Лексикостатистика и славянские языки

Алексей Касьян

В статье дается обзор основных экспериментов по применению лексикостатистических методов к классификации славянских языков, а также кратко разбираются теоретические и практические проблемы, с которыми сталкивается лексикостатистика. Русский перевод статьи «Lexicostatistics and Slavic languages» для энциклопедии: Marc L. Greenberg et al. (eds.), *Encyclopedia of Slavic Languages and Linguistics Online*. Brill.

Введение

Лексикостатистика (Swadesh 1952; Swadesh 1955; Tischler 1973; Embleton 2000; Starostin 2000; McMahon & McMahon 2005; Starostin 2010) в широком понимании – это формализованная математическая процедура, которая оценивает по лексическим данным степень близости между языками. Можно сказать, что лексикостатистика измеряет *лексические дистанции* между языками. Оцениваемая близость может отражать как генеалогическое родство, так и контакты в зависимости от того, как мы составляем лексические списки, предназначенные для компьютерного анализа (выкидываем ли мы заимствования или нет), и какой метод обсчета используется. Дополнительная процедура, которая датирует узлы дерева, иногда называется *глоттохронологией*. В настоящее время иногда предлагается (конечно же, в рекламных целях) оставить термин лексикостатистика для лексикостатистического анализа традиционными дистантными методами (NeighborJoining или UPGMA, см. ниже), а для недавно введенных в лингвистический оборот признаковых методов (прежде всего байесовский вывод) использовать другие ярлыки.

Графический результат лексикостатистического анализа – это или *дерево* (граф, в котором любая пара вершин соединена одним и только одним путем), или филогенетическая *сеть* (граф, в котором пара вершин может иметь и в нормальном случае имеет более одного пути между собой). Конечные узлы (листья) такого дерева или сети представляют собой анализируемые языки.

Деревья (Jacques & List 2019) в идеальном случае должны отражать непосредственную историю человеческих популяций, где узлы с разделением языков соответствуют расселению и миграциям. Корень дерева представляет собой праязык всех языков, включенных в анализ, а промежуточные узлы – это праязыки отдельных языковых групп и подгрупп. Иными словами, деревья несут генеалогическую информацию.

Сети (Heggarty, Maguire & McMahon 2010; Huson & Scornavacca 2011) представляют информацию двух видов: генеалогическую и контактную. Это значит, что сети отображают как общие черты, унаследованные от общего предка (праязыка), так и общие черты, заимствованные из одного языка в другой (или же параллельно независимо развившиеся). Надо подчеркнуть, что сети графически не разделяют эти два типа сигнала близости языков.

Подавляющее большинство вычислительных методов и компьютерных программ, сегодня используемых для лексикостатистики, были импортированы из биологии, прежде всего из генетики, а не разработаны для нужд лингвистов. Открытым остается вопрос, какие из многочисленных имеющихся математических алгоритмов реконструкции филогении лучше удовлетворяет естественной языковой эволюции, и, к сожалению, лингвисты редко задаются этим вопросом. Представляется, что пока не разработан такой алгоритм, который в достаточной степени отражал бы основные особенности эволюции лексики. Однако с практической точки зрения можно утверждать следующее:

1. По крайней мере, такие методы в целом подходят для лексикостатистического анализа: NeighborJoining, максимальная экономия, байесовский вывод (байесовский вывод является наиболее популярным на сегодняшний день, но результат такого анализа очень сильно зависит от предварительно заданных условий и настроек, Yanovich 2020).
2. Адекватность получаемых деревьев прежде всего зависит от аккуратности и лексикографического качества входных лексических списков, нежели от того или иного математического алгоритма или софта (Kassian 2015).

Как уже было упомянуто выше, лексикостатистический анализ основан на лексических списках, которые вручную составляют лингвисты для данных языков. Обычно используются различные выборки из базисного словаря, поскольку культурный

словарь в общем случае менее устойчив и в большей степени подвержен заимствованию.

Наиболее популярной такой выборкой является так называемый список Сводеша, который был сначала сформулирован в 200-словной версии (Swadesh 1952), а затем сокращен до 100-словника (Swadesh 1955). На сегодняшний день было предложено и используется теми или иными лингвистами большое количество вариаций сводешевского списка (List et al. 2023), из них наиболее подходящими как для работы со словарями, так и для полевой работы оказываются списки, где концепты сопровождены семантическими спецификациями и диагностическими контекстами: это 110-словник, используемый в Московской школе сравнительно-исторического языкознания (Kassian et al. 2010) и 207-словник, разработанный Микаелем Данном (Michael Dunn) и Кейт Беллами-Дворак (Kate Bellamy-Dworak) для индоевропеистического проекта Bouckaert et al. 2012 (к сожалению, насколько мне известно, этот 207-словник остается не опубликованным).

Высокое лексикографическое качество и семантическая унификация входных лексических списков являются ключевыми условиями надежного филогенетического результата. Необходимо отдавать себе отчет, что компьютерная программа выдаст какую-нибудь филогению в любом случае, хорошие списки были поданы на вход или плохие (этот принцип называется “garbage in, garbage out”), таким образом, если в некой статье вы видите красивое дерево или сеть для языковой группы, это совсем не значит, что данная филогенетическая схема автоматически заслуживает доверия. К сожалению, многие авторы лексикостатистических исследований делают основной упор на усложнение математического аппарата в ущерб подготовке языковых данных.

В заключение надо подчеркнуть, что лексикостатистическое разделение (бифуркация на дереве) обозначает первую лексическую замену в используемой версии списка (как, например, в русском языке слово *око* было недавно заменено на *глаз*). На этой стадии дивергенции диалекты еще остаются взаимопонимаемыми и, если находятся в ситуации контакта, легко переживающими общие инновации в фонетике, морфологии и синтаксисе.

Применение к славянским языкам

Славянские языки – это очень хорошо изученная языковая группа, ее трехчастная структура – западная подгруппа, восточная и южная – в целом не вызывает вопросов, даже если позиция отдельных идиомов, например, древненовгородского языка, и вызывает споры. Таким образом славянская группа – это хороший полигон для проверки лексикостатистических методов и подходов. Ниже я перечисляю и комментирую основные лексикостатистические эксперименты со славянскими языками (более широкий обзор с дополнительными историческими экскурсами можно найти в Blažek 2020).

