В. Д. Соловьев
Казанский государственный университет

Типологическая схожесть языков как метод изучения языковой эволюции[1]

Существующие типологические классификации языков опираются на один или небольшое число признаков. С появлением больших типологических баз данных, таких как The World Atlas of Language Structures или «Языки мира», становится возможным строить классификации, учитывающие сотни и тысячи признаков одновременно. Для построения классификаций могут быть применены различные математические алгоритмы, что создает предпосылки создания объективной классификации. В статье приводятся первые результаты в этом направлении. К числу наиболее интересных неожиданных результатов можно отнести типологическое различие языков Африки и Евразии, а также типологическую близость чукотско-камчатских языков с языками синокавказской макросемьи. Предложено несколько (близких) вариантов классификации языков Северной Африки, Европы и Северной и Центральной Азии. Обсуждается возможность использования этого подхода для установления сверхдальнего родства языков, основанная на двух соображениях. Во-первых, близость одновременно по сотням признаков статистически не может быть результатом случайного совпадения. Во-вторых, по крайней мере, некоторые грамматические свойства являются очень стабильными и несут информацию о языковых состояниях древности.

Keywords: типологические классификации, типологические базы данных, математические алгоритмы кластеризации, гипотезы сверхдальнего родства

1. Введение

В любой научной дисциплине классифицируются изучаемые объекты. Хорошо известна таксономическая классификация всех живых существ в биологии. Предмет лингвистики — языки также могут быть классифицированы, причем разными способами. Наиболее продвинутой является генеалогическая классификация языков. Другой подход к классификации — по степени похожести структуры языков вне зависимости от их происхождения — менее развит. Видимо, первой серьезной попыткой классификации языков по структуре является морфологическая классификация, построенная во второй половине 19-го века в работах Шлегеля, Гумбольдта и Шлейхера [Кузнецов 1954]. Эта классификация сохраняет свое значение и поныне, однако, она учитывает лишь один аспект языковой структуры — способ соединения морфем, так что языки, попавшие по этой классификации в один класс, могут радикально отличаться друг от друга в других аспектах.

[1] Работа выполнена при поддержке РФФИ, грант № 10-06-00087а.

Современная типология акцентирует внимание на корреляциях логически независимых типологических признаков и выделяет пучки таких признаков, которые также могут послужить основой для классификации языков. Детально изученный пример — порядок глагола (V) и прямого дополнения (O). Этот параметр может принимать разные значения: VO, OV, свободный порядок. Оказалось, что порядок V и O строго коррелирует с наличием в языке предлогов или послелогов и целым рядом других параметров, связанных с порядком слов и морфем [Comrie 1989]. Но и в этом случае, порядок глагола и прямого дополнения не является доминирующим, определяющим все или хотя бы многие свойства языка.

Остается неясным, возможна ли холистическая классификация языков, разбивающая все языки на несколько непересекающихся групп, так, что внутри групп языки типологически однородны, а между языками разных групп имеются существенные типологические различия, причем, по широкому спектру признаков, охватывающему все основные языковые уровни. Такая классификация будет невозможна, если окажется, что языки равномерно заполняют все пространство возможных комбинаций типологических свойств, так что любые границы между группами языков будут чисто условны. Априори наиболее вероятным представляется наличие нескольких центров сгущения языков, между которыми располагаются языки с промежуточными свойствами. Группы языков вокруг центров сгущения, если таковые будут найдены, и образуют типологические типы в холистической классификации.

К попытке начать развивать такой подход можно отнести предложение Трубецкого [Трубецкой 1987] трактовать индоевропейские языки, как ставшие родственными в результате конвергенции, когда они приобрели 6 общих признаков: отсутствие сингармонизма, консонантизм начала слова не беднее консонантизма середины и конца слова, наличие приставок, наличие аблаутных чередований гласных, наличие чередования согласных в грамматических формах (sandhi), аккузативность. При последовательном развитии этого подхода можно было бы попытаться найти аналогичные описания и для других семей. Так тюркские языки можно было бы охарактеризовать как агглютинативные, с финальным положением глагола, сингармонизмом и какими-то еще свойствами.

Этот подход оказался забыт по ряду причин. Прежде всего, отвергается базовая идея Трубецкого о причинах родственности индоевропейских языков. Далее, были обнаружены языки, которые обладают всеми 6 вышеперечисленными свойствами, но не относятся к индоевропейской семье. Наконец, в рамках подхода, связывающего группу языков с набором характеристических свойств, просто не было получено каких-либо значимых результатов, которые бы привлекли внимание широкого круга лингвистов.

Важный момент, на который следует обратить внимание, состоит в следующем. До последнего времени у лингвистов фактически не было возможности сравнивать языки одновременно по десяткам и тем более сотням и тысячам различных грамматических свойств. Человеческий мозг не способен учитывать одновременно такое число абстрактных переменных. Ситуация радикально изменилась с появлением больших типологических баз данных, таких как The World Atlas of Language Structures [WALS] и «Языки мира» [Поляков & Соловьев 2006], а также адекватных средств обработки больших массивов грамматических данных.

На повестку дня вновь может быть поставлен вопрос о холистической классификации языков в вышеописанном смысле, причем классификация языков может быть осуществлена на основе одновременного учета сотен и более грамматических свойств с применением объективных математических и компьютерных методов анализа.

Если таким способом будет получено некоторое разбиение языков на классы типологически близких языков, то на повестку дня встанет вопрос: чем объясняется эта бли-

зость — общностью происхождения или сближением структуры языков за счет заимствований при длительных ареальных контактах. Теоретически схожесть может оказаться и случайной, однако, если речь идет о сотнях и тысячах признаков, то крайне маловероятно, чтобы столько параметров (большинство из которых независимы друг от друга) приняли близкие значения.

Если типологически близкие языки находятся в далеко отстоящих друг от друга регионах и отсутствуют какие-либо данные (археологические, генетические) об их контактах в древности, то заимствования между ними также будут маловероятны, что позволит предположить общность происхождения. Таким образом, изучение типологической близости языков с помощью типологических баз данных и математических алгоритмов может послужить дополнительным методом в изучении языкового родства.

Вероятно, пионерскими работами в этом русле является статьи Б. Комри и М. Сисоу [Comrie & Cysouw 2006; Cysouw & Comrie 2009]. В данной статье будут приведены первые результаты, полученные в этом направлении.

2. Данные и методы

В настоящее время существует целый ряд типологических баз данных, современный обзор приведен в [Everaert & Musgrave 2009]. Все они, кроме двух, являются специализированными, т. е. посвящены определенным аспектам структуры языков. Лишь две базы данных — WALS и «Языки мира» — охватывают все основные разделы грамматики.

WALS — завершенный в 2005 г. крупный международный проект, выполненный под руководством М. Хаспельмата, Б. Комри и др. Он включает компьютерную базу данных, доступную на CD-дисках и через Интернет, а также бумажное издание. WALS содержит данные по 2560 языкам (из всех семей) и 142 признакам, из которых 128 грамматические. Каждый признак может принимать от 2 до 9 значений, в среднем, примерно, 5. Некоторым недостатком WALS является то, что не все языки описаны по всем признакам, более того, в среднем язык в WALS имеет всего около 45 признаков.

Две отличительные особенности придают проекту большую значимость. Во-первых, для каждого признака построена карта Земного шара, на которой кружочками разного цвета обозначены языки с различными значениями выбранного признака. Хотя идея графического изображения географического распределения признаков предлагалась и ранее, но впервые она была реализована столь масштабно. Во-вторых, база данных снабжена чрезвычайно удобным интерфейсом, позволяющим легко ориентироваться в огромном массиве информации — WALS включает огромное число справочных статей по различным признакам и языкам. Поисковые средства позволяют находить нужную информацию по комбинации признаков, генерировать новые карты с заданными свойствами.

База данных «Языки мира» создана в Институте языкознания РАН по материалам одноименной серии монографий, доступна в Интернете по адресу www.dblang2008.narod.ru и подробно описана в [Поляков & Соловьев 2006].

Она содержат описания 315 языков Евразии по 3821 признаку, относящихся к следующим сферам языка: фонетика, морфология, синтаксис. Все признаки бинарные. Бинарность представления означает, что для каждого языка и для каждого признака в базе данных фиксируется только наличие или отсутствие этого признака в языке, но не степень его проявления. Таким образом, с математической точки зрения БД представляет собой прямоугольную бинарную матрицу размером 315×3821, содержащую более миллиона бит информации. Разделяемый WALS и «Языки мира» общий принцип — рав-

ноправие всех признаков, т. е. признаки не делятся на более и менее важные, не вводится никаких весовых коэффициентов.

В базе данных «Языки мира» представлены следующие языковые семьи и языковые сообщества: австроазиатские, австронезийские, алтайские, афразийские, индоевропейские, кавказские, палеоазиатские, синотибетские, уральские, хуррито-урартские языки, а также языки изоляты: айнский, нивхский, бурушаски, шумерский, эламский. Языки Европы, Северной и Центральной Азии описаны очень полно. Австроазиатские, австронезийские, синотибетские группы представлены только отдельными языками.

Программная оболочка позволяет осуществлять не только поиск, но и другие содержательные операции. Для проведения квантитативных типологических исследований важной является операция сравнения, позволяющая по двум выбранным языкам найти все признаки (и подсчитать их число), по которым эти языки совпадают и по которым они отличаются.

Число признаков, по которым два языка различаются, называется расстоянием между ними (метрика Хемминга или городская метрика). Ясно, что чем расстояние меньше, тем языки типологически ближе друг к другу. Например, расстояние в базе данных «Языки мира» между белорусским и македонским языками (оба славянские) равно 243, а между белорусским и бирманским — 401.

Полученные числовые данные могут быть использованы для выделения кластеров типологически близких языков. Для этих целей применяются методы кластерного анализа, наиболее популярными из которых являются филогенетические алгоритмы. Они активно используются в биологии при построении систематики живых организмов. Пионером в области применения филогенетических алгоритмов в лингвистике является, видимо, Т. Варноу, начавшая их использовать еще в 1997 г. [Warnow 1997].

Алгоритм NeighborNet, применяемый в данной работе, описан в [Bryant et al. 2005]. Он строит звездообразные структуры, в которых классифицируемые объекты располагаются по окружности так, что языки с меньшим расстоянием располагаются ближе друг к другу. Существуют и другие подходы, один из которых будет затронут в разделе 4.

3. Типологическая классификация языков с применением филогенетических алгоритмов

3.1. Данные WALS.

В работе [Comrie & Cysouw 2006] изучается типологического разнообразия языков Новой Гвинеи. Выбрано 48 наиболее полно описанных языков этого региона, представляющих практически все традиционно выделяемые языковые группы Новой Гвинее. Для сравнения их с языками всего мира следующим образом выбрано 47 языков. Сначала выбрано 150 наиболее полно описанных в WALS языков, не более чем по одному из каждой языковой группы, сопоставимой с романскими языками по возрасту. Затем из них случайным образом выделено 47. Далее к выбранным 95 языкам применен алгоритм NeighborNet. Результат приведен на рис. 1. Жирным шрифтом выделены языки Новой Гвинеи.

Хорошо видно, что языки Новой Гвинеи не образуют компактную группу, а рассредоточены практически по всему языковому пространству. Языки Новой Гвинеи Maybrat, Abun, Arapesh близки к африканским языкам Hausa, Sango и Luvale соответственно, язык Wahgi — к языку американских индейцев Yaqui, язык Imonda — к языку австралийских аборигенов Kayardild, языки Hua, Awtuw — к азиатским языкам Burmese и Bu-

rushaski соответственно и так далее. Другими словами, типологическое разнообразие языков Новой Гвинеи вполне сопоставимо с типологическим разнообразием языков всего мира.

Дает ли диаграмма рис. 1 какую-либо информацию о возможной холистической типологической классификации языков? Ясно видно деление всех языков на 2 группы: левая часть диаграммы (от Zulu до Guarani) и правая часть (от Marind до Gooniyand), между которыми располагается большой зазор, указывающий на большую разницу в типологических свойствах языков этих групп. Легко видеть, что языки левой части диаграммы характеризуются порядком VO, а правой — OV. В [Comrie & Cysouw 2006] авторы обращают внимание на это деление, указывая, однако, что это вполне может быть и артефакт описания. И действительно, в WALS непропорционально много признаков посвящено порядку слов: 17 из 142 признаков, т. е. 12%. Это могло послужить причиной такого деления, т. к. порядок глагола и прямого дополнения, как уже отмечалось, коррелирует с другими признаками, связанными с порядком слов.

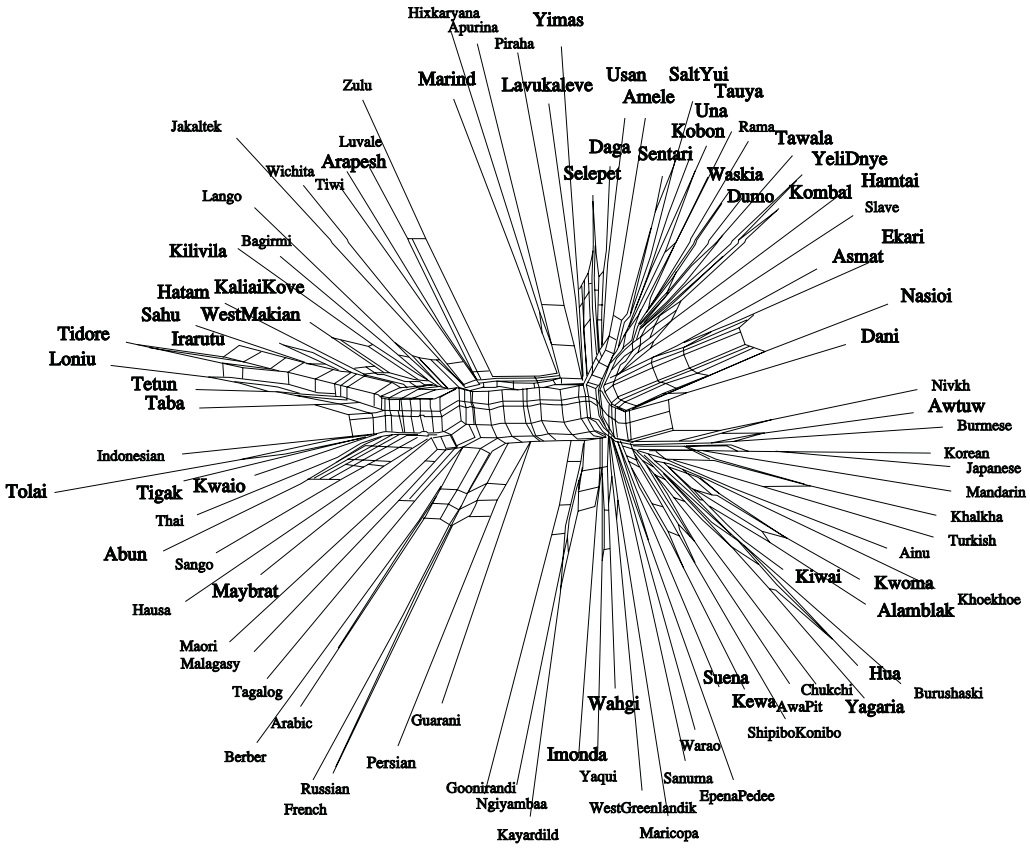

Рис.1. *Языки Новой Гвинеи vs. языки остального мира (цит. по: Comrie, Cysouw 2006)*

В работе [Cysouw & Comrie 2009] сравнивались языки Африки и Евразии. С помощью алгоритма NeighborNet построена аналогичного вида диаграмма, приведенная на рис. 2.

На ней представлено по одному языку из большинства языковых групп (уровня ветви индоевропейской семьи) этих континентов, включая изоляты, как самостоятельные группы. Здесь ситуация совсем иная — языки Африки (выделены жирным шрифтом), кроме Khoekhoe, группируются отдельно, языки Евразии отдельно. Таким образом, существуют явные типологические различия на континентальном уровне. Причины этого пока не вполне ясны. Но одно из очевидных возможных объяснений состоит в том, что в пределах континентов действуют механизмы конвергенции за счет заимствований.

На этой диаграмме также отражается деление по признаку порядка слов. Африканские языки делятся на 2 группы: справа от Beja до Harar Oromo языки имеют порядок слов VO, вверху от Middle Atlas Berber до Supyire — порядок OV (с несколькими исключениями). Языки Евразии образуют два больших кластера: внизу от Hindi до Persian имеют порядок VO (с несколькими исключениями), слева от Turkish до Mundari — OV. Brahui и Hungarian располагаются посередине между этими группами.

В целом же авторы этих статей отмечают, что представленные данные не выявляют четко выраженных кластеров языков, так что остается большая свобода исследователя в построении типологической классификации.

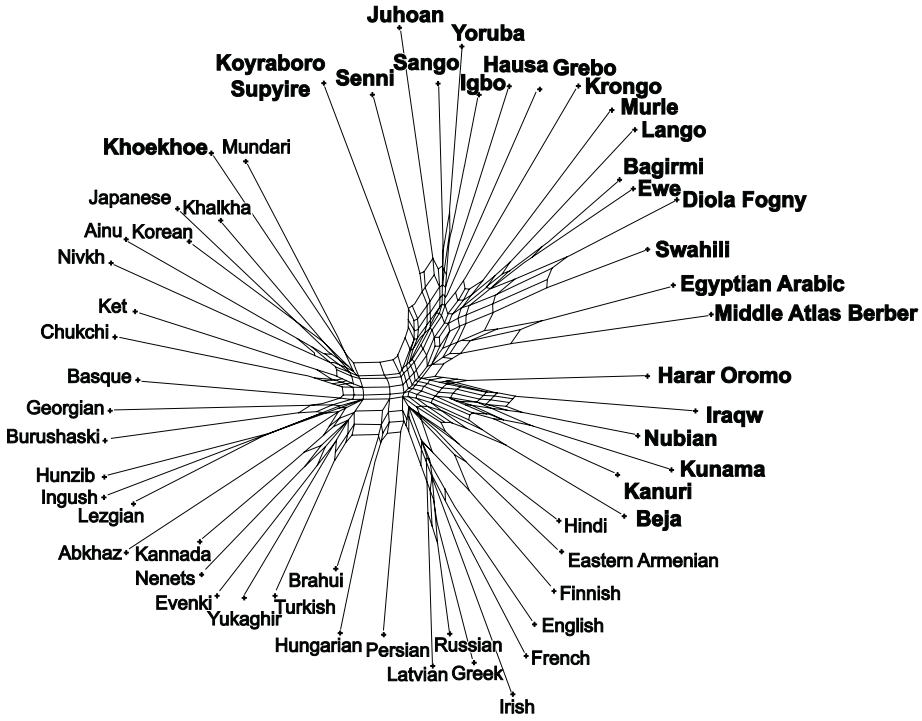

Рис. 2. *Языки Африки и Евразии (цит. по: Cysouw, Comrie 2009, p. 194)*

3.2. Данные «Языки мира».

Для оценки влияния порядка глагола и прямого дополнения мы перепроверим эти результаты с помощью базы данных «Языки мира». Основная идея состоит в применении к тем же языкам того же алгоритма и сравнение результатов. К сожалению, база данных «Языки мира» не содержит описания столь многих, как WALS, языков Африки. В ней на настоящий момент присутствуют лишь афразийские языки. Зато большинство используемых на диаграмме 2 языков Евразии в базе данных «Языки мира» присутствует. В нескольких случаях отсутствующие языки заменены на близкородственные: Eastern Armenian на Armenian, Egyptian Arabic на Modern Arabic, Middle Atlas Berber на Zenaga, Hunsib на Avar, Khalkha на Halh Mongolian, Mundari на Khmer. Результат применения алгоритма кластеризации показан на диаграмме 3.

Опишем основные структурные особенности диаграммы рис. 3.

1. На этой диаграмме расположение языков более равномерное, чем на предыдущих; нет никакого четкого разделения на кластеры. Это можно рассматривать как косвенный аргумент в пользу гипотезы моногенеза и/или подтверждение значительного вклада заимствований в уменьшение расстояний между группами языков, расходящимися в ходе представляемой в древовидной форме эволюции.

2. Нет никакого зазора между языками с порядками слов VO и OV. Таким образом, это деление, прослеживающееся на предыдущих диаграммах, является, как отмечалось выше, скорее всего, результатом непропорционально большого числа признаков в WALS, связанных с порядком слов. Более сбалансированный набор признаков базы данных «Языки мира» не обнаруживает этой дихотомии (см. рис. 3).

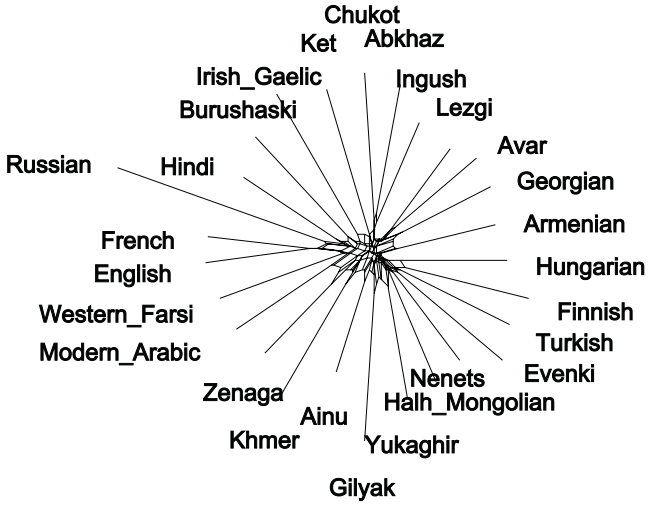

Рис. 3. *Языки Африки и Евразии по данным серии «Языки мира»*

3. Некоторые языки размещаются на диаграммах этого типа очень странным образом. Например, на рис. 3 ирландский язык (кельтская ветвь индоевропейской семьи) размещен между кетским и бурушаски. Эти языки не являются в действительности типологически близкими, они не являются родственными и располагаются географически очень далеко друг от друга. Вероятной причиной такого размещения является недостаточность данного метода графического представления информации для совершенно точного отображения типологической близости между всеми парами языков. С математической точки зрения база данных «Языки мира» дает представление языков точками в 3821-мерном пространстве признаков. В то же время диаграмма, которую строит NeighborNet, эквивалентна 1-мерному представлению (по окружности). Естественно, что при сворачивании 3821-мерного пространства в 1-мерное могут иметь место искажения.

4. Африканские языки арабский и зенага (афразийская семья) на диаграмме, разумеется, располагаются рядом друг с другом. Но так как данных слишком мало — всего 2 языка с африканского континента, то трудно говорить о типологическом разделении языков Африки и Евразии на основе этих данных.

5. Проанализируем возможное группирование других языков. Слева четкую группу образуют русский, французский и английский языки, разделяющие как общее происхождение в составе индоевропейской семьи, так и общее географическое положение на европейском континенте. В [Comrie & Cysouw 2006] Б. Комри и М. Сисоу также отмечают типологическую близость русского и французского, выявляемую на диаграмме рис. 1.

Рядом с этими языками располагаются хинди и западный фарси (персидский), что отражает их принадлежность к индоевропейской семье. При этом хинди и персидский находятся на периферии группы индоевропейских языков, что соответствует их более раннему отделению по сравнению с европейской ветвью [Gray & Atkinson 2003] и более далекому географическому положению. Интересно, что хинди и персидский расположены не рядом друг с другом, как можно было бы ожидать на основании их родства в рамках индоиранской ветви, а по разные стороны от европейской части индоевропей-

ской семьи. Точно такое же расположение зафиксировано и на диаграмме рис. 2 по данным WALS. Это указывает на то, что индийские и иранские языки действительно имеют значительные типологические различия, большие, чем славянские, германские и романские. В любом случае, не учитывая ошибочно расположенного ирландского языка, слева на диаграмме языки сгруппировались по общему происхождению.

Справа-внизу располагаются шесть языков, относящихся к алтайской и уральской семьям. Языки этих семей действительно имеют много общих типологических черт. Гипотеза о родственности этих языков поддержана, в частности, в работах [Starostin 2007; Грунтов и др. 2006]. Противники этой гипотезы считают, что похожие черты языков этих семей могут быть следствием ранних (на стадии протоязыков) заимствований. В настоящее время не представляется возможным сказать, какая из этих двух гипотез верна. Можно считать, что эти языки сгруппированы по ареально-генетическому принципу.

Очень интересным является группирование шести языков справа-вверху от абхазского до армянского. Все они располагаются на Кавказе, относясь к 4 различным семьям: индоевропейской, картвельской, нахско-дагестанской и абхазо-адыгской. Таким образом, здесь явно имеет место ареальное группирование. Можно отметить, что грузинский и армянский находятся на одном конце этой группы ближе к урало-алтайским языкам, что коррелирует с их генеалогической общностью в соответствии с ностратической гипотезой [Дыбо 1978]. В то время как абхазский и ингушский языки находятся на другом конце, ближе к кетскому, в соответствии с гипотезой о генеалогической близости северокавказских и енисейских языков [Старостин 2007].

Два языка внизу: гиляк (нивхский) и айну являются изолятами. Их генетические связи не установлены, но географически они близки друг к другу — расположены на Дальнем Востоке. Между ними и урало-алтайским филумом (но ближе к первым) расположен юкагирский, который в некоторых работах трактуется как изолят [Николаева & Хелимский 1997], в некоторых относится к уральской семье [Nichols 1992]. Он также располагается на Дальнем Востоке. Т. е. это размещение языков имеет видимые ареальные основы. В этой же части диаграммы находится и кхмерский язык (австроазиатская семья), не имеющий явных ареальных или родственных связей с другими языками на диаграмме. Но так как алгоритм все равно должен был его куда-то поместить, то это место рядом с другими азиатскими изолятами кажется наиболее удачным.

Наконец, в верхней части диаграммы расположено еще несколько изолятов — бурушаски, кетский и чукотский. Бурушаски расположен рядом с хинди, к которому он и географически близок, так что здесь возможны контактные влияния. В [Эдельман 1997] отмечены значительные заимствования в бурушаски из индоиранских языков. Кетский язык расположен близко к группе северокавказских языков, что коррелирует с уже упомянутой гипотезой о дальнем родстве синокавказских и енисейских языков в пределах бореальной макросемьи.

Чукотский язык сам по себе не изолят, а входит в небольшую чукотско-камчатскую семью, но генеалогические связи этой семьи не установлены. Так что чукотский язык не имеет установленных родственных связей ни с каким из языков, представленных на данной диаграмме. Трудно себе представить и ареальные контакты между чукотско-камчатскими и северокавказскими языками, с которыми он оказался рядом на рис. 3, в обозримом прошлом.

Между тем чукотско-камчатские языки действительно имеют ряд общих типологических черт с енисейскими и северокавказскими языками, что отмечалось ранее. Это двусторонняя агглютинация, наличие категории рода, полиперсональный глагол и некоторые другие [Володин 1997]. Обнаруженная типологическая близость чукотского языка с абхазским и кетским может служить основание для выдвижения гипотезы о

сверхдальнем родстве чукотско-камчатской семьи и бореальной макросемьи. Отметим, что на диаграмме рис. 2 по данным WALS чукотский язык расположен рядом с кетским, что дает другой аргумент в пользу этой гипотезы.

В итоге, выделяются, хотя и не столь явно, 5 основных кластеров языков по типологической близости: индоевропейский, урало-алтайский, кавказский (возможно, с чукотским и кетским), дальневосточный (несколько изолятов) и афразийский.

Посмотрим, выделяются ли найденные типологические группы и на диаграмме рис. 2 по данным WALS. На ней также четко видны афразийский (справа, в составе большого африканского кластера), индоевропейский (внизу-справа), кавказский (слева) и дальневосточный (вверху-слева) кластеры. В составе последнего присутствуют также японский и корейский языки (их нет в базе данных «Языки мира»). Неожиданно алтайские и уральские языки не образовали одной группы, оказались разбросаны по разным частям диаграммы. Например, типологически близкие Turkish и Khalkha (алтайская семья) оказались на разных концах диаграммы. Это является проявлением недостатков алгоритма классификации. На диаграмме рис.1 эти языки оказались рядом.

В некоторых случаях, вероятно, на классификации сказываются и определенные проблемы самой WALS. В работе [Polyakov et al. 2009] проводилось сопоставление аналогичных классификаций на основе WALS и «Языки мира» и показано, что классификация на основе «Языки мира» точнее отражает ареально-генеалогические связи. В [Polyakov et al. 2009] высказано мнение, что данные WALS сильно зашумлены, в первую очередь, пробелами в данных. Все же, в первом приближении, обнаруженная типологическая классификация подтверждается данными WALS.

4. Алгоритм ординации

Как отмечалось выше, алгоритм NeighborNet строит, по сути, одномерное представление близости языков. Другие математические средства позволяют создавать иные представления, в том числе 2-мерные, располагая языки на плоскости так, что типологически близкие языки размещаются близко друг к другу. В числе этих средств алгоритм ординации [R-Project]. В [Поляков & Соловьев 2006] построена ординация для множества языков из нижеприведенного списка (по техническим причинам отсутствуют языки с номерами 1, 25 и 33) с использованием базы данных «Языки мира».

2	венгерский	14	бурятский	27	эстонский	40	испанский
3	финский	15	азербайджанский	28	македонский	41	итальянский
4	ассамский	16	вепсский	29	немецкий	42	галисийский
5	дари	17	хантыйский	30	бенгальский	43	абхазский
6	ительменский	18	турецкий	31	румынский	44	белорусский
7	португальский	19	бирманский	32	лезгинский	45	болгарский
8	грузинский	20	армянский	34	корякский	46	датский
9	бурушаски	21	багвалинский	35	персидский	47	нивхский
10	аккадский	22	агульский	36	таджикский	48	шугнанский
11	норвежский	23	могольский	37	чукотский	49	польский
12	английский	24	калмыцкий	38	туркменский		
13	исландский	26	монгорский	39	татарский		

Результаты приведены на рис. 4. Линии разделяют возможные кластеры. Слева располагаются европейские языки. Внизу — иранские (в эту область попали также бурушаски и аккадский). В центре-вверху — плотная группа из алтайских, уральских, кавказских и некоторых индоиранских языков. Это образование можно назвать центральноазиатским кластером. В нем языки нескольких семей сближаются, очевидно, вследствие ареальных контактов. Маленькую группу образуют три языка с номерами 34, 37 и 43 — корякский, чукотский и абхазский. Наконец, языки №19 и 47 (бирманский и нивхский) — изоляты в данной выборке.

Как видим, метод ординации дает несколько иное, укрупненное деление на кластеры типологически близких языков. Иранские языки, хотя и выделены на рис. 4 в отдельный кластер, но он располагается рядом с кластером европейских языков, так что здесь вполне правомерно объединение их в единый кластер индоевропейских языков. Вновь абхазский язык сближается с языками чукотско-камчатской семьи, что дает очередное подтверждение гипотезы об их сверхдальнем родстве.

Полученная методом ординации классификация коррелирует, но не совпадает с приведенными в предыдущем параграфе.

5. Заключение

Целью статьи является не столько представление конкретных результатов, которые еще далеки от окончательных, сколько очертить общий путь, двигаясь по которому можно получить новые типологические классификации языков. На этом пути должны быть решены многие методологические проблемы. Было показано, что математические методы и алгоритмы не являются идеальными, выдаваемые ими результаты являются лишь приближенными. В этом направлении в последние годы ведется интенсивная работа: в [Wichmann & Saunders 2007] проведено сопоставление различных филогенетических алгоритмов, в [Поляков & Соловьев 2006] сравнивались различные меры близости, в [Polyakov et al. 2009] — различные источники данных.

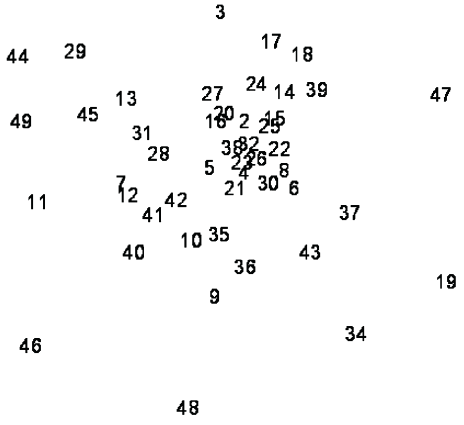

Рис. 4. *Ординация тестового множества языков*

Несмотря на указанные трудности, даже первоначальные результаты внушают оптимизм. Полученные разными методами типологические классификации коррелируют между собой. Типологическая близость во всех упоминавшихся в статье кластерах может быть объяснена общностью происхождения или ареальными контактами. В первом приближении на пространстве Северной Африки, Европы, Северной и Центральной

Азии можно выделить следующие основные кластеры типологически близких языков: индоевропейский, урало-алтайский, кавказский, дальневосточный и афразийский. Некоторые из полученных результатов вполне ожидаемы, как например, выделение в особую группу индоевропейских языков. Насколько эта классификация удачна — покажут будущие исследования (см. рис. 4).

На этом пути могут быть получены и новые неожиданные результаты, например, типологическая близость чукотско-камчатской семьи с синокавказской макросемьей. Эта близость может быть следствием сверхдальнего родства (на временной глубине порядка 15 тыс. лет до н. э.), не устанавливаемого классическим сравнительно-историческим методом. Типологические черты языков, по крайней мере, некоторые, являются более стабильными, чем лексемы [Nichols 2007], что позволяет выявлять сверхдальнее языковое родство или древние языковые контакты.

В целом все направление с использованием больших типологических баз данных и математических методов анализа данных представляется чрезвычайно перспективным и может привести к формированию новых разделов типологии и получению общезначимых результатов.

А. В. Дыбо
РГГУ, Москва

К сожалению, следует констатировать, что в статье В. Д. Соловьева с самого начала имеется логическая ошибка. Эта же ошибка, конечно, имеется и у Н. С. Трубецкого, которого призывает вспомнить автор. По определению, родственные языки — это, как известно, языки, восходящие к общему языку-предку. Вот примерные основания этого определения:

Между звуковой оболочкой слова и его значением нет необходимой связи, т. е. в общем случае минимальный знак языка (морфема) является символом (в пирсовском смысле). Тогда однотипное сходство внешних оболочек при относительном тождестве значений большого количества знаков между разными языками — т. е. гипотеза о случайном совпадении здесь неприемлема — означает, что эти знаки в разных языках суть отображения одного и того же набора знаков. Если эти знаки относятся к «малопроницаемым» подсистемам языка, т. е. гипотеза о массовом заимствовании неприемлема, то следует считать, что рассматриваемые языки являются отображениями одного и того же языка. Этот прототипический язык может быть построен как язык-посредник, содержащий глубинные формы, из которых по непротиворечивому набору правил могут быть построены формы представленных языков. Построенный таким образом глубинный язык (реально обычно строятся какие-то фрагменты языковой системы) называется праязыком данной группы языков, а группа называется родственными языками. В частности, доказательство родства некоторого множества языков фактически заключается в построении представительного фрагмента праязыковой системы (и, соответственно, в предъявлении этого фрагмента и набора правил вывода поверхностных языковых форм). Интерпретация такой ситуации — историческая, т. е. генетически родственными языками называются те два или несколько языков, которые представляют собой результаты различных линий эволюции одного и того же языка, существовавшего раньше и тем самым являющегося для этих языков праязыком.

Таким образом, языки по определению не могут стать родственными в результате конвергентного развития. То сходство, которое может между ними образоваться в последнем случае, нельзя называть языковым родством во избежание терминологической путаницы. Некоторые авторы называют его сродством. Т. е., если бы Н. С. Трубецкой дал себе труд как-то формализовать свою задачу, он должен был бы написать, что собирается показать, что отношения между индоевропейскими языками являются не родством, а сродством.

Теоретически возможна, конечно, и такая модель развития, при которой носители нескольких (абсолютно неизвестных нам) языков в результате тесных контактов между собой развили некоторый общий язык, не являющийся по определению генетическим потомком какого-либо из тех, прежних языков, или же являющийся генетическим потомком нескольких из них, из которого в дальнейшем дивергентным путем развились языки-потомки — индоевропейские языки. Трудностью здесь является то, что на практике специалистам по сравнительно-историческому языкознанию еще не довелось столкнуться с чистыми случаями того, чтобы принципиально нельзя было определить, генетическим потомком какого из контактировавших языков является данный язык, и отнести сходства с остальными языками к заимствованиям. Т. е., в надежных случаях никогда не удается построить для одной группы родственных языков несколько разных, но равно представительных праязыковых подсистем. Отмечу, что проблема генеалогической принадлежности креольских языков и пиджинов неоднократно дискутировалась между нами с В. И. Беликовым, и последнему пока не удалось привести пример языка «смешанного происхождения», для которого с помощью сравнительно-исторического анализа нельзя было бы определить принадлежность к определенной языковой семье.

Типологические сходства по определению (см. выше) не являются доказательными для языкового родства. Они могут показывать либо сродство языков, образовавшееся в результате контактов, либо сохранившиеся от праязыкового состояния сходства, либо — третья возможность, которая, как мне кажется, недостаточно учитывается автором статьи, — сходства, обусловленные структурной близостью человеческих языков, чисто типологического характера (пример: левое ветвление побуждает к отсутствию префиксов, или что-нибудь подобное), или — четвертая возможность, также недостаточно учитываемая автором, — случайно возникать в разных языках в силу совершенно различных процессов (пример из области материальных сходств — новогреческое и полинезийское название глаза, греч. *мати*, полинез. *мата*). Четвертая возможность для типологических сходств чисто арифметически значительно более вероятна, чем для материальных (которые учитываются при доказательстве родства). В самом деле, типологические признаки обычно принимают небольшое число значений. Возьмем бинарный (левое-правое ветвление) или кватернарный (эргативный — номинативный — …) признак. У каждого языка есть только одна из двух или одна из четырех возможностей, вероятность случайного попадания 0,5 или 0,25. Если же взять возможность совпадения названия глаза из четырех букв, то легко подсчитать, что при принятии алфавита, допустим, из 40 букв, вероятность случайного совпадения между двумя языками будет 0,00000390625. Сравнительный же метод требует значительного числа таких совпадений.

База данных «Языков мира», по которой ведется подсчет, состоит из 3821 бинарного признака. Чтобы получить вероятность совпадения, равную случаю с *одним* четырехбуквенным «глазом», нужно совпадение около 18 бинарных признаков. Если для доказательства родства требовать совпадения 600 четырехбуквенных слов с заданными значениями (что, в общем, соответствует обычной практике сравнительно-исторического языкознания), — и именно для таких случаев считать совпадение неслучайным, то бинарных признаков потребуется около 18×600 = 10800 признаков. Впрочем, и такое совпадение не будет демонстрировать родства — по определению, а покажет только неслучайность сходства.

Читателю предлагается оценить степень случайности типологических расстояний между любыми двумя выбранными языками по базе «Языков мира» (македонским и белорусским — 245, т. е. совпадают 3576 признаков или бирманским и белорусским — 401, т. е. совпадают 3420 признаков).

Из свойств диаграммы на рис. 3 выше, первое —

«1. На этой диаграмме расположение языков более равномерное, чем на предыдущих; нет никакого четкого разделения на кластеры. Это можно рассматривать как косвенный аргумент в пользу гипотезы моногенеза и/или подтверждение значительного вклада заимствований в уменьшение расстояний между группами языков, расходящимися в ходе представляемой в древовидной форме эволюции».

Конечно, ничего про моногенез или заимствования, как я уже объясняла, это свойство не говорит, а говорит о том, что подобранные для описания языков бинарные типологические признаки обладают высокой степенью взаимной независимости, что хорошо говорит о составителях схемы описания в «Языках мира».

Вторая особенность и ее причины автором, видимо, проинтерпретированы правильно.

Дальнейшее описание кластеризации по алгоритму NeighborNet показывает в основном свойства этого алгоритма. То же, в, общем, можно сказать о плоскостном представлении близости языков.

В заключении работы говорится: «Даже первоначальные результаты внушают оптимизм. Полу-

ченные разными методами типологические классификации коррелируют между собой. Типологическая близость во всех упоминавшихся в статье кластерах может быть объяснена общностью происхождения или ареальными контактами».

У меня как у компаративиста приведенные примеры классификаций ни в малой степени не вызывают оптимизма (заметим на полях, что, например, армянский язык не попал в индоевропейский кластер ни по какой классификации, а ведь это даже не хеттский; разброд наблюдается и среди относительно близких генетически индоиранских языков), а типологическая близость в кластерах может быть объяснена еще и случайностью, как упоминалось выше.

Добавлю, что утверждение «Типологические черты языков, по крайней мере, некоторые, являются более стабильными, чем лексемы» основано на работе Дж. Николс, результаты которой крайне сомнительны из-за особенностей выборки анализируемых языков и чрезвычайно узкой признаковой базы сопоставления.

Итак, по-моему, пока количество типологических признаков для классификации языков мира, по неслучайности равносильной генетической, явно недостаточно, а то, что полученная на достаточном количестве признаков классификация совпадет или как-то будет скоррелирована с генетической классификацией, данное предварительное исследование не показывает.

Г. С. Старостин
РГГУ, Москва

Статья В. Д. Соловьева посвящена достаточно «горячей» теме в современной лингвистике — значимости (потенциальной) сравнительных данных, накопленных в закромах лингвистической типологии, для решения спорных вопросов в области генеалогической классификации языков. Специально подчеркиваю слово «спорный», хотя в самом тексте статьи оно и не упоминается, потому что там, где с задачей классификации успешно и непротиворечиво справляется классическая компаративистика, никакие альтернативные методы, по сути, не требуются, что имплицитно признает и автор статьи: так, отмечаемый им факт «странного» типологического сближения ирландского языка с кетским и бурушаски на рис. 3 не предлагается (слава богу!) интерпретировать как серьезный аргумент в пользу ирландско-кетского родства, а лишь как свидетельство определенного несовершенства компьютерного алгоритма (положим, что так).

Работа по сближению типологии и компаративистики ведется уже давно; начало ее можно, по-видимому, отсчитывать со ставших классическими (вне зависимости от того, заслуженно или нет) работ Дж. Николс 1990-х гг., в последнее время же построение «историко-типологических» компьютерных моделей переживает своеобразный бум — так, совсем недавно на эту же тему была опубликована очередная статья Р. Грея [Greenhill et al. 2010].

Объяснить такую неожиданную популярность можно тремя факторами. Во-первых, это устойчивое представление о том, что сравнительный метод как основной способ установления генетического родства языков «изжил» себя как парадигма, по крайней мере там, где речь касается временной глубины, превышающей 6—8 тысяч лет (концепция, активно пропагандируемая, в том числе, и Дж. Николс), и вызванное этим в чем-то, пожалуй, даже благородное желание помочь «застрявшим» компаративистам, неспособным своими силами проникнуть в столь глубокое прошлое. Во-вторых, это стремление к структурированию и осмыслению самого массива типологических данных, в том числе и с точки зрения тех исторических связей, которые его образуют. Наконец, в-третьих, это своего рода синкретическая тенденция применить к лингвистической таксономии те же методы, которые давно и успешно применяются в других областях, прежде всего — в биологии.

Безусловно, хотелось бы надеяться, что со временем методика анализа данных, описываемая как в дискуссионной работе, так и в других аналогичных исследованиях, будет усовершенствована и станет реальным подспорьем для компаративиста. К сожалению, на данном этапе можно только констатировать факт абсолютной практической неприменимости полученных результатов, несмотря на несколько натянутый оптимизм, демонстрируемый автором.

Зададимся простым вопросом: действительно ли полученные результаты можно считать, хотя бы в каком-то одном отношении, *полезными* для разрешения спорных вопросов сравнительно-исторического языкознания? Диаграмма, приведенная на рисунке 3, схематически суммирует данные по целому ряду языков Евразии, потенциально связанных между собой дальним родством (в рамках таких гипотез, как ностратическая, сино-кавказская и ряд других), причем здесь возможны два типа ситуаций — (а) члены одной макросемьи, отделившись друг от друга, продолжают существовать на географически смежных территориях, (б) оказываются в ситуации отсутствия каких-либо контактов в результате дальней миграции одной или нескольких составляющих первоначального этноса. Какие *новые* выводы относительно этих двух типов ситуаций позволяет нам сделать диаграмма?

По ситуациям типа (а) — очевидно, никаких. Алгоритм «грамотно» располагает в непосредственной близости друг от друга языки уральской и алтайской семей, но, как отмечает автор, диаграмма позволяет утверждать лишь то, что «эти языки сгруппированы по ареально-генетическому принципу». Но что такое «ареально-генетический принцип», как не признание того факта, что отмеченные типологические сходства могут объясняться либо контактами, либо общим происхождением, либо и тем и другим, т. е., по сути, не добавляют к нашим текущим представлениям об урало-алтайских языках ничего нового?

Тот факт, что абхазский и нахско-дагестанские языки располагаются рядом друг с другом, что могло бы косвенно подтверждать идею общесеверокавказского родства, полностью обесценивается тем, что неподалеку от них рядом друг с другом расположены языки хинди и бурушаски. Последнее объясняется контактными влияниями индоарийских (а также дардских) языков на бурушаски, но что в таком случае мешает нам считать и абхазо-дагестанское «соседство» результатом интенсивных языковых контактов?

По ситуациям типа (б) результаты оказываются лишь маргинально более ценными. *Допустим*, что типологическая близость языков, устанавливаемая по данному алгоритму, не является продуктом случайного развития (момент сам по себе чрезвычайно спорный, но здесь, по крайней мере, можно легко понять, как усовершенствовать алгоритм за счет подключения еще большего количества бинарных или многозначных признаков), а всегда отражает либо конвергенцию, либо родство, либо и то, и другое. В этом случае факт расположения кетского языка в непосредственном соседстве от северокавказских, казалось бы, должен был порадовать любого сторонника сино-кавказской гипотезы; поскольку енисейские языки расположены вне зоны возможных контактов с кавказскими языками, речь может идти только об опции родства.

Однако что делать с оказавшимися на том же конце диаграммы чукотско-камчатскими языками? Автор говорит о возможности «сверхдальнего родства чукотско-камчатской семьи и бореальной макросемьи», но это довольно расплывчатое замечание, т. к. в гипер-гипотетическую «бореальную макросемью» в том виде, в котором она сегодня представляется в рамках Московской школы компаративистики, по сути, входят все без исключения языки, представленные на диаграмме, что совершенно не объясняет особую близость чукотского языка именно к кавказско-енисейской ее части. Может быть, соседство чукотского с кетским следует понимать, наоборот, как ареальное влияние енисейских языков, ранее имевших более широкое распространение, на чукотско-камчатскую семью? Но в таком случае почему и енисейские языки, в свою очередь, не могли подвергнуться влиянию какого-нибудь древнего «кавказоидного» субстрата, повлиявшего на их типологическую структуру?

Отметим, наконец, что ситуации типа (б) на представленных диаграммах вообще возникают *крайне редко*. Как правило, рядом друг с другом предпочитают располагаться языки одного и того же ареала, что особенно наглядно видно на рис. 2, где языки Африки, несмотря на признаваемое всеми специалистами значительное генетическое разнообразие этого региона, четко противопоставлены языкам Евразии (диаграмма по алгоритму Б. Комри и др., но, скорее всего, схожие результаты были бы получены и автором статьи при условии подключения африканских данных). Нет ни малейшего сомнения в том, что, если бы к работе алгоритма были подключены данные по Юго-Восточной Азии, сино-тибетские языки встали бы рядом с австроазиатскими и тайскими, т. к. по своим типологическим характеристикам они почти неотличимы (хотя к генетическому происхождению этих языков они и не имеют никакого отношения).

Не подлежит сомнению, что диаграммы, полученные по методике как автора работы, так и Б. Комри и других исследователей, в целом неслучайны и имеют право на сущностную интерпретацию: они представляют собой способ построения формальной, объективной типологической классификации языков, являясь значительным шагом

вперед по сравнению с «ручными», серьезно ограниченными классификациями типологов предыдущего поколения. Но для того, чтобы иметь возможность спроецировать их в историческую плоскость, необходимо опираться в первую очередь на такие типологические признаки, которые являются в языке наиболее устойчивыми, т. е. чрезвычайно редко (в идеале — никогда не) меняют свое значение на противоположное под ареальным влиянием окружающих их языковых семей.

Факт самого существования таких признаков, однако, не бесспорен. В статье он обозначен лишь в паре мест, и то скорее вскользь — в резюме говорится, что «некоторые грамматические свойства являются очень стабильными и несут информацию о языковых состояниях древности», а в конце утверждается, что «типологические черты языков, по крайней мере, некоторые, являются более стабильными, чем лексемы». Никаких конкретных примеров таких типологических черт, однако, не приводится; вместо этого читатель отсылается к презентации Дж. Николс на соответствующую тему, но и там было бы тщетно обнаружить такие примеры, тем более что, по утверждению самой Дж. Николс, «typology… for purposes of discriminating genealogical from other relatedness… has weak resolution at all time depths». Вместо «типологических» черт она предпочитает делать упор на «полу-типологические» («semi-typological characters») — формального определения для этого понятия не предлагается, но, судя по приводимым примерам (личные показатели в алгонкинских языках, маркеры рода и числа в афразийских языках, местоименная парадигма n : m в «американских языках»), речь идет об элементарном распределении фонетически сходных грамматических морфов в языках мира, что, строго говоря, уже не имеет непосредственного отношения к типологии как таковой (отсюда и оговорка «semi-typological», на наш взгляд, не очень успешная), а представляет собой, по сути, разновидность гринберговского метода массового сравнения, ограниченного грамматическим материалом.

Поиск «устойчивых» типологических признаков — задача чрезвычайно интересная (хотя плодотворное ее разрешение сомнительно), но как раз про нее в статье ничего не говорится; меж тем, до тех пор, пока точного определения таких признаков не существует, генетическая интерпретация получаемых типологических схем неправомочна. Даже «рекомендательное» использование их как потенциальных индикаторов родства, судя по представленным результатам, даст примерно сопоставимое количество истинных и ложных «наводок», и в этом смысле практическая ценность предлагаемой методики близка к нулю.

Думается, что, при условии как дальнейшего усовершенствования самого алгоритма анализа, так и расширения типологической базы данных и привлечения материала максимально большого числа языков, можно будет создать прочную основу для построения единой формальной *ареально-типологической* классификации языков мира — задачи в целом не менее важной, чем построения классификации генетической. Но остается совершенно непонятным, для чего вообще настаивать на специальной значимости данной работы для выяснения генетических связей между языками.

Возможно, в каком-то смысле это объясняется неправильно проведенной аналогией между лингвистикой и биологией; настоящим лингвистическим аналогом классификационных признаков в биологии все-таки должны быть не абстрактные структурные особенности того или иного языка, а его морфемный инвентарь (или хотя бы то, что у Дж. Николс называется «полутипологическими» признаками). Только на таком материале представляется возможным получить сколь-либо существенные сведения относительно дальнего родства между теми или иными языковыми семьями.

В. Д. Соловьев
Казанский государственный университет

Прежде всего, следует отметить, что упоминание Н. С. Трубецкого в статье — не более чем историческая реминисценция. Оно, по сути, несущественно и может быть полностью удалено из статьи без какого-либо ущерба для понимания. Видимо, я неудачно представил материал, необоснованно сместив акцент на работу Трубецкого. Родство в статье понимается, конечно же, в классическом

смысле. Ни определение понятия «родство», ни справедливость древовидной модели эволюции языков не являются предметов данной статьи. Главный тезис статьи: созданные несколько лет назад большие типологические базы данных могут быть полезны и для установления родства.

Разумеется, исследования в этом направлении находятся в самом начале, и именно сейчас дискуссия, организованная редакцией журнала, является очень полезной. Далее я остановлюсь на двух моментах: методологии исследований и причинах, побудивших использовать типологические (как синоним будет использоваться слово 'грамматические') данные в компаративистике.

В дискуссии выявились четыре методологические проблемы предлагаемого подхода. Это: объем имеющихся типологических данных, стабильность грамматических признаков, влияние заимствований, корректность используемых филогенетических алгоритмов.

Общая идея оценки объемов лексических и грамматических данных, предложенная А. В. Дыбо, конечно, правильна. Однако приведенные в ее реплике подсчеты очень приблизительны и могут быть уточнены. Вероятность случайного совпадения двух 4-буквенных слов в двух языках оценивается ей как 0,00000390625, исходя из полного перебора всех комбинаций букв. При этом учитываются такие «слова», как *шщйь* и даже *ъъъъ*. Вряд ли в каком-либо языке мира используются такие комбинации букв для обозначения какого-либо понятия. Ясно, что следует подсчитывать не все вообще сочетания букв, а лишь приемлемые в человеческих языках. Сделать это не просто, но отмечу, что, если взять русский язык, то получится следующая картина. Общее число комбинаций из 4 букв в 33-буквенном алфавите — 1185921, в то же время в максимально полном 144-тысячном словаре русского языка, используемом в компьютерной программе Aspell (http://ru.wikipedia.org/wiki/GNU_Aspell), зафиксировано только 1987 четырехбуквенных слова, т. е. примерно в 600 раз меньше числа всех комбинаций.

Далее, А. В. Дыбо в своих подсчетах ориентируется на сопоставление 600 слов в сравниваемых языках. Однако многие компаративистские исследования оперируют с 200-, 100-, 40- (проект [ASJP]) и даже 35-словными (Яхонтов) списками, и полученные на этом пути результаты не отбрасываются как заведомо неверные. Если учесть эти соображения, то окажется, что уже объем грамматических данных, накопленных в базе данных «Языки мира», значительно превосходит объем лексических данных. Это гарантирует невозможность случайной близости языков.

Следующий вопрос — о стабильности типологических признаков. Это ключевой вопрос развиваемого подхода. Данную точку зрения разделяют и участники дискуссии. Г. С. Старостин пишет: «Поиск "устойчивых" типологических признаков — задача чрезвычайно интересная (хотя плодотворное ее разрешение сомнительно), но как раз про нее в статье ничего не говорится; меж тем, до тех пор, пока точного определения таких признаков не существует, генетическая интерпретация получаемых типологических схем неправомочна».

Данная область развивается чрезвычайно быстро и даже за то краткое время, пока моя статья лежала в редакции журнала, появилось несколько важных работ по стабильности типологических признаков [Greenhill et al. 2010; Wichmann & Holman 2009; Соловьев & Фасхутдинов 2009], которые нельзя оставить без внимания и которые в значительной степени проясняют ситуацию.

Важными являются два вопроса: какова скорость изменения грамматических признаков по сравнению с лексическими и какие из грамматических признаков более стабильны. Отметим, во-первых, упомянутую Г. С. Старостиным работу Р. Грея [Greenhill et al. 2010], в которой показано, что скорость изменения типологических признаков примерно такая же, как и лексических. Независимая оценка темпов эволюции типологических признаков приведена в монографии С. Вихмана и Е. Холмана [Wichmann & Holman 2009], в которой суммированы результаты предшествующих работ этих авторов, а также Николс, Камхолца и др. Если в 100-словном списке Сводеша за 1000 лет меняется в среднем 14 слов, то из признаков в базе данных WALS за то же время меняется в среднем 19, т. е. скорость изменения типологических признаков, хотя и несколько выше, чем лексических, но сопоставима с последней. Любопытно, что если исключить изменения признаков в результате заимствований, то результаты вновь оказываются близкими. Согласно С. А. Старостину [Старостин 1989], число исконных (не обусловленных заимствованиями) замен в 100-словном списке равно 5 за 1000 лет, согласно же [Wichmann & Holman 2009], исконных типологических замен — 7. Уже эта близость значений показывает, что использование грамматических признаков уместно.

Казалось бы, можно сделать вывод, что использование типологических черт не дает преимуществ по сравнению с лексическими. Однако следует иметь в виду, что для лексического уровня языка

стабильность подсчитывалась для ядра лексики — наиболее стабильных слов, обозначающих важнейшие концепты человеческой культуры, в то время как для грамматики были взяты все признаки, включенные в базу данных, а не только самые стабильные. Таким образом, напрашивается следующий шаг — выделение ядра грамматики — наиболее стабильных признаков.

В литературе дано несколько строгих определений стабильности грамматических признаков, позволяющих оценить стабильность количественно. Наиболее проработанными являются независимо предложенные подходы Е. Масловой [Маслова 2004] и С. Вихмана, Е. Холмана [Wichmann & Holman 2009]. В работе [Wichmann & Holman 2009] приведен список высокостабильных признаков из WALS (с количественной оценкой стабильности). Возникает естественный вопрос, является ли метод определения стабильности из [Wichmann & Holman 2009] корректным.

Мною в статье [Соловьев & Фасхутдинов 2009] на материале нашей базы данных проведено сопоставление четырех различных методов: Е. Масловой, С. Вихмана и Е. Холмана, одного метода, восходящего к идеям Николс, и предложенного мной оригинального метода оценки стабильности на основе аппроксимации с помощью филогенетического алгоритма Maximal Parsimony [Semple & Steel 2003] числа изменений значений грамматического признака в ходе эволюции. Оказалось, что все 4 метода подсчета дают коррелирующие результаты. В работах [Соловьев & Фасхутдинов 2009; Polyakov et al. 2009] показано, что стабильность, подсчитанная по данным «Языки мира», коррелирует со стабильностью, подсчитанной по данным WALS, а также хорошо согласуется и с результатами, опубликованными ранее в типологической литературе.

Например, к высокостабильным все методы подсчета отнесли такие признаки как: наличие рода, наличие предлогов, оппозиция инклюзив/эксклюзив и др. Таким образом, в настоящее время уже имