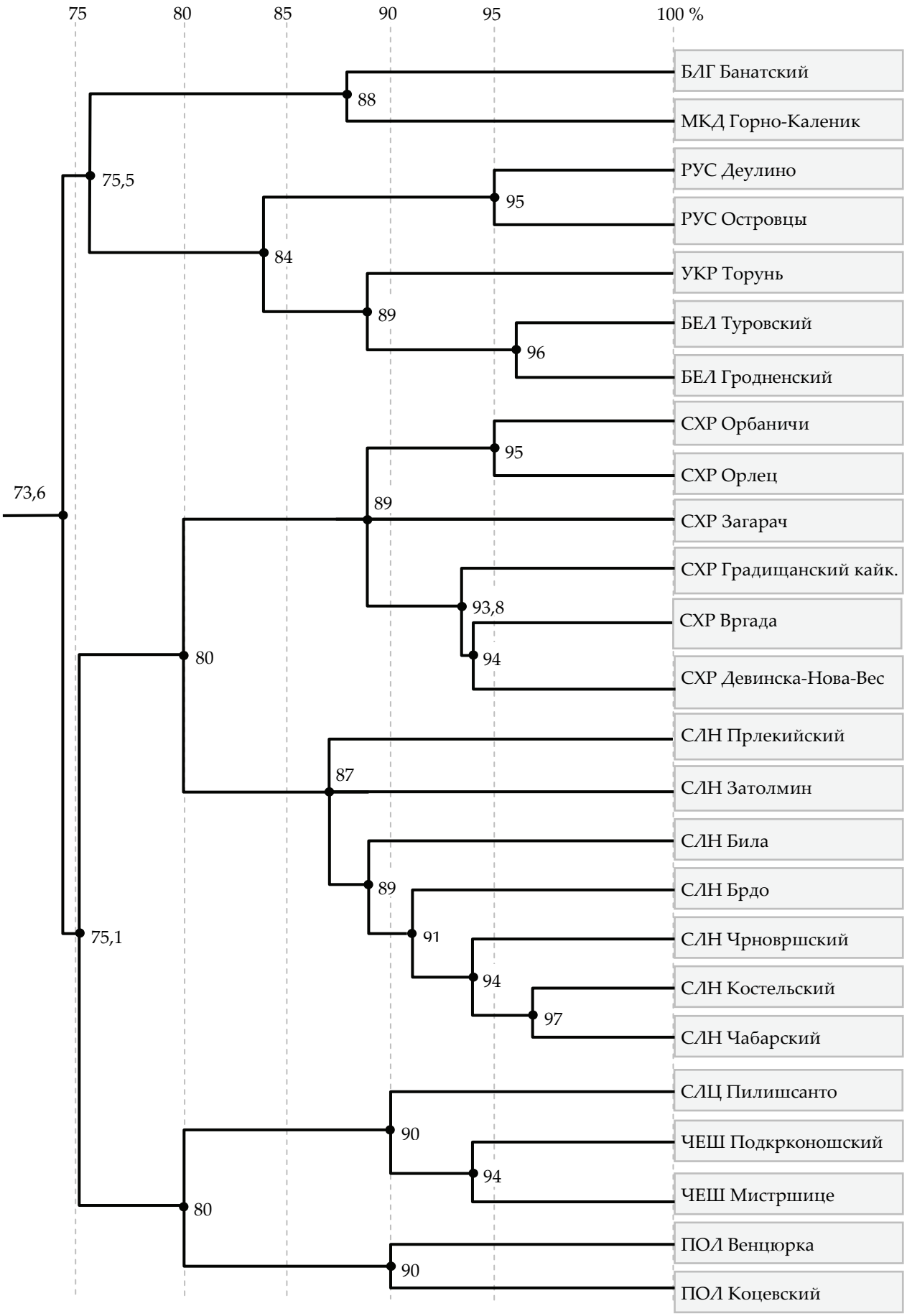


Рисунок 2. Лексикостатистическая классификация 25 славянских языков, полученная с помощью Starling (метод «Standard»). Значения на шкале и рядом с узлами древа соответствуют процентам совпадений.

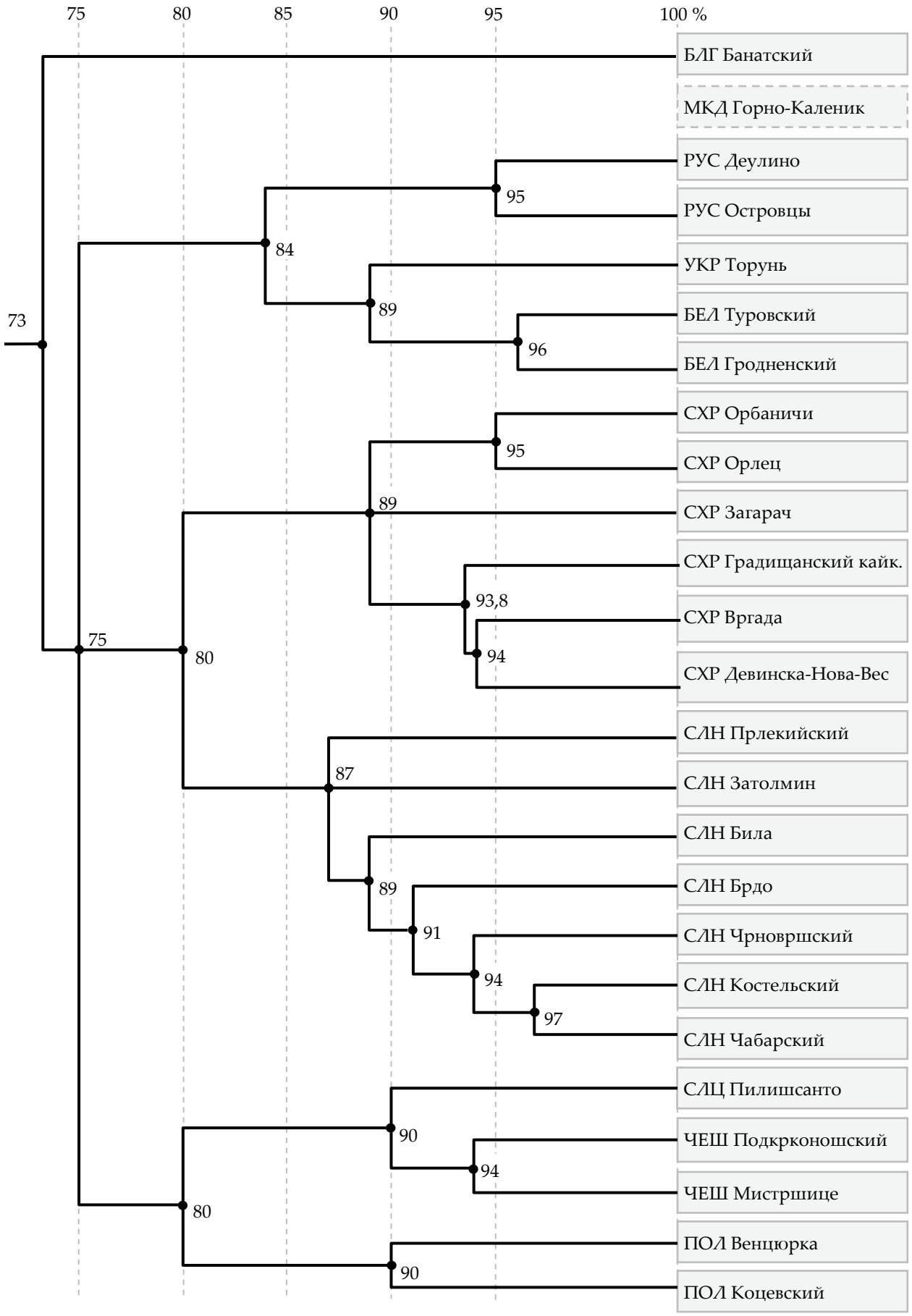


Рисунок 3. Генеалогическое древо 24 славянских идиомов, после исключения македонского. Значения на шкале и рядом с узлами древа соответствуют процентам совпадений.

среднего абсолютного отклонения[12], общий смысл которой поясним на следующем примере (см. табл. 2 и рис. 4):

Языки	A	B	C
A	-	90	84
B	90	-	86
C	84	86	-

Таблица 2. Проценты совпадений между списками языков A, B и C

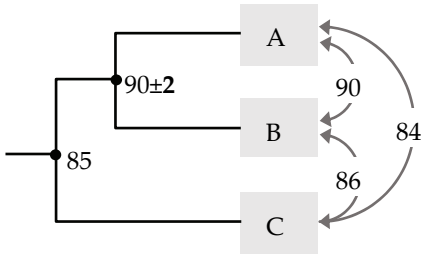


Рисунок 4. Генеалогическое древо языков A, B и C, полученное на основе таблицы

Согласно исходным данным (табл. 2), языки A и B являются ближайшими родственниками и образуют первый узел древа с процентом совпадений N_{AB}=90, после чего к ним остается присоединить оставшийся идиом C (рис. 4). При этом количество совпадений между языками A и C и между языками B и C отличается на два слова (N_{AC}=84; N_{BC}=86). Если мы исключим возможность заимствований и повторных сближений, то в рамках классической модели дивергенции такое расхождение можно объяснить только неравномерностью процесса замен в базисной лексике идиомов A и B, а именно: либо ускоренным лексическим изменением языка A, либо, наоборот, замедленным изменением языка B. Следовательно, в первом случае (с учетом двух «лишних» замен), количество общих слов в списках языков A и B составит N_{AB}=90+2=92, а во втором N_{AB}=90–2=88. Таким образом расчетная доля совпадений для узла A — B может лежать в диапазоне от 88% до 92%, что численно соответствует величине абсолютного отклонения E_{AB}, которая, для данного примера[13], рассчитывается по формуле:

$$E_{AB} = | N_{AC} - N_{BC} | = | 84 - 86 | = 2$$

Перейдем теперь к анализу полученных ранее классификаций славянских языков на основе средних абсолютных отклонений, рассчитанных для каждого узла (рис. 5 и рис. 6). Нетрудно заметить, что в строении обоих деревьев присутствует большое количество узлов, абсолютные отклонения которых накладываются друг на друга. Более того, среди них можно выделить несколько групп, в которых диапазоны отклонений включают в себя сами расчетные значения, полученные для соседних узлов. Особенно показательным примером является последовательное объединение идиомов Брда, Билы, Затолмина и Прлекии внутри словенской подгруппы, образовавших непрерывную цепочку из трех узлов с взаимным перекрытием. Причем в случае с Билой абсолютное отклонение составляет 2,6% и охватывает сразу два соседних узла.

[12] Методика расчета среднего абсолютного отклонения подробно излагается в работах Васильев 2010: 538–540; Васильев, Коган 2013: 160.

[13] В общем случае (для группы с произвольным количеством идиомов) используется более сложная формула — см. Васильев 2010: 540.

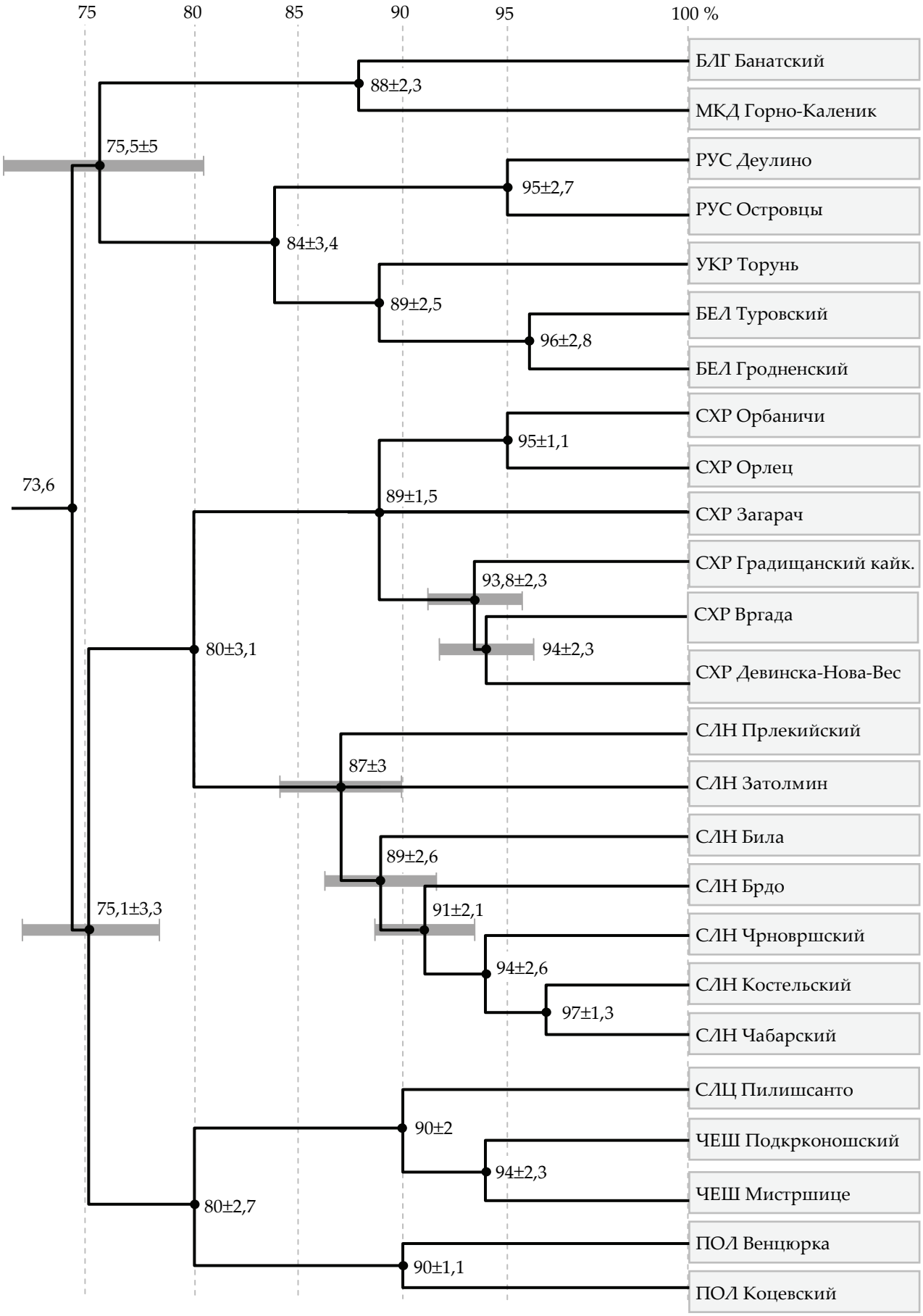


Рисунок 5. Генеалогическое древо 25 славянских идиомов с указанными процентами совпадений и значениями средних абсолютных отклонений. Диапазоны отклонений, перекрывающие соседние узлы, показаны отрезками.

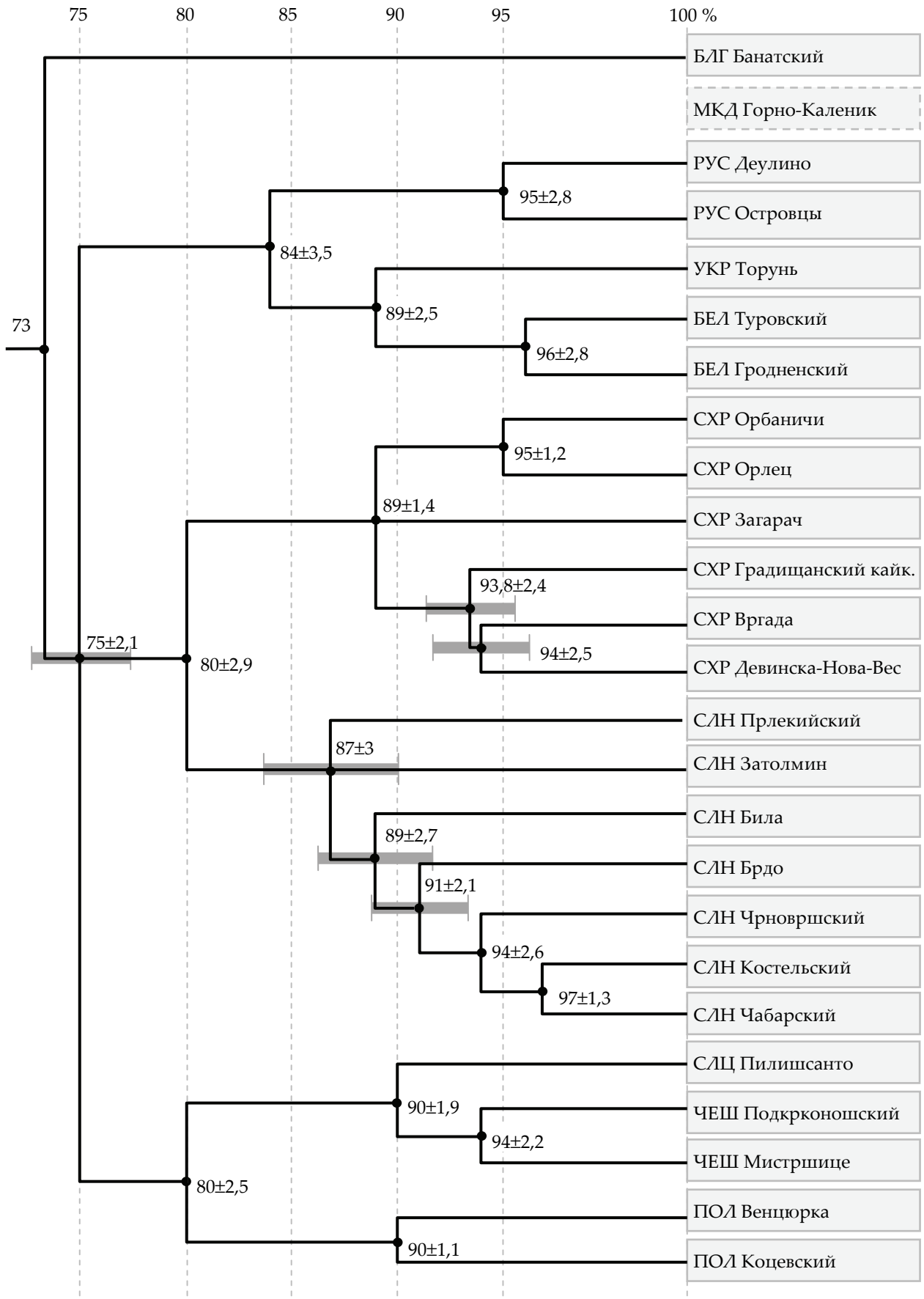


Рисунок 6. Генеалогическое древо 24 славянских языков после исключения македонского. Рядом с узлами указаны значения процентов совпадений и средних абсолютных отклонений. Диапазоны отклонений, перекрывающие соседние узлы, показаны отрезками.

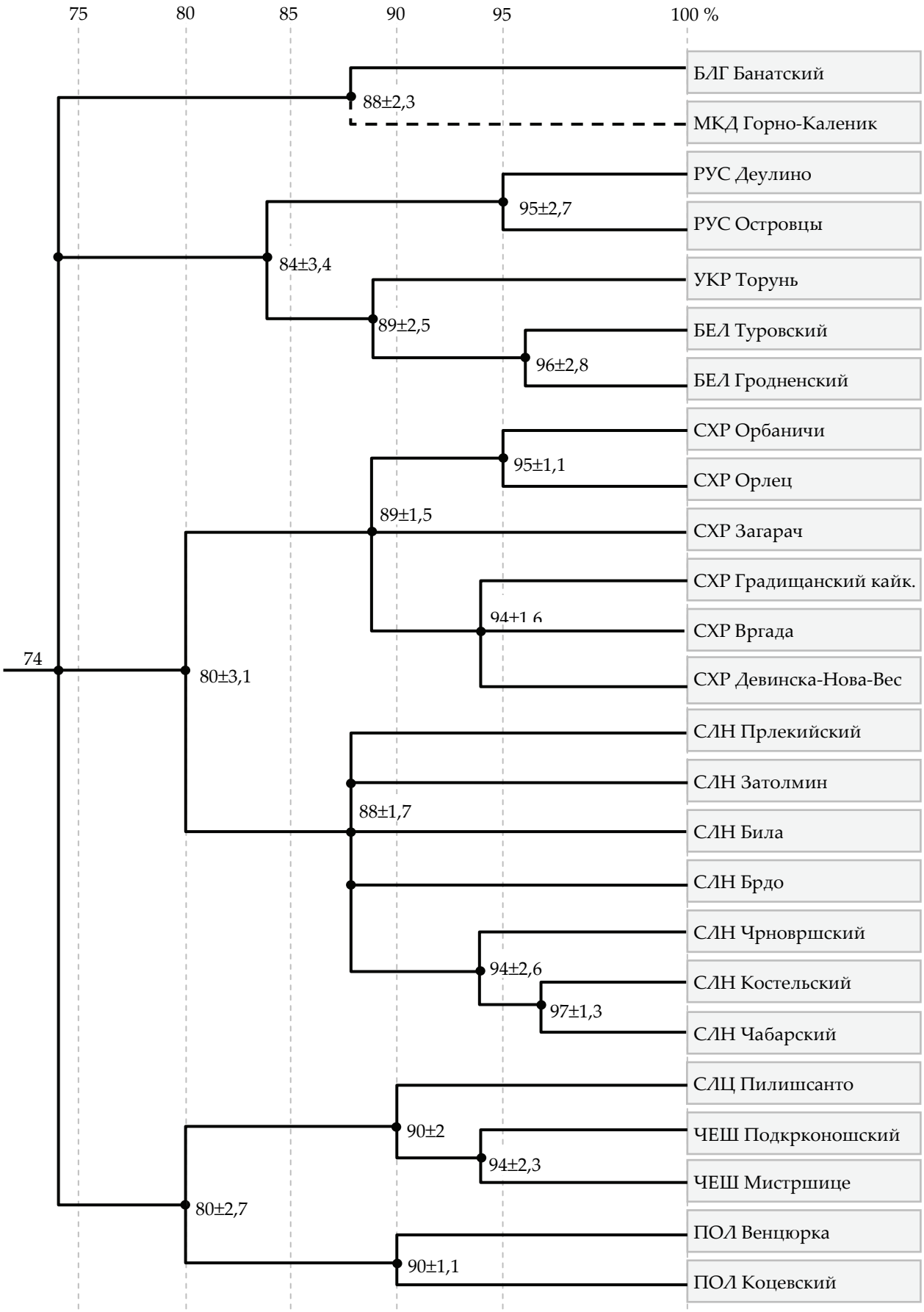


Рисунок 7. Генеалогическое древо славянских языков после объединения узлов с взаимно перекрывающимися диапазонами отклонений. Рядом с узлами приведены значения процентов совпадений и средних абсолютных отклонений.

Похожую ситуацию мы наблюдаем в основании деревьев. На рис. 5 корневой узел (со значением 73,6) перекрывается отклонениями обоих последующих узлов, в один из которых входят восточнославянский и болгаро-македонский таксоны, а в другой — западнославянский и сербохорватско-словенский таксоны. Аналогично на рис. 6 узел, соответствующий отделению болгарского, лежит в пределах отклонений узла, объединяющего все остальные ветви древа[14].

В этом и аналогичных примерах разница между процентами совпадений близкорасположенных узлов лежит в пределах погрешности, что свидетельствует об их *статистической неразличимости* на основе имеющихся данных. Следовательно, такие узлы можно[15] рассматривать как некий континуум и заменить их одним или несколькими более крупными узлами, после чего рассчитать для них новые значения долей совпадений, а также средние абсолютные отклонения. При этом, если узлы, полученные в результате объединения, также окажутся взаимно перекрывающимися, то процедуру можно повторять до тех пор, пока расстояние между любыми соседними узлами не будет превышать величину отклонений[16].

Применяя описанную методику к рассматриваемым классификациям, мы обнаружим, что после объединения узлов с перекрывающимися средними отклонениями и пересчета соответствующих значений оба древа оказались идентичными и приобрели вид, показанный на рис. 7. Очевидным образом это устраняет вариативность, вызванную изъятием македонского говора, поскольку конфигурация древа теперь остается неизменной вне зависимости от того, какую выборку (полную или сокращенную) мы используем. Кроме того, таксономическое «прореживание» позволило сократить количество фиктивных узлов[17] в группе словенских говоров, а также других ветвях древа и тем самым значительно смягчить проблему «топологического шума», обозначенную ранее.

Учитывая эффективность предложенной методики в случае с македонским, можно предположить, что покажет хорошие результаты и в остальных двух случаях, выявленных нами при исследовании топологии древа. Напомним, что они оба связаны с сербохорватскими говорами: чакавским острова Вргада и градищанским кайкавским. В отличие от предыдущего примера, исключение каждого из этих идиомов привело только к локальным изменениям внутри самой сербохорватской подгруппы и не затронуло остальные ветви древа. Тем не менее, в каждом случае это заметно повлияло на конечный вид классификации (см. рис. 9). В частности, после изъятия чакавского (рис. 9б) структура древа меняется до неузнаваемости: ранее плотная группа, состоявшая из кайкавского и двух чакавских говоров (Вргада и Девинска-Нова-Вес) распадается, причем один из них (Вргада) объединяется с другим чакавским (Орбаничи), второй образует

[14] Определить величину среднего абсолютного отклонения для корневого узла древа невозможно в силу особенностей методики — а именно, отсутствия «внешних» (по отношению к образованному узлу) языков, относительно которых можно было бы выполнить расчеты.

[15] Подчеркнем — статистическая неразличимость узлов не обязывает нас к их объединению, а только *указывает* на такую возможность. Поэтому, при наличии дополнительных (содержательных) аргументов в пользу дифференциации, близкие узлы не следует объединять, даже если диапазоны их отклонений взаимно перекрываются.

[16] Описанная методика представляет по сути своеобразный «топологический фильтр», который позволяет устранить случайные «помехи» в виде незначимых узлов, вызванных статистическими отклонениями в исходных лексических данных и «засоряющих» полезную структуру древа.

[17] Как мы уже отмечали, появление этих узлов вызвано недостатками бинарного принципа кластеризации, заложенного в методике «присоединения соседей», который не позволяет объединить более двух таксонов за один раз.

группу со штокавским (Загарач), а кайкавский становится обособленным идиомом, первым отделившимся от всей группы. В случае с градищанским (рис. 9в) мы не наблюдаем каких-либо радикальных изменений в топологии, однако обратим внимание на расположение главного узла, связывающего основные три ветви подгруппы. Его значение увеличилось сразу на 3 процента (с 89% до 92%), в результате чего произошло его сближение с узлом Вргада — Девинска-Нова-Вес (94%). Если теперь в каждом из трех фрагментов (а, б, в) мы объединим узлы с перекрывающимися отклонениями (они обведены пунктиром), то получим три несовпадающие топологии, что очевидно свидетельствует о неэффективности нашей методики в данном случае.

Как показывает дальнейший анализ, причина неудачи кроется в еще одном неучтенном факторе, а именно — особом способе подсчета процентов совпадений для узлов древа, реализованном в Starling. Согласно описанию методики расчетов в работе (Бурлак, Старостин 2005: 163–167), при объединении близкородственных языков (с долей общей лексики более 70%), следует выбирать не среднюю, а минимальную долю совпадений между ними. Авторы объясняют это тем, что «при близком родстве языков возможно вторичное их сближение, при котором трудно отличить более поздние заимствования от исконно родственной лексики»[18]. Поясним этот принцип на уже знакомом нам примере с языками A, B и C и рассчитаем долю совпадения для узла, связывающего идиом C с группой A+B.

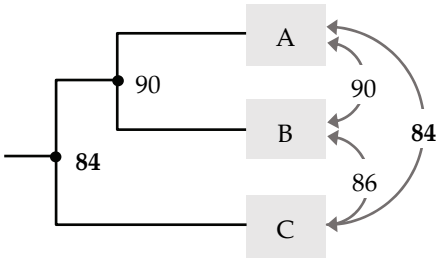
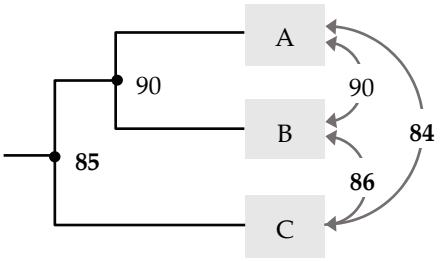


Рисунок 8. а) Расчет процентов совпадений по минимальному значению (Starling).



б) Расчет процентов совпадений по среднему значению.

Поскольку процент совпадений между списками идиомов A – C и B – C не совпадает и больше 70%, то мы, следуя данному правилу, должны выбрать наименьшее из двух значений — т. е. $N_{мин} = N_{AC} = 84$ (рис. 8а). Отметим, что выбранное значение будет отличаться от среднего процента совпадений, который для тех же языков составит $N_{ср} = (N_{AC} + N_{BC})/2 = (86+84)/2 = 85$ (рис. 8б). Причем это отличие может быть существенным, если разница между минимальным и максимальным долями совпадений окажется больше. Например, если мы примем количество общих слов в языках A и C (N_{AC}) равным 80, то среднее и минимальная доли совпадений будут отличаться уже на 3 слова: $N_{ср} = (86+80)/2 = 83$.

Несмотря на справедливость доводов, приводимых в пользу выбора минимального значения, использование такого подхода в предложенном виде трудно признать оправданным. Как уже говорилось выше, расхождения в процентах совпадений между объединяемыми языками или группами языков могут быть вызваны не только вторичным сближением между ними[19], но и — в значительно большей степени — самим случайным характером процесса лексических замен, в результате которого в лексике двух родственных языков за один и тот же выбранный промежуток времени может измениться разное

[18] Там же: 164.

[19] Которое проявляется в невыявленных поздних заимствованиях, завышающих процент совпадений при сравнении списков.

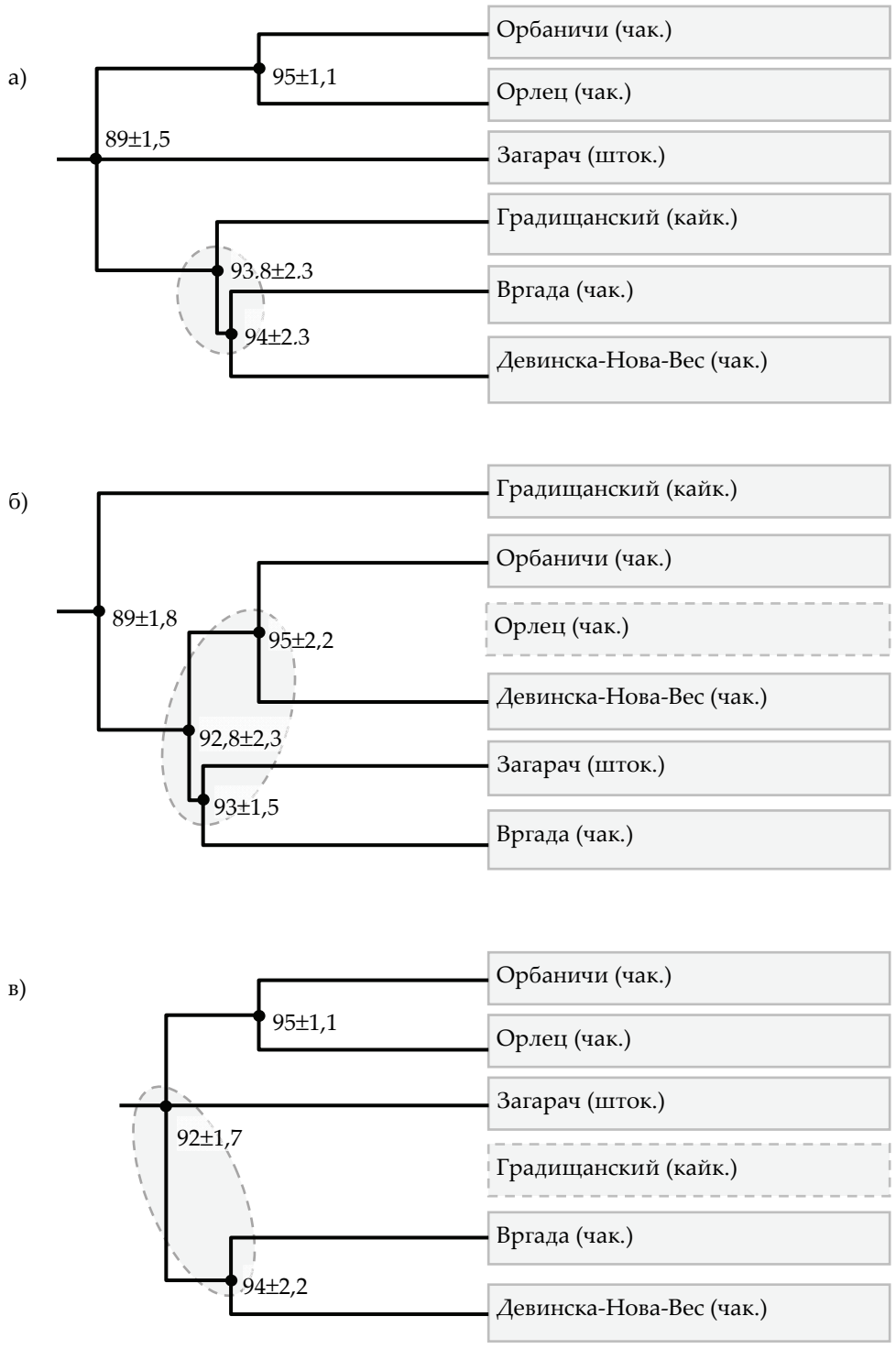


Рисунок 9. Древо, полученное для трех разных выборок сербохорватских идиомов:

а) полный список языков;

б) после исключения чакавского говора д. Орлец;

в) после исключения градищанского кайкавского.

Доли совпадений для узлов рассчитаны по наименьшим значениям.

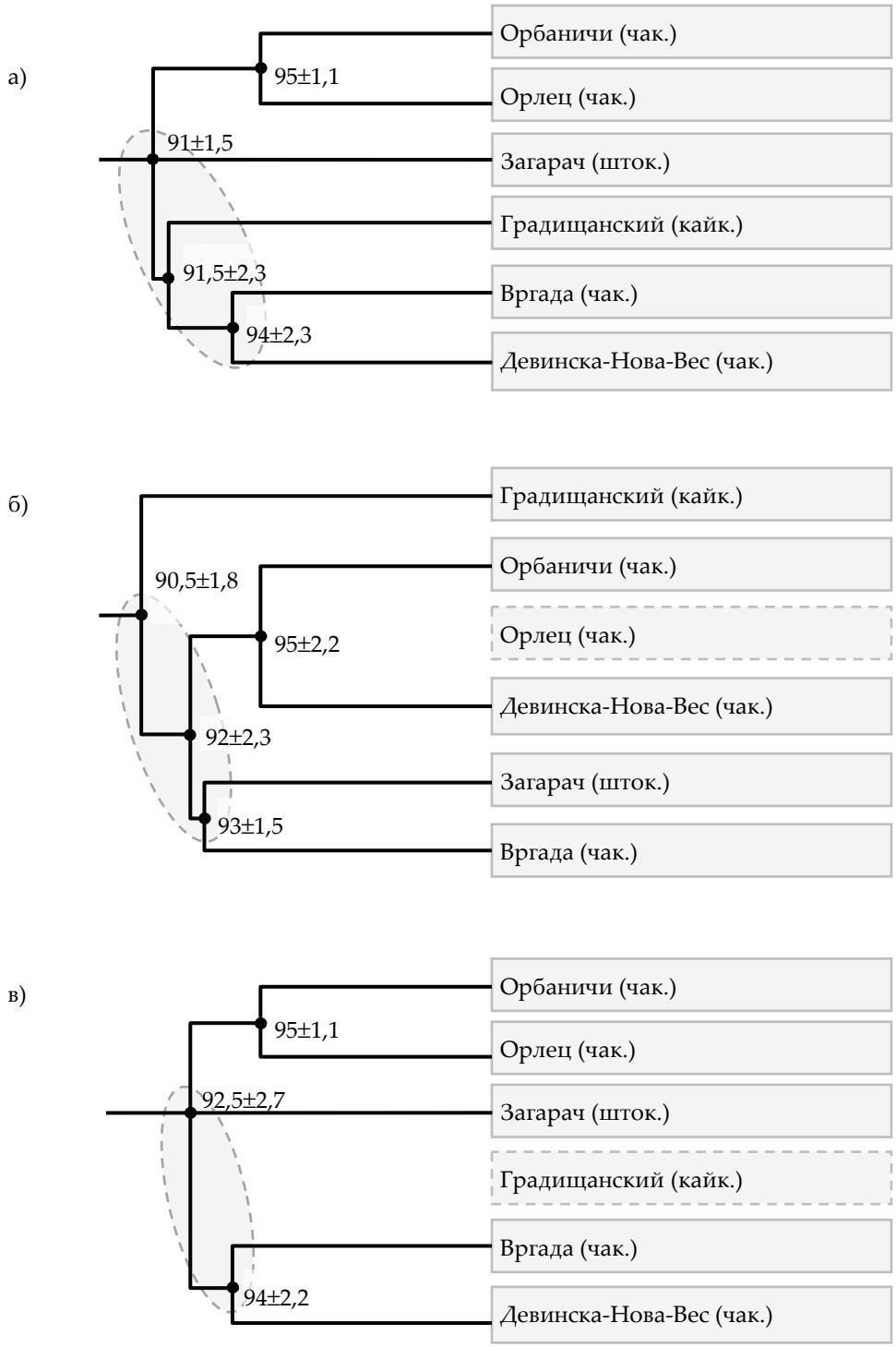


Рисунок 10. Древо, полученное, полученная для трех выборок сербохорватских идиомов после пересчета долей совпадений по средним значениям:

а) полный список языков;

б) после исключения чакавского говора д. Орлец;

в) после исключения градищанского кайкавского.

количество значений. Подчеркнем, что данная неравномерность (в отличие от вторичных сближений) не зависит от условий дивергенции языков и привносит неизбежную погрешность в *любые* лексикостатистические расчеты, как при построении деревьев, так и при расчете глоттохронологических датировок[20]. Другими словами, расхождение между долями совпадений в рамках этих погрешностей не является аномалией само по себе, а лишь отражает случайную природу лексического процесса и в большинстве случаев не требует корректировки. Поэтому попытка устранения расхождений путем отбрасывания больших значений на практике приводит к систематическому искажению исходных данных[21], в результате чего мы получаем заведомо заниженные проценты совпадений для абсолютного большинства языков, имеющих более 70% общей лексики[22].

Посмотрим, насколько существенным оказалось это искажение в случае с классификацией сербохорватской группы. Для этого пересчитаем все доли совпадений по средним значениям (рис. 10) и сравним их с рассмотренными ранее.

Прежде всего отметим, что группировка идиомов во всех трех деревьях осталась прежней. В то же время, как и ожидалось, переход к средним значениям привел к увеличению долей совпадений в основании деревьев, что отразилось в заметном сокращении расстояний между узлами. Так, в первом фрагменте с полным набором идиомов (рис. 10а), разрыв между первым и вторым узлом сократился с 5% до 0,5%, что фактически означает их полное совпадение. Во втором и третьем случаях (рис. 10б, в) это расстояние уменьшилось соответственно с 4% до 1,5% и с 2% до 1,5%, в результате чего разница между узлами оказалась в переделах статистической погрешности. Благодаря этим, на первый взгляд, несущественным изменениям, после объединения перекрывающихся узлов (обведены пунктиром) конфигурация двух деревьев (рис.10а и 10в) стала полностью идентичной, а третьего (с исключенным говором д. Орлец, рис. 10б) — очень близкой к ним[23]. Таким образом, переход к средним значениям при расчете долей совпадений, а также последующее устранение незначимых узлов древа позволили добиться топологической стабильности и прозрачности древа во всех трех случаях, выявленных нами в ходе анализа.

Вернемся теперь к исходной классификации (рис. 2) и повторим обе вышеописанные процедуры (пересчет долей совпадений по средним значениям и объединение перекрывающихся узлов) для полного генеалогического древа 25 славянских идиомов. Результаты вычисления долей совпадений по средним значениям приведены на рис. 11. Сравнение полученного древа с рис. 5 наглядно демонстрирует, насколько существенным

[20] Количественная оценка этой неравномерности для разных временных интервалов дана в статье Васильев, Саенко 2016: 272–275, а также Васильев, Саенко 2017: 128–133. Как показывают проведенные расчеты и результаты моделирования, представленные в статье, именно этот вероятностный характер процесса замен имеет определяющее значение для точности лексикостатистических расчетов.

[21] Здесь нужно добавить, что в подобной ситуации неопределенности (т. е. когда невозможно установить, какие из данных достоверны, а какие — искажены), в статистике принято использовать именно среднее значение. Применительно к нашему случаю это означает, что, если мы не можем установить факт влияния внешних факторов (будь то повторное сближение или согласованные изменения в лексике языков), то любые отклонения следует считать статистическими и, следовательно, использовать средние доли совпадений, так как замедление или ускорение процесса замен равновероятно. Предлагаемый же подход, очевидно носит не статистический, а детерминированный характер.

[22] Просматривая Таблицу 3, нетрудно убедиться, что к ним относятся почти все рассматриваемые славянские идиомы.

[23] Отделение Девинска-Нова-Вес от Вргады и присоединение к говору д. Орбаничи объясняется тем, что в отсутствие орлецкого говора, они становятся ближайшими родственниками (95% совпадений) среди оставшихся идиомов, и поэтому связываются первыми.

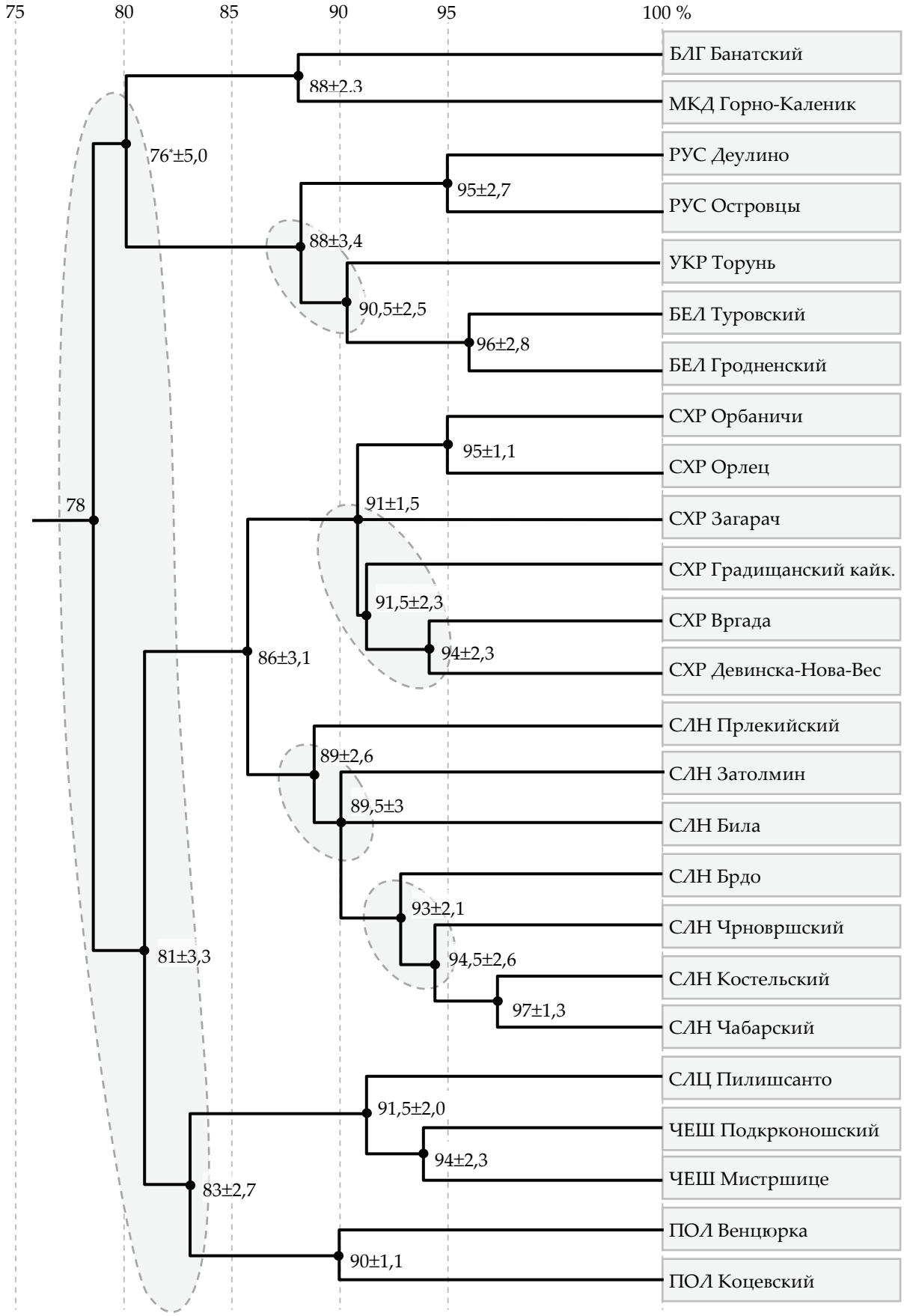


Рисунок 11. Генеалогическое древо 25 славянских идиомов с процентами совпадений, рассчитанными по средним значениям. Узлы с взаимно перекрывающимися диапазонами средних абсолютных отклонений обведены пунктиром.

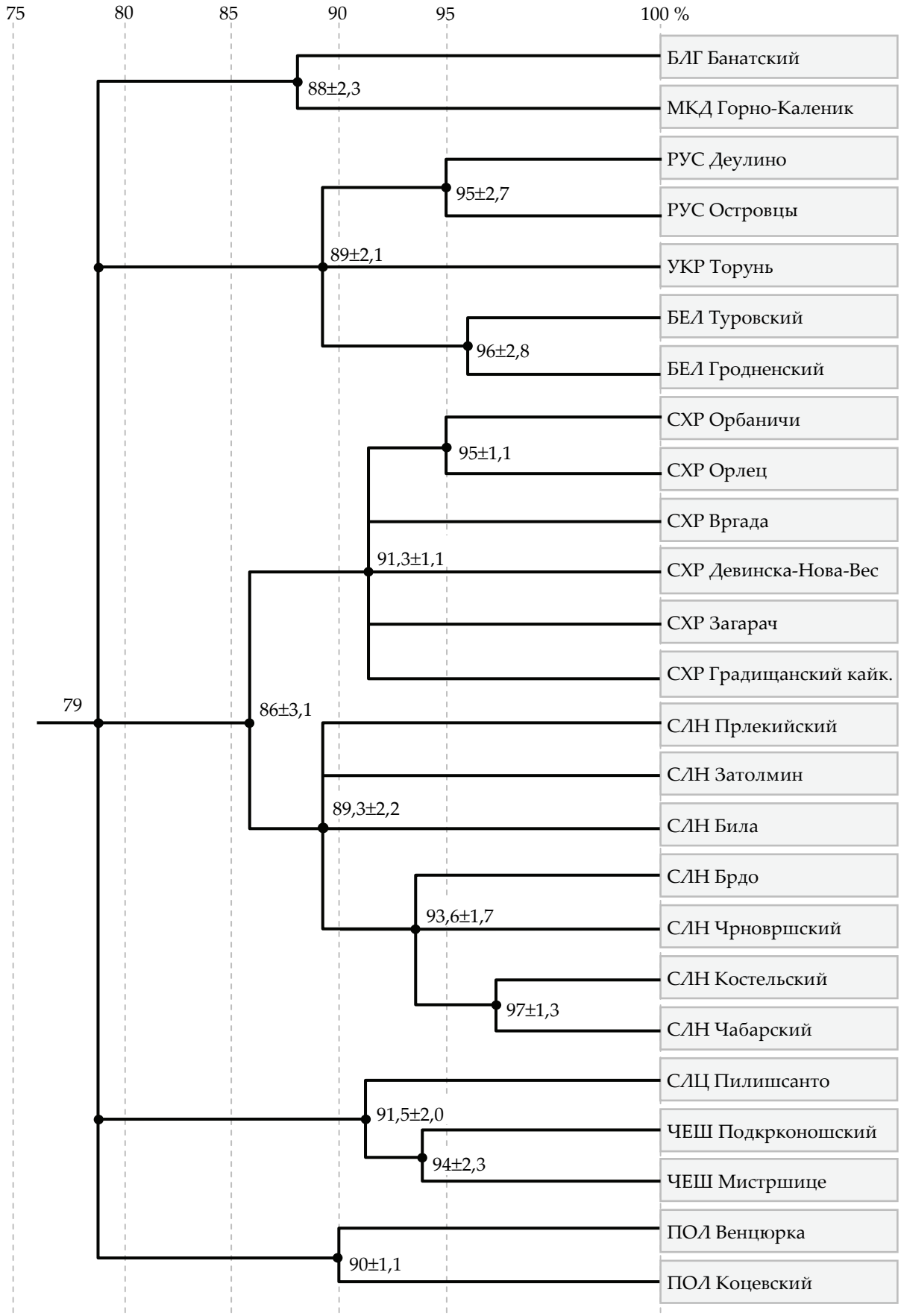


Рисунок 12. Генеалогическое древо 25 славянских идиомов с процентами совпадений, рассчитанными по средним значениям, после объединения взаимно перекрывающихся узлов.

может быть отличие между наименьшими и средними процентами совпадений. Так значение узла, объединяющего словенскую и сербохорватскую ветви, увеличилось на 6% (с 80 до 86%), а узла восточнославянских языков — на 4% (с 84 до 88%). При этом доли совпадений некоторых узлов после пересчета стали очень близкими: например, словенские прлекийский, затолминский и Била или рассмотренные выше сербохорватские идиомы. Наконец, в одном случае изменения в долях совпадений потребовали пересмотра самой структуры древа: в результате увеличения значения корневого узла с 73,6 до 78% он оказался правее узла, связывающего восточнославянскую и болгаромакедонскую ветви со средним процентом совпадений 76%, что противоречит найденной топологии[24].

Если мы перейдем к итоговому виду классификации (рис. 12), объединив узлы с перекрывающимися отклонениями в соответствии с предложенной методикой, то увидим, что различия между деревьями (ср. с рис. 7) станут еще более заметными. Вследствие уменьшения расстояний между узлами некоторые из них оказалась статистическим неразличимыми, что позволило заменить их одним общим узлом, объединяющим сразу три (например, восточнославянская подгруппа), четыре (словенская и сербохорватская подгруппы) или даже пять ветвей (а также корневой узел древа). Важно отметить, что благодаря переходу к средним значениям нам удалось не только снизить «бинарность» топологии, вызванную несовершенством процедуры объединения таксонов, но и упростить содержательную интерпретацию древа[25].

Если сравнивать полученное древо с исходным (рис. 2), то можно отметить целый ряд положительных изменений в итоговой топологии. Узлы перестали быть строго бинарными, что приблизило древо ко классификациям, построенным традиционными методами. Другим важным преимуществом стало отсутствие «лесенки» в сербохорватской и словенской ветвях: 110-словных списков явно недостаточно для построения точной классификации столь близких диалектов. Нельзя не отметить, что те говоры, которые все же объединены в бинарные узлы (Орбаничи и Орлец, костельский и чабарский), действительно очень близки друг к другу как в географическом, так и генетическом отношении.

Исчезли фантомные корневые объединения, вызванные тем, что алгоритм Starling вынуждает всё сводить к бинарному виду. На первый взгляд, некоторую проблему может представлять то, что по сравнению с первым древом западнославянская ветвь распалась на две части: польскую и чешско-словацкую. Однако имеется не так много древних фонетических инноваций, объединяющих западнославянские языки и противопоставляющих их всем остальным. В первую очередь это аффрикативизация *ť > *c и *ď > *ʒ, а также переход *x по второй и третьей палатализациям в *š, а не *s'[26]. В то же время, допустим, рефлексация сочетаний вида TorT и TolT сближает чешско-словацкую под-

[24] Данный пример свидетельствует о том, что использование минимальных значений вместо средних приводит не только к систематическому занижению долей совпадений, но влияет также на саму последовательность объединения таксонов, а следовательно — непосредственно сказывается на полученной конфигурации древа.

[25] Прежде всего это проявляется в уменьшении количества незначимых узлов, которые невозможно сопоставить каким-либо фактическим или предполагаемым событиям в истории развития языков.

[26] Часто в качестве западнославянских черт рассматривают сохранение групп *tl и *dl, а также *kw и *gw перед *ě и *i. Однако эти особенности не являются эксклюзивно западнославянскими и, что важнее, представляют из себя архаизмы, а не инновации, что серьезно снижает их ценность для генеалогической классификации.

группу с южнославянскими языками, а не лехитскими и лужицкими. Таким образом, западнославянскую подгруппу (как и южнославянскую) нельзя считать таксоном с доказанным статусом.

Фактически, именно отсутствием объединения болгаро-македонского и сербохорватско-словенского таксонов в южнославянскую подгруппу, а чешско-словацкого и лехитского — в западнославянскую, полученное нами древо и отличается от общепринятой классификации славянских языков, представленной в (Иванов 1990: 95).

Оговорим, что полученная нами схема не может служить в качестве аргумента против необходимости выделения южно- и западнославянской подгрупп, поскольку она базируется на неполном материале: отсутствуют данные лужицких, кашубского, словинского и полабского языков, а также торлакского наречия. Кроме того, как уже отмечалось ранее, даже выполнение условия о взаимном перекрытии соседних узлов не обязывает нас к их объединению, а лишь указывает на такую возможность. Поэтому при наличии дополнительных доводов в пользу сохранения западнославянской общности, она может быть выделена на итоговом древе.

В завершение ещё раз напомним, что предложенная методика не является самостоятельным методом классификации и не формирует её структуру, а применяется к *уже построенному* генеалогическому древу и позволяет оценить достоверность его топологии на основе статистических расчетов. С одной стороны, это несколько ограничивает ее возможности[27], но с другой — делает ее универсальной и дает возможность использовать ее для анализа любых лексикостатистических классификаций, вне зависимости от способа их построения.

IV. Заключение

Подведем итоги нашего исследования, сформулировав основные полученные теоретические выводы и практические результаты:

a. Накопленный опыт лексикостатистической классификации языков с помощью системы Starling свидетельствует о том, что, несмотря на общую правдоподобность и полезность получаемых результатов, в построенных генеалогических деревьях обнаруживаются внутренние несоответствия и артефакты, не поддающиеся объяснению или противоречащие известным данным.
b. Наиболее распространенными из этих несоответствий являются:

- *проблема вариативности* — неустойчивость конфигурации древа при изменении количества или состава идиомов;
- *проблема избыточной кластеризации* — древо содержит большое количество близкорасположенных узлов, интерпретация которых проблематична или невозможна.

c. Основная причина возникновения указанных проблем заключается в несовершенстве методики построения древа, реализованной в Starling, которая, с одной стороны, игнорирует вероятностный характер лексических расчетов, а с другой — привносит фиксированные поправки в исходные данные, что приводит к появлению статистических погрешностей, а также систематических ошибок в полученной классификации.

[27] Так как она, очевидным образом, не может изменить порядок объединения узлов, хотя и указывает на необходимость такого изменения (как в случае с болгаро-македонским/восточнославянским узлом на рис. 11).

d. В работе предложена методика анализа классификации, позволяющая количественно оценить возникающие погрешности на основе величины среднего абсолютного отклонения, а также минимизировать их влияние на строение древа с помощью процедуры устранения недостоверных узлов.
e. Применение методики для анализа генеалогического древа 25 славянских идиомов показало, что она позволяет эффективно решать обе выявленные проблемы, связанные с нестабильностью топологии и избыточной кластеризацией древа, а также значительно улучшить полученную классификацию древа с точки зрения ее содержательной интерпретации.
f. Универсальность разработанной методики дает возможность применять ее для анализа любых генеалогических деревьев, полученных на основе лексикостатистических расчетов, независимо от того, каким способом они были построены.
g. Все основные процедуры методики хорошо формализуются и могут быть реализованы в виде дополнительного модуля Starling, или самостоятельной утилиты, удобной для практического использования при анализе и уточнении лексикостатистических классификаций.

Литература

Бурлак, С. А., С. А. Старостин. 2005. *Сравнительно-историческое языкознание*. Москва: Академия.

Васильев, М. Е. 2010. Об использовании лексического критерия для построения генеалогической классификации. *Бюллетень Общества востоковедов РАН* 17: 530–572.

Васильев, М. Е., А. И. Коган. 2013. К вопросу о восточнодардской языковой общности. *Вопросы языкового родства* 10: 149–177.

Васильев, М. Е., М. Н. Саенко. 2016. К вопросу о точности глоттохронологии: датирование процесса лексических замен по данным романских языков. *Вопросы языкового родства* 14/3–4: 259–277.

Васильев, М. Е., М. Н. Саенко. 2017. К вопросу о точности глоттохронологии: датирование языковой дивергенции по данным романских языков. *Вопросы языкового родства* 15/1–2: 114–135.

Грунтов, И. А., О. М. Мазо. 2015. Классификация монгольских языков по лексикостатистическим данным. *Вопросы языкового родства* 13/3–4: 205–255.

Иванов Вяч. Вс. 1990. Генеалогическая классификация языков. В: *Лингвистический энциклопедический словарь*. М.: Советская энциклопедия: 93–98.

Николаев, С. Л., М. Н. Толстая. 2001. *Словарь карпатоукраинского торуньского говора*. М.: Институт славяноведения РАН.

Поздняков, К. И. 2014. О пороге родства и индексе стабильности в базисной лексике при массовом сравнении: атлантические языки. *Вопросы языкового родства* 11: 187–225.

ССРНГ 1969 = *Словарь современного русского народного говора (д. Деулино Рязанского района Рязанской области)*. Москва: Наука

Стойков, С. 1968. *Лексиката на банатския говор*. София: Издателство на Българската академия на науките.

Стойков, С. 2002. *Българска диалектология*. София: Проф. Марин Дринов.

Сцяшковіч, Т. Ф. 1972. *Матэрыялы да слоўніка Гродзенскай вобласці*. Мінск: Навука і тэхніка.

Сцяшковіч, Т. Ф. 1983. *Слоўнік Гродзенскай вобласці*. Мінск: Навука і тэхніка.

ТС = *Тураўскі слоўнік*. 1982–1987. Тамы 1–5. Мінск: Навука и тэхніка.

Ћупић Д., Ћупић Ж. 1997. *Речник говора Загарача*. Београд: Српска академија наука и уметности, Институт за српски језик САНУ.

Хонселаар, З. 2001. *Говор деревни Островцы Псковской области*. Amsterdam: Rodopi.

References

Bachmannová, Jarmila. 2016. *Slovník podkrkonošského nářečí*. Praha: Academia.

Blažek, Václav. Klasifikace slovanských jazyků. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.). *CzechEncy — Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/KLASIFIKACE SLOVANSKÝCH JAZYKŮ.

Burlak, Svetlana A., Sergey A. Starostin. 2005. *Sravnitel'no-istoricheskoje jazykoznanije*. Moscow: Academia.

Čujec Stres, Helena. 2011–2014. *Slovar zatolminskega govora*. Dela 1–2. Zatolmin: Stres inženiring.

Ćupić, Drago, Željko Ćupić. 1997. *Rečnik govora Zagarača*. Beograd: Srpska akademija nauka i umetnosti, Institut za srpski jezik SANU.

Gregor, Ferenc. 1975. *Der slowakische Dialekt von Pilisszántó*. Budapest: Akadémiai Kiadó.

Gregorič, Jože. 2014. *Kostelski slovar*. Ljubljana: Založba ZRC.

Gruntov, Ilya A., Olga M. Mazo. 2015. Klassifikacija mongol'skix jazykov po leksikostatisticheskim dannym. *Journal of Language Relationship* 13/3–4: 205–255.

Hill, Peter. 1991. *The Dialect of Gorno Kalenik*. Columbus: Slavica Pub.

Honselaar, Zep. 2001. *Govor derevni Ostrovtsy Pskovskoy oblasti*. Amsterdam: Rodopi.

Houtzagers, Peter. 1985. *The Čakavian dialect of Orlec and the islands of Cres*. Amsterdam: Brill.

Houtzagers, Peter. 1999. *The Kajkavian Dialect of Hidegség and Fertőhomok*. Amsterdam: Rodopi.

Ivanov, Vyacheslav V. 1990. Genealogicheskaja klassifikacija jazykov. In: *Lingvisticheskij enciklopedicheskij slovar'*: 93–98. Moscow: Sovetskaja enciklopedija.

Jurišić, Blaž. 1973. *Rječnik govora otoka Vrgade*. Zagreb: Izdavački zavod Jugoslavenske akademije; GZH — ZRINSKI.

Kalsbeek, Janneke. 1998. *The Čakavian Dialect of Orbanići near Žminj in Istria*. Amsterdam — Atlanta: Rodopi.

Kogan, Anton I. 2016. Genealogical classification of New Indo-Aryan languages and lexicostatistics. *Journal of Language Relationship* 14/3–4: 227–258.

Kucała, Marian. 1957. *Porównawczy słownik trzech wsi małopolskich*. Wrocław: Zakład im. Ossolińskich — Wydawnictwo PAN.

Malina, Ignát. 1946. *Slovník nářečí mistřického*. Praha: Česká akademie věd a umění.

Malnar, Slavko. 2008. *Rječnik govora čabarskog kraja*. Čabar: Matica hrvatska — Ogranak u Čabru.

Nikolaev, Sergey L., Marfa N. Tolstaja. 2001. *Slovar' karpatoukrainskogo torun'skogo govora*. Moscow: Institut slavianovedenija RAN.

Pozdniakov, Konstantin I. 2014. O poroge rodstva i indekse stabil'nosti v bazisnoj leksike pri massovom sravnenii: atlanticheskije jazyki. *Journal of Language Relationship* 11: 187–225.

Pronk, Tijmen. 2009. *The Slovene Dialect of Egg and Potschach in the Gailtal, Austria*. Amsterdam — New York: Rodopi.

Rajh, Bernard. 2010. *Gúčati po antùjoško. Gradivo za narečni slovar severozahodnoprleškega govora*. Maribor: Filozofska fakulteta — Mednarodna založba Oddelka za slovanske jezike in književnosti.

Saitou, Naruya, Masatoshi Nei. 1987. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* 4 (4): 406–425.

Sciaškovič, Tatjana F. 1972. *Materyjały da słoŭnika Hrodzienskaj vobłasci*. Minsk: Navuka i technika.

Sciaškovič, Tatjana F. 1983. *Słoŭnik Hrodzienskaj vobłasci*. Minsk: Navuka i technika.

SSRNG 1969 = Ossovetski I.A. (ed.). *Slovar' sovremennogo russkogo narodnogo govora (d. Deulino Riazanskogo rajona Riazanskoy oblasti)*. Moskva: Nauka.

Starostin, George. 2010. Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship* 3: 79–116.

Steenwijk, Han. 1992. *The Slovene dialect of Resia: San Giorgo*. Amsterdam — Atlanta: Rodopi.

Stojkov, Stojko. 1968. *Leksikata na banatskija govor*. Sofija: Izdatelstvo na Bălgarskata akademija na naukite.

Stojkov, Stojko. 2002. *Bălgarska dialektologija*. Sofija: Prof. Marin Drinov.

Sychta, Bernard. 1980–1985. *Słownictwo kociewskie*. Tomy 1–3. Wrocław — Warszawa — Kraków — Gdańsk: Zakład narodowy im. Ossolińskich — Wydawnictwo PAN.

Tominec, Ivan. 1964. *Črnovrški dialect*. Ljubljana: Slovenska akademija znanosti in umetnosti.

TS = Kryvitski, A. A. (ed.). 1982–1987. *Turaŭski słoŭnik*. Tamy 1–5. Minsk: Navuka i technika.

Vasilyev, Mikhail E. 2010. Ob ispol'zovanii leksicheskogo kriterija dl'a postrojenija genealogicheskoj klassifikacii. *B'ulleten' Obshchestva vostokovedov RAN* 17: 530–572.

Vasilyev, Mikhail E., Anton I. Kogan. 2013. K voprosu o vostochnodardskoj jazykovoj obshchnosti. *Journal of Language Relationship* 10: 149–177.

Vasilyev, Mikhail E., Mikhail N. Saenko. 2016. K voprosu o tochnosti glottoxronologii: datirovanije processa leksicheskix zamen po dannym romanskix jazykov. *Journal of Language Relationship* 14/3–4: 259–277.

Vasilyev, Mikhail E., Mikhail N. Saenko. 2017. K voprosu o tochnosti glottoxronologii: datirovanije jazykovoj divergencii po dannym romanskix jazykov. *Journal of Language Relationship* 15/1–2: 114–135.
Vážný, Václav. 1927. *Čakavské nářečí v slovenském Podunají*. Bratislava: Filosofická fakulta University Komenského.
Vydrin, Valentin. 2009. On the Problem of the Proto-Mande Homeland. *Journal of Language Relationship* 1: 107–142.

Mikhail Vasilyev, Mikhail Saenko. An analysis of the topology and estimation of accuracy for lexicostatistical classifications (on the data of Slavic languages)

Today, lexicostatistical methods are widely used in comparative-historical linguistics to establish linguistic kinship and build genealogical classifications. In works by Russian comparative linguists the most common technique is construction of phylogenetic trees obtained with the aid of the Starling software, developed by Sergei Starostin at the end of the 20th century. Starostin's algorithm was based on a modified method of "neighbor joining" and yielded satisfactory or plausible results in the vast majority of cases. At the same time, many researchers have pointed out a number of significant shortcomings in the obtained classifications, the most serious of which are the instability of the tree caused by even minimal changes in the number of idioms, as well as detection of a large number of fictitious taxa and nodes that are poorly explained or even contradict existing concepts. This article provides a detailed examination of these shortcomings based on the example of a new lexicostatistical classification for 25 Slavic lects. Upon detailed analysis, we propose a special procedure that allows to minimize the negative effect of identified deficiencies on the structure of the tree, making use of statistical analysis of the resulting topology and capable of identifying unreliable nodes within it. The technique is simple enough to to be practically implemented in the form of an additional Starling component or a separate application.

Keywords: lexicostatistics; neighbor-joining method; genealogical classification; mean absolute deviation.

Приложение

	BAN	GKM	ZAG	ORB	ORL	VRG	DNV	BRG	CVS	KOS	ZTL	RES	GLT	PRL	CAB	PKC	MIS	PLS	WLP	KGP	TUR	GRD	TOR	DEU	OST
BAN	1	0.88	0.8	0.76	0.77	0.77	0.77	0.76	0.73	0.77	0.68	0.72	0.75	0.73	0.76	0.74	0.72	0.71	0.67	0.68	0.72	0.73	0.76	0.73	0.75
GKM	0.88	1	0.84	0.79	0.79	0.8	0.8	0.78	0.74	0.77	0.71	0.74	0.73	0.72	0.75	0.77	0.74	0.72	0.68	0.68	0.77	0.79	0.79	0.77	0.77
ZAG	0.8	0.84	1	0.89	0.89	0.93	0.92	0.89	0.84	0.88	0.8	0.83	0.85	0.83	0.88	0.82	0.8	0.8	0.76	0.78	0.81	0.82	0.85	0.79	0.8
ORB	0.76	0.79	0.89	1	0.95	0.93	0.95	0.9	0.87	0.91	0.8	0.84	0.85	0.86	0.9	0.83	0.82	0.83	0.79	0.81	0.79	0.82	0.83	0.76	0.8
ORL	0.77	0.79	0.89	0.95	1	0.92	0.93	0.89	0.86	0.89	0.81	0.85	0.85	0.84	0.88	0.83	0.82	0.82	0.78	0.77	0.78	0.81	0.82	0.75	0.78
VRG	0.77	0.8	0.93	0.93	0.92	1	0.94	0.89	0.87	0.91	0.8	0.84	0.85	0.84	0.89	0.81	0.81	0.82	0.78	0.79	0.8	0.83	0.85	0.77	0.81
DNV	0.77	0.8	0.92	0.95	0.93	0.94	1	0.94	0.9	0.94	0.84	0.89	0.89	0.88	0.93	0.86	0.84	0.83	0.81	0.82	0.81	0.84	0.85	0.77	0.8
BRG	0.76	0.78	0.89	0.9	0.89	0.89	0.94	1	0.87	0.88	0.8	0.87	0.86	0.86	0.87	0.82	0.82	0.8	0.78	0.78	0.78	0.81	0.83	0.76	0.81
CVS	0.73	0.74	0.84	0.87	0.86	0.87	0.9	0.87	1	0.94	0.94	0.9	0.94	0.91	0.95	0.83	0.79	0.79	0.79	0.8	0.76	0.77	0.77	0.76	0.8
KOS	0.77	0.77	0.88	0.91	0.89	0.91	0.94	0.88	0.94	1	0.88	0.89	0.91	0.91	0.97	0.86	0.85	0.84	0.81	0.83	0.8	0.82	0.82	0.76	0.79
ZTL	0.68	0.71	0.8	0.8	0.81	0.8	0.84	0.8	0.94	0.88	1	0.87	0.88	0.87	0.91	0.79	0.75	0.73	0.76	0.75	0.73	0.75	0.74	0.73	0.77
RES	0.72	0.74	0.83	0.84	0.85	0.84	0.89	0.87	0.9	0.89	0.87	1	0.89	0.86	0.89	0.82	0.79	0.78	0.76	0.79	0.79	0.8	0.81	0.75	0.81
GLT	0.75	0.73	0.85	0.85	0.85	0.85	0.89	0.86	0.94	0.91	0.88	0.89	1	0.89	0.93	0.84	0.81	0.82	0.79	0.8	0.75	0.77	0.79	0.75	0.8
PRL	0.73	0.72	0.83	0.86	0.84	0.84	0.88	0.86	0.91	0.91	0.87	0.86	0.89	1	0.9	0.83	0.85	0.82	0.82	0.82	0.78	0.8	0.81	0.76	0.79
CAB	0.76	0.75	0.88	0.9	0.88	0.89	0.93	0.87	0.95	0.97	0.91	0.89	0.93	0.9	1	0.86	0.83	0.82	0.8	0.81	0.78	0.79	0.79	0.76	0.79
PKC	0.74	0.77	0.82	0.83	0.83	0.81	0.86	0.82	0.83	0.86	0.79	0.82	0.84	0.83	0.86	1	0.94	0.9	0.8	0.81	0.78	0.81	0.82	0.76	0.79
MIS	0.72	0.74	0.8	0.82	0.82	0.81	0.84	0.82	0.79	0.85	0.75	0.79	0.81	0.85	0.83	0.94	1	0.93	0.84	0.85	0.77	0.83	0.83	0.74	0.75
PLS	0.71	0.72	0.8	0.83	0.82	0.82	0.83	0.8	0.79	0.84	0.73	0.78	0.82	0.82	0.82	0.9	0.93	1	0.83	0.85	0.79	0.86	0.83	0.78	0.78
WLP	0.67	0.68	0.76	0.79	0.78	0.78	0.81	0.78	0.79	0.81	0.76	0.76	0.79	0.82	0.8	0.8	0.84	0.83	1	0.9	0.78	0.84	0.84	0.77	0.77
KGP	0.68	0.68	0.78	0.81	0.77	0.79	0.82	0.78	0.8	0.83	0.75	0.79	0.8	0.82	0.81	0.81	0.85	0.85	0.9	1	0.78	0.83	0.83	0.75	0.77
TUR	0.72	0.77	0.81	0.79	0.78	0.8	0.81	0.78	0.76	0.8	0.73	0.79	0.75	0.78	0.78	0.78	0.77	0.79	0.78	0.78	1	0.96	0.89	0.88	0.89
GRD	0.73	0.79	0.82	0.82	0.81	0.83	0.84	0.81	0.77	0.82	0.75	0.8	0.77	0.8	0.79	0.81	0.83	0.86	0.84	0.83	0.96	1	0.92	0.9	0.91
TOR	0.76	0.79	0.85	0.83	0.82	0.85	0.85	0.83	0.77	0.82	0.74	0.81	0.79	0.81	0.79	0.82	0.83	0.83	0.84	0.83	0.89	0.92	1	0.84	0.87
DEU	0.73	0.77	0.79	0.76	0.75	0.77	0.77	0.76	0.76	0.76	0.73	0.75	0.75	0.76	0.76	0.76	0.74	0.78	0.77	0.75	0.88	0.9	0.84	1	0.95
OST	0.75	0.77	0.8	0.8	0.78	0.81	0.8	0.81	0.8	0.79	0.77	0.81	0.8	0.79	0.79	0.79	0.75	0.78	0.77	0.77	0.89	0.91	0.87	0.95	1

Таблица 3. Исходная таблица долей совпадений между 110-словными списками 25 славянских идиомов.

Условные обозначения: BAN — Банатский болгарский; GKM — Македонский говор д. Горно-Каленик; ZAG — Штокавский сербохорватский племени Загарач; ORB — Чакавский сербохорватский говор д. Орбаничи; ORL — Чакавский сербохорватский говор д. Орлец и острова Црес; VRG — Чакавский сербохорватский говор острова Вргада; DNV — Чакавский сербохорватский говор д. Девинска-Нова-Вес; BRG — Градищанский кайкавский сербохорватский; CVS — Чрноврш-ский словенский; KOS — Костельский словенский; ZTL — Затолминский словенский; RES — Резьянский словенский д. Била.; GLT — Словенский д. Брдо; PRL — Прлешский словенский; CAB — Словенский говор Чабара и окрестностей; PKC — Подкрконошский чешский; MIS — Моравский чешский д. Мистршице; PLS — Словацкий говор д. Пилишсанто; WLP — Малопольский диалект д. Венцюрка; KGP — Коцевский великопольский; TUR — Белорусские говоры Турова и окрест-ностей; GRD — Белорусские говоры Гродненской области; TOR — Украинский говор д. Торунь; DEU — Русский говор д. Деулино; OST — Русский говор д. Островцы