В **Kushniarevich et al. 2015** (S2 File “Linguistics: Datasets; Methods; Results”) приводится датированное славянское дерево, которое в целом совпадает с мнением славистов и не содержит никаких явных несообразностей. Дерево устойчивое, все узлы имеют высокую статистическую поддержку. Анализ в Kushniarevich et al. 2015 строится на высококачественных 110-словниках, частично составленных по авторитетным словарям, частично собранным в поле от информантов в соответствии с семантическими спецификациями, предложенными в Kassian et al. 2010. Этимологизированные списки были обсчитаны основными алгоритмами (Bayesian inference, NeighborJoining, UPGMA, Maximum Parsimony). Итоговое консенсусное дерево (рис. 1) показывает троичное разделение около 100 г. н.э. на западную, восточную и южную подгруппы, а затем в середине 1-го тысячелетия эти подгруппы почти синхронно распадаются каждая на две или три ветви – происходит так называемая славянизация Европы.

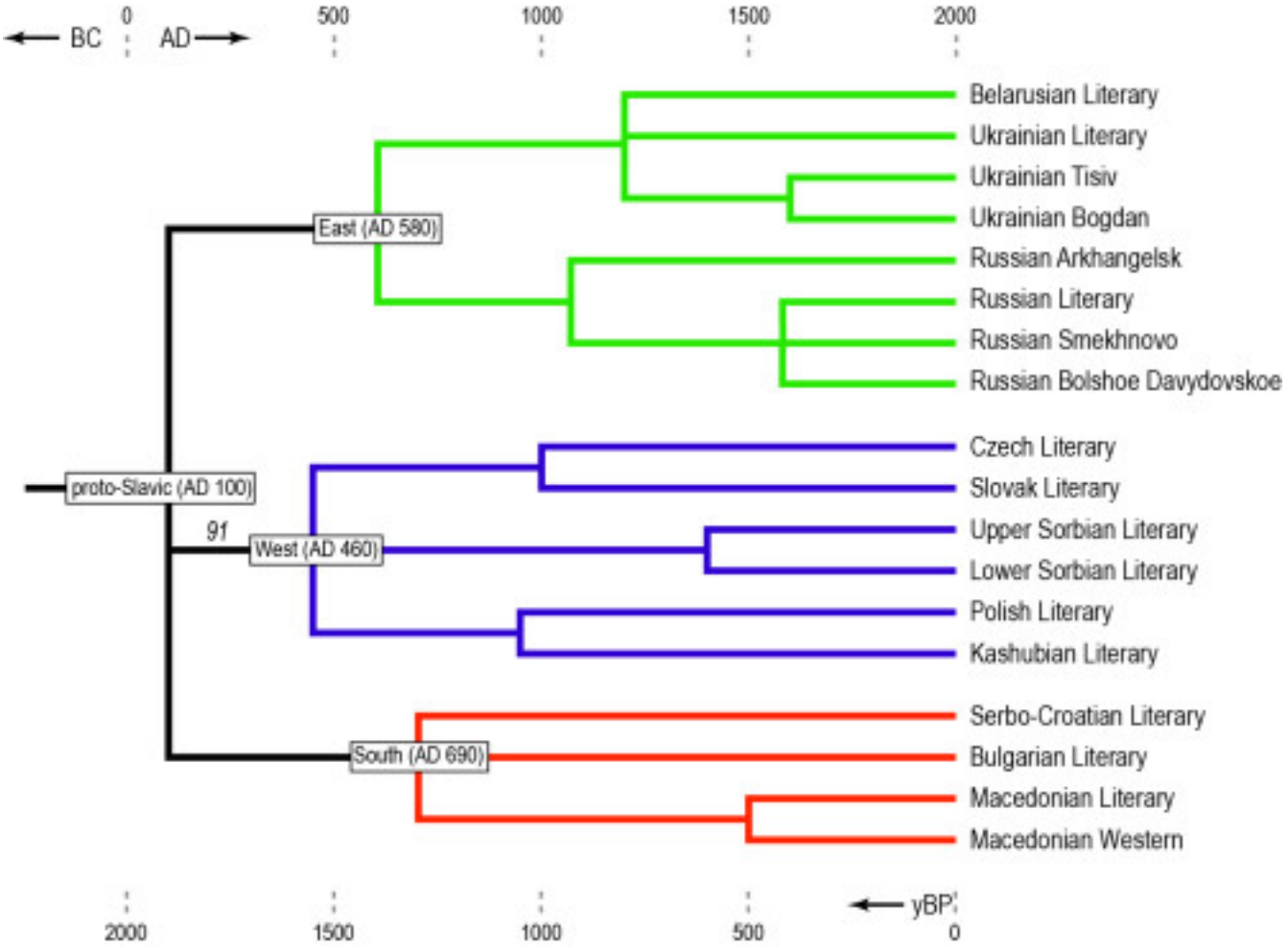

Рис. 1. Kushniarevich et al. 2015: датированное лексикостатистическое дерево славянских языков. 110-словники. Несколько алгоритмов: Bayesian inference (no constraints), NeighborJoining, UPGMA, Maximum Parsimony. Тройные узлы получены в результате объединения соседних бинарных узлов, если временно́е расстояние между ними ≤ 300 лет. Байесовские апостериорные вероятности даны курсивом рядом с узлами (не указаны для стабильных узлов с P ≥ 0.95). Длина ветвей показывает абсолютную хронологию. (Цит. по: Kushniarevich et al. 2015, Fig. G in S2 File)

В Kushniarevich et al. 2015 отдельно обсуждается проблема словенского языка, который был вынуждено исключен из лексикостатистического анализа. Словенский, который в своем нынешнем виде должен рассматриваться как член южнославянской подгруппы, имеет некоторые специфические изоглоссы с западнославянскими языками, прежде всего со словацким. Эти общие инновации характерны в первую очередь для северо-западных диалектов словенского, но некоторые из них распространяются дальше на юг, захватывая не только оставшуюся словенской область, но и кайкавские диалекты сербо-хорватского, а иногда даже и чакавские диалекты. Словенско-западнославянские изоглоссы относятся как к фонетике и морфологии (Greenberg 2000: 36, 40–42 с доп. лит., также Kurkina 1985; Bezlaj 2003), так и к базисной лексике (Kushniarevich et al. 2015, Fig. K, L, M in S2 File). Промежуточная позиция словенского между южной и западной подгруппами видна на филогенетической сети, рис. 2. Мало сомнений, что южнославянские языки (прежде всего словенский) и западнославянские языки (прежде всего чешско-словацкие диалекты) тесно контактировали на территории Паннонии до прибытия туда в X в. венгров. Происхождение словенского языка остается не до конца ясным. Возможны по меньшей мере три сценария. (1) Словенский – исконно южнославянский язык, близкородственный сербо-хорватскому, но попавший под сильное влияние западнославянских диалектов в конце 1-го тысячелетия н.э. (2) Словенский – исконно западнославянский язык, который сначала оказался отрезан от западнославянского центра переселением германцев и венгров в конце 1-го тысячелетия н.э., а затем неизбежно оказался под прессингом сербо-хорватских диалектов. (3) Словацкий или некоторые из современных словацких диалектов исторически происходят из южнославянской подгруппы. Лексикостатистический анализ высококачественных сводешевских списков северо-западных словенских диалектов мог бы прояснить ситуацию.

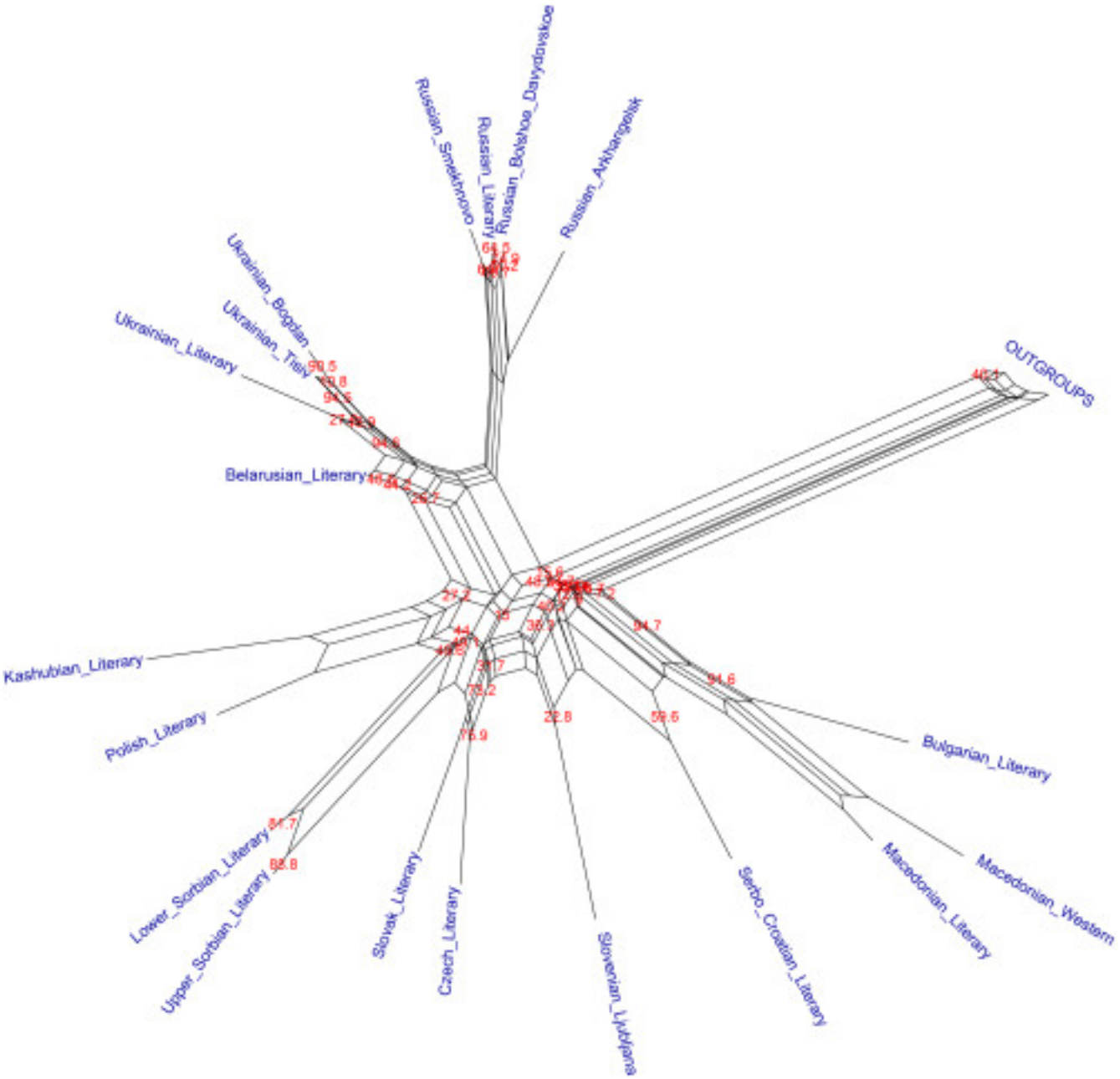

Рис. 2. Kushniarevich et al. 2015: сеть NeighborNet славянских языков, включая люблянское койне словенского. 110-словники. Значения бутстрэпа даны при узлах; не даны для устойчивых узлов со значением ≥ 95%. (Цит. по: Kushniarevich et al. 2015, Fig. M in S2 File)

В целом, можно сказать, что лингвистический анализ в Kushniarevich et al. 2015 показывает, как должна работать лексикостатистика: высококачественные входные данные, применение нескольких конкурирующих филогенетических алгоритмов, на выходе – устойчивое дерево, у которого топология и хронология соответствуют ранее установленным фактам и не противоречат традиционным взглядам специалистов.

В **Novotná & Blažek 2007a; Novotná & Blažek 2007b** приведено датированное дерево славянских языков на основе сводешевских 100-словников, собранных и проэтимологизированных авторами, а затем обсчитанных алгоритмом Starling NeighborJoining algorithm, рис. 3. Устойчивость дерева не ясна, статистическая поддержка узлов не приведена. Топология дерева не содержит каких-либо явных ошибок, однако узлы четко сдвинуты в сторону омоложения датировок, что противоречит историческим свидетельствам и нашим общим представлениям. В результате такого омоложения, например, исчезла лексикостатистическая разница между польским и кашубским (оба языка склеены в одном конечном узле). Причина таких неглубоких дат в том, что списки Новотной и Блажека этимологически-ориентированы и таким образом искусственно архаизированы, в то время как глоттохронологический модуль пакета Starling калиброван для более единообразных входных данных. Некоторые примеры из русского списка: в слоте ‘ashes’ стоит архаичное слово *пепел*, которое сейчас применяется в основном к пеплу сигарет или пеплу после кремации, а нейтральное русское слово для ‘ashes’ – это инновация *зола*. В слоте ‘belly’ стоят два синонима: архаичное *брюхо* и инновативное *живот*, на самом деле в сводешевском списке должно присутствовать, конечно же, только *живот*. Помимо всего прочего, такой этимологический подход противоречит

эксплицитному принципу, сформулированному Сводешем: “for each test item the common, everyday equivalent is listed for each language” (Swadesh 1955: 122).

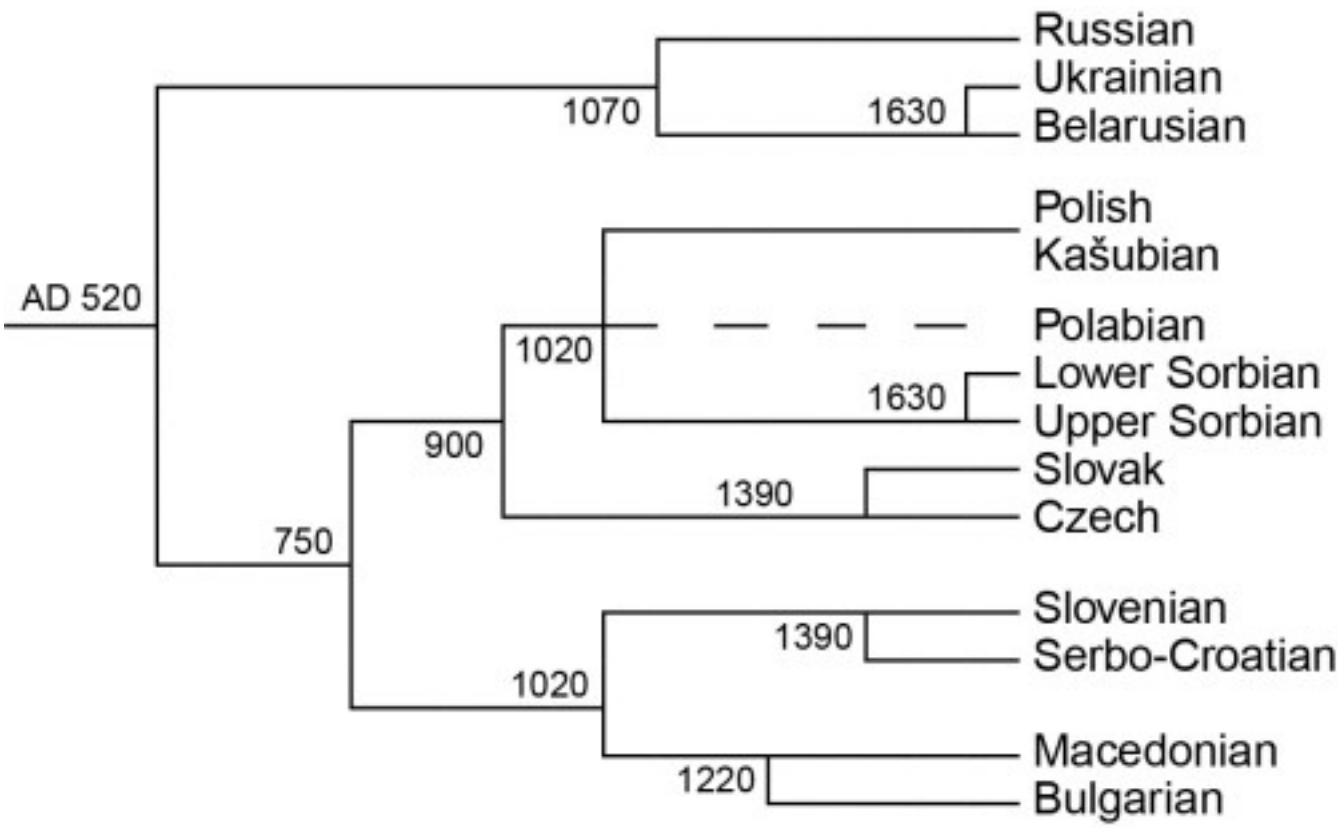

Рис. 3. Novotná & Blažek 2007a: датированное лексикостатистическое дерево славянских языков. 100-словники, алгоритм Starling NeighborJoining. Длина ветвей показывает абсолютную хронологию. (Цит. по: Novotná & Blažek 2007a: 201)

В **Dyen, Kruskal & Black 1992** дается классификация индоевропейской семьи на основе сводешевских 200-словников, собранных примерно для 85 индоевропейских языков, включая и славянские. Сами списки были составлены и проэтимологизированы Изидором Дайеном еще в 1960-е гг., в настоящее время они доступны в виде текстового файла “Comparative Indoeuropean Database Collected by Isidore Dyen”, размещенного на различных интернет-сайтах. Списки были обсчитаны алгоритмом UPGMA. В Dyen, Kruskal & Black 1992 не приводится дерево, вместо этого дается иерархически организованный список языков и публикуется некоторая сложная для восприятия диаграмма, на основе чего читателю предлагается самостоятельно представить себе древесную классификацию (кладограмму). Славянский фрагмент кладограммы см. на рис. 4. Как можно видеть, полученная филогения совершенно неудовлетворительна. Основная причина такого неубедительного результата – некачественные входные списки, содержащие значительное число лексикографических ошибок. Некоторые ошибки в славянской части базы Дайена разобраны в Kushniarevich et al. 2015 (S2 File, p. 27-29). Результат Dyen, Kruskal & Black 1992 контрастирует с правдоподобно выглядящим славянским деревом в Kushniarevich et al. 2015 (Fig. D in S2 File), полученным тем же алгоритмом UPGMA.

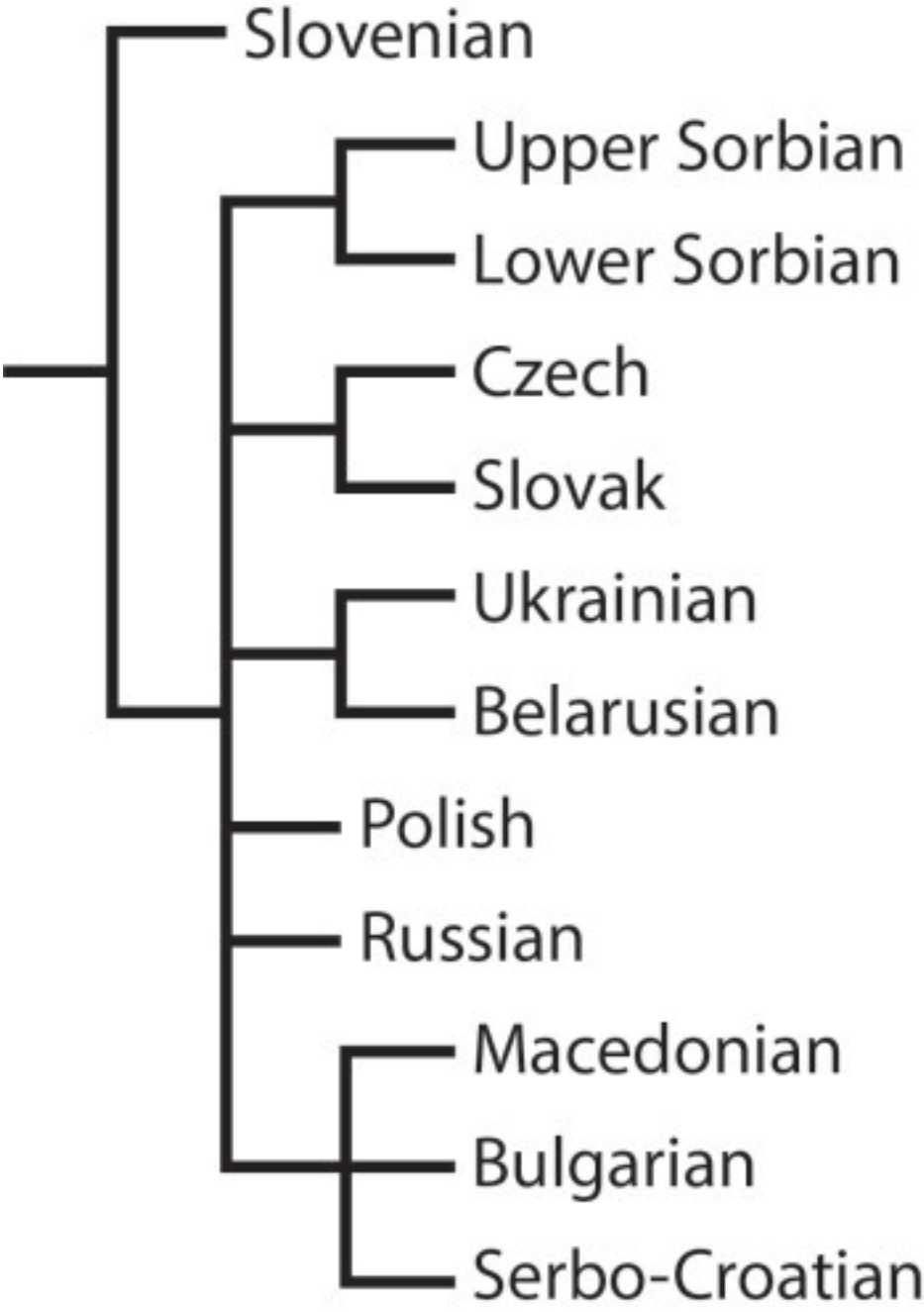

Рис. 4. Кладограмма славянских языков по Dyen, Kruskal & Black 1992: 88, 200-словники, UPGMA.

В **Gray & Atkinson 2003** и **Bouckaert et al. 2012** (с существенными уточнениями в 2013 г.: *Science* 342, p. 1446) приводятся индоевропейские деревья, полученные методом байесовского вывода.

Анализ в Gray & Atkinson 2003 строится на 200-словниках из базы Дайена. Славянская часть дерева, рис. 5, неудовлетворительна в плане топологии (восточнославянские помещены внутрь западнославянской клады), хронологии (700 г. н.э. – распад праславянского) и устойчивости (многие узлы имеют низкую статистическую поддержку).

Анализ в Bouckaert et al. 2012 построен на 207-словниках. Лексическая база данных “Indo-European Lexical Cognacy Database” (IELex) была подготовлена специально для данного филогенетического проекта. IELex – это исправленная и расширенная версия базы Дайена, туда были добавлены некоторые индоевропейские языки и расширен список источников (в частности были использованы анонимные сводешевские списки из *Wikipedia: the free encyclopedia*). По сравнению с базой Дайена, в EILex было исправлено какое-то количество лексикографических ошибок, но при этом некоторые новые ошибки были добавлены, см. обсуждение в Kushniarevich et al. 2015 (S2 File, p. 30-32). Славянское дерево в Bouckaert et al. 2012 лучшее соответствует нашим представлениям о славянских языках, чем дерево в Gray & Atkinson 2003, но тем не менее и оно содержит две явные ошибки: польский – восточнославянский язык, а словенский – аутлайер в южнославянской подгруппе.

Причина таких слабых результатов у двух данных экспериментов с байесовским выводом та же, что и в случае дерева из Dyen,

Kruskal & Black 1992, рассмотренного выше: низкое лексикографическое качество входных списков, например, многочисленные польские заимствования в украинском списке были размечены как исконные украинские формы, этимологически родственные соответствующим польским словам.

Статья Bouckaert et al. 2012 требует некоторых экстра-лингвистических комментариев. Статья вышла 24 августа 2012 г. (*Science* 337), представленное в ней индоевропейское дерево содержало слишком много явных ошибок, чтобы лингвисты могли принять такой результат. Далее, 20 декабря 2013 г. (*Science* 342: 1446, без DOI, без формы цитирования) этот научный коллектив сообщил, что в кодировании данных была выявлена значимая техническая ошибка, и представил уточненную версию индоевропейского дерева. Дерево-2013 и в самом деле выглядит более приемлемо, чем дерево-2012, однако, к сожалению авторы (вопреки своему утверждению) не проапдейтили некоторые файлы-приложения на сайте *Science*, в частности, не был размещено новый файл NEXUS, а в файле *.tre содержится только дерево-2012 без дерева-2013. Эти детали следует иметь в виду, ссылаясь на статью Bouckaert et al. 2012.

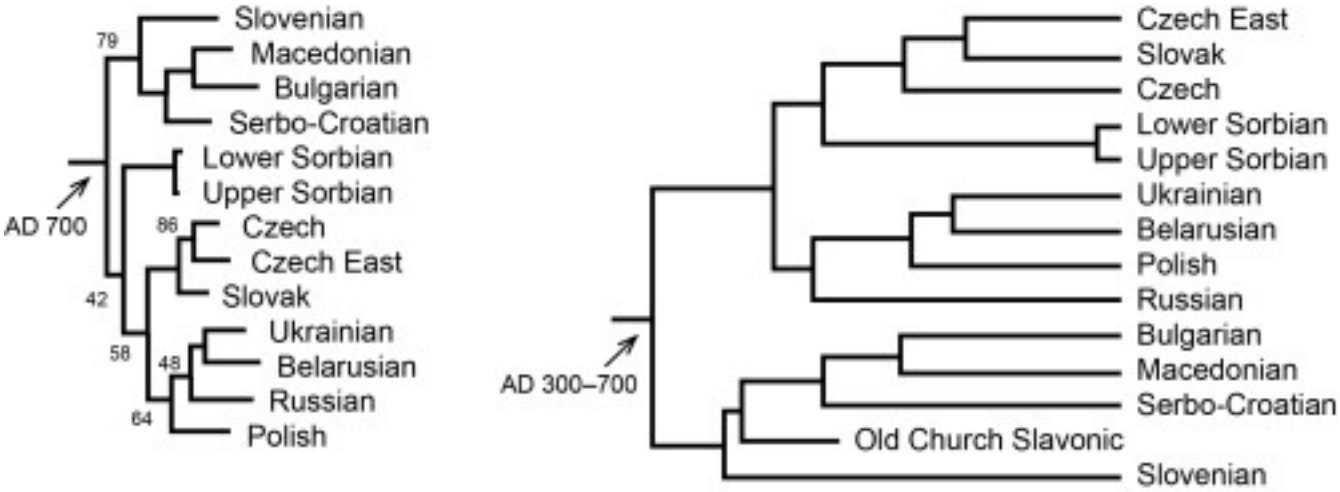

Рис. 5. Слева: славянская часть индоевропейского дерева из Gray & Atkinson 2003. 200-словник, байесовский вывод. Байесовские апостериорные вероятности даны курсивом рядом с узлами (не указаны для стабильных узлов с $P \geq 0.95$). Длина ветвей показывает скорость лексических замен. Right: славянская часть индоевропейского дерева из Bouckaert et al. 2012 (с апдейтом 2013 г., Fig. S1). 207-словники, байесовский вывод. Статистическая поддержка узлов не известна. Длина ветвей показывает абсолютную хронологию.

В **Chang et al. 2015** опубликовано индоевропейское дерево на основе тех же 207-словников из базы IELex, обработанных методом байесовского вывода, но с некоторыми дополнительными заранее заданными ограничениями по сравнению с анализом в Bouckaert et al. 2012. На первый взгляд, полученная славянская филогения, рис. 6, выглядит осмысленной: тройная структура группы [Южные [Восточные, Западные]] без перетасовки отдельных языков, например, польского. Однако, на самом деле каждая из трех славянских подгрупп была предварительно задана авторами как жесткая клада. Иными словами, сначала мы говорим компьютерной программе, что хотим получить на выходе, затем программа выдает желаемый результат.

Это иллюстрирует одну существенную с практической точки зрения вещь. Если вы видите статью с филогенией, полученной байесовским методом, следует помнить, что компьютерные пакеты для байесовского вывода (например, MrBayes или BEAST) имеют гибкие настройки, многие из которых могут напрямую повлиять на итоговое дерево. Необходимо внимательно ознакомиться с техническим разделом статьи и с приложениями к ней, чтобы понять, были ли заданы перед анализом какие-либо ограничения и какие элементы дерева – результат этих ограничений, а какие элементы действительно представляются собой результат исследования.

Рис. 6. Славянская часть датированного индоевропейского дерева из Chang et al. 2015. 207-словники, байесовский вывод. Четыре красные вертикальные полосы маркируют четыре предварительно заданные клады: славянскую, южнославянскую, западнославянскую и восточнославянскую. Все узлы имеют байесовскую апостериорную вероятность ≥ 0.95. Длина ветвей показывает абсолютную хронологию. (Цит. по: Chang et al. 2015, Fig. 2)

В **Vasilyev & Saenko 2020** представлено дерево славянских языков на основе 25 диалектных 110-словников, проэтимологизированных и затем обсчитанных методом Starling NeighborJoining. Устойчивость дерева не ясна, статистическая поддержка узлов не была приведена.

Главное новшество эксперимента – специально разработанная математическая процедура, по которой сначала рассчитывается среднее абсолютное отклонение в количестве общей лексики для каждого узла, а затем соседние перекрывающие друг друга бинарные узлы склеиваются в один общий небинарный узел (что приводит к политомии). Второе новшество – авторы принципиально отказались от включения в анализ литературных языков, потому что литературные и стандартизированные языки имеют свойство искусственно консервировать лексику.

Итоговое дерево, рис. 7, не представляется убедительным. Можно предположить, что источник проблем лежит в следующем. Васильев и Саенко (2020) правы, что литературная норма часто имеет архаизированный словарь, не во всем соответствующий живому словоупотреблению, что приводит к омоложению дат на дереве и потенциально к топологическим искажениям. Однако славянские диалектные словари, используемые в Vasilyev & Saenko 2020, страдают от недостатков другого рода. Во-первых, некоторые из этих словарей дифференциальные (охватывают только ту лексику, которая отличается от литературной нормы), что делает поиск дефолтных слов для сводешевских концептов в таком словаре затруднительным. Во-вторых, что более важно, значительная часть словарей, использованных авторами, составлена местными языковыми энтузиастами, которые желают описать и сохранить свой родной говор. Часто подразумеваемый принцип таких лексикографов-носителей – это показать, что описываемый диалект значимо отличается от литературного языка, чем больше находится различий, тем лучше. В частности, такая установка приводит к тому, что в таких описаниях словарь может быть перекошен в пользу особенных локальных слов и выражений. Подобный подход оказывается очень полезен для диалектологических, этимологических и этнографических исследований, но вреден для лексикостатистики, потому что вместо базовых слов для сводешевских концептов ошибочно могут быть отобраны редкие и маргинальные слова. Ошибки такого рода ведут наоборот к удревнению дат на дереве. Этот эффект, видимо, и привел к появлению веера на дереве Васильева и Саенко.

Рис. 7. Лексикостатистическое дерево славянских языков Васильева и Саенко. 110-словники, метод Starling NeighborJoining. Длина ветвей показывает относительную хронологию. (Цит. по: Vasilyev & Saenko 2020: 340)

Проект **Automated Similarity Judgment Program (ASJP)** (Wichmann, Holman & Brown 2022) представляет собой совершенно особый подход. Вместо традиционного этимологического анализа (как во всех экспериментах, описанных выше), ASJP полагается на фонетическое сходство между сравниваемыми словами. Авторы собрали 40-словники для огромного количества языков по всему миру, около 8 000 списков. Словоформы в списках единообразно транскрибированы, между ними измерены дистанции Левенштейна. Получаемая таким образом матрица расстояний между языками обсчитывается алгоритмом NeighborJoining, и строится дерево (устойчивость дерева не ясна, статистическая поддержка узлов не приводится). ASJP – это продолжающийся проект, куда добавляются новые языки, а уже имеющиеся списки могут обновляться и уточняться.

Славянская часть дерева ASJP (дерево v.5 от 2021 г., Müller et al. 2021), рис. 8, неудовлетворительна. Сложно установить без дополнительных тестов, что именно вызывает такие сильные искажения: лексикографические и транскрипционные ошибки во входных списках или особенности алгоритма, или же что-либо еще. Действительно, хотя ошибки в русском списке, отмеченные в Kushniarevich et al. 2015 (S2 File, p. 33-34) для базы данных ASJP v.16, были исправлены в нынешней v.20, но остальная часть славянского раздела базы ASJP вряд ли подверглась сплошной проверке. С другой стороны, не исключено, что специфический алгоритм ASJP сам по себе недостаточно чувствителен для реконструкции филогении таких близкородственных и контактирующих друг с другом языков, как славянские (ср. схожие проблемы, например, с германскими языками на дереве ASJP).

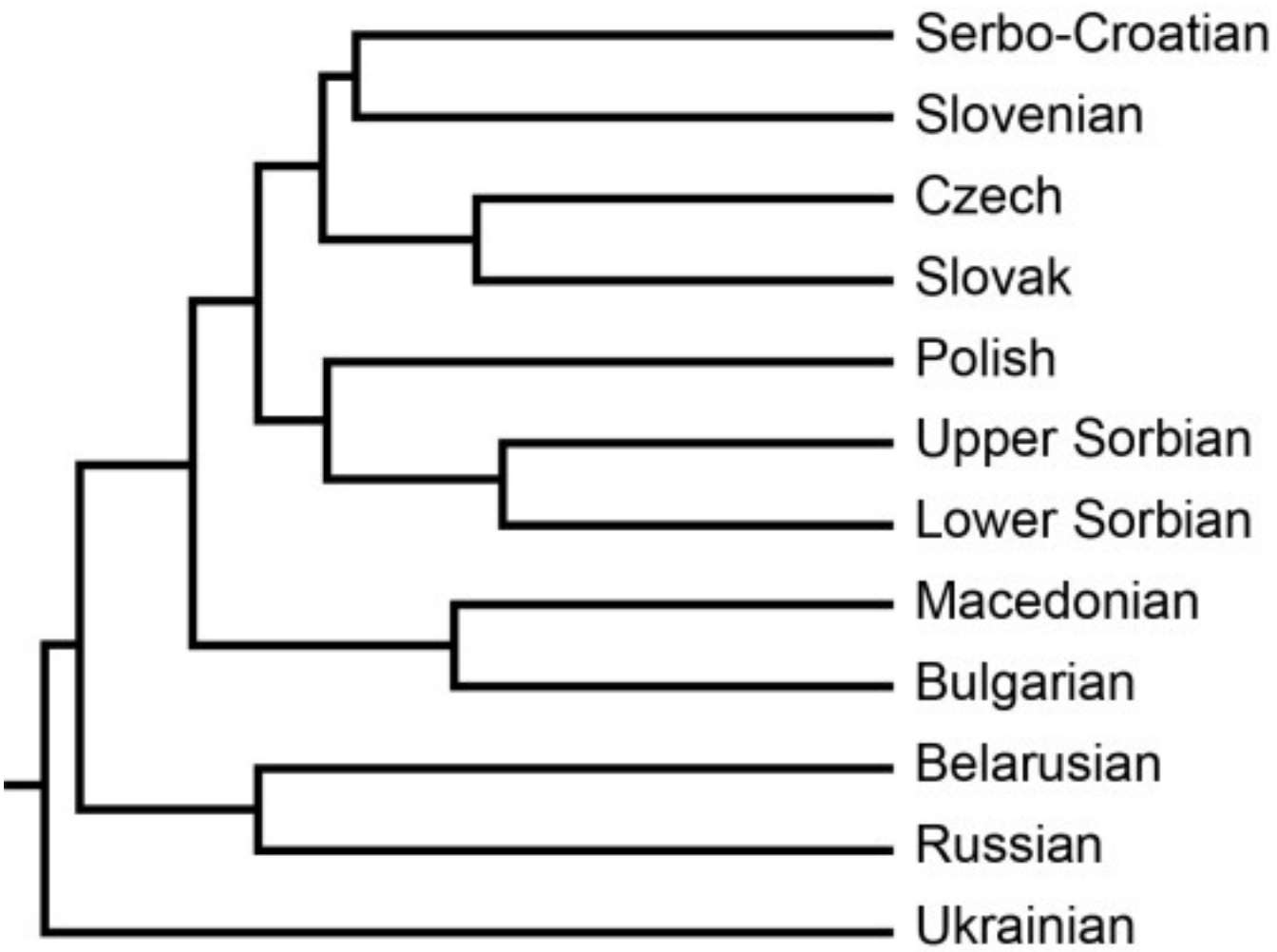

Рис. 8. Лексикостатистическое дерево славянских языков проекта ASJP. Неэтимологизированные унифицировано транскрибированные 40-словники, дистанции Левенштейна, алгоритм NeighborJoining. Длина ветвей показывает фонетическое расхождение. (Цит. по: Müller et al. 2021, в упрощенной форме, т.е. опущены несколько языков, не существенных для настоящего обзора: искусственный язык словио, дублирующие паразитические списки, напр. Russian-2, микро-языки нашта и нинильчик)

References

Bezlaj, France. 2003. Položaj slovenščine v okviru slovanskih jezikov. *Zbrani jezikoslovni spisi*, vol. 1, 268–277. Ljubljana: Založba ZRC.

Blažek, Václav. 2020. Classification of Slavic languages: evolution of developmental models. *Slavia Occidentalis* 77(1). 33–64. doi:10.14746/so.2020.77.3.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard & Q. D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337. 957–960. doi:10.1126/science.1219669.

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. doi:10.1353/lan.2015.0005.

Dyen, Isidore, Joseph B. Kruskal & Paul Black. 1992. *An Indoeuropean classification: a lexicostatistical experiment* (Transactions of the American Philosophical Society, New Series 82(5)). Independence Square, Philadelphia: The American Philosophical Society.

Embleton, Sheila. 2000. Lexicostatistics/Glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In Colin

Renfrew, April McMahon & Larry Trask (eds.), *Time depth in historical linguistics*, vol. 2, 143–166. Cambridge, England: McDonald Institute for Archaeological Research.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426. 435–439. doi:10.1038/nature02029.

Greenberg, Marc L. 2000. *A historical phonology of the Slovene language* (Historical Phonology of the Slavic Languages 13). Heidelberg: Universitätsverlag C. Winter.

Heggarty, Paul, Warren Maguire & April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559). 3829–3843. doi:10.1098/rstb.2010.0099.

Huson, Daniel H. & Celine Scornavacca. 2011. A survey of combinatorial methods for phylogenetic networks. *Genome Biology and Evolution* 3. 23–35. doi:10.1093/gbe/evq077.

Jacques, Guillaume & Johann-Mattis List. 2019. Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). *Journal of Historical Linguistics* 9(1). 128–167. doi:10.1075/jhl.17008.mat.

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

Kassian, Alexei S., George Starostin, Anna Dybo & Vasily Chernov. 2010. The Swadesh wordlist. An attempt at semantic specification. *Journal of Language Relationship* 4. 46–89.

Kurkina, Lyubov V. 1985. Praslavjanskie dialektnye istoki južnoslavjanskoj jazykovoj gruppy [Proto-Slavic dialect origin of the South Slavic language group]. *Voprosy yazykoznaniya* 4. 61–71.

Kushniarevich, Alena, Olga Utevska, Marina Chuhryaeva, Anastasia Agdzhoyan, Khadizhat Dibirova, Ingrida Uktveryte, Märt Möls, et al. 2015. Genetic heritage of the Balto-Slavic speaking populations: a synthesis of autosomal, mitochondrial and Y-chromosomal data. *PLOS ONE* 10(9). 1–19. doi:10.1371/journal.pone.0135820.

List, Johann Mattis, Annika Tjuka, Mathilda Van Zantwijk, Frederic Blum, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill & Robert Forkel. 2023. CLLD Concepticon 3.1.0. Zenodo. doi:10.5281/ZENODO.7777629. https://concepticon.clld.org/ (23 April, 2023).

McMahon, April M. S. & Robert McMahon. 2005. *Language classification by numbers* (Oxford Linguistics). Oxford NY: Oxford University Press.

Müller, André, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Pamela Brown, et al. 2021. ASJP world language trees of lexical similarity: Version 5 (October 2021). https://asjp.clld.org/download.

Novotná, Petra & Václav Blažek. 2007a. Glottochronology and its application to the Balto-Slavic languages. *Baltistica* 42(2). 185–210.

Novotná, Petra & Václav Blažek. 2007b. Glottochronology and its application to the Balto-Slavic languages. *Baltistica* 42(3). 323–346.

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

Starostin, Sergei A. 2000. Comparative-historical linguistics and lexicostatistics. In Colin Renfrew, April McMahon & Larry Trask (eds.), *Time depth in historical linguistics*, vol. 2, 223–265. Cambridge, England: The McDonald Institute for Archaeological Research.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96. 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* (21). 121–137.

Tischler, Johann. 1973. *Glottochronologie und Lexikostatistik* (Innsbrucker Beiträge Zur Sprachwissenschaft). Innsbruck: Inst. f. Sprachwiss. d. Univ. Innsbruck.

Vasilyev, Mikhail & Mikhail Saenko. 2020. Analiz topologii i ocenka točnosti leksikostatističeskix klassifikacij (na primere

slavjanskix jazykov) [An analysis of the topology and estimation of accuracy for lexicostatistical classifications (on the data of Slavic languages)]. *Journal of Language Relationship* 18(4). 320–347.

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2022. The ASJP Database (version 20). https://asjp.clld.org/ (23 February, 2023).

Yanovich, Igor. 2020. Phylogenetic linguistic evidence and the Dene-Yeniseian homeland. *Diachronica* 37(3). 410–446. doi:10.1075/dia.17038.yan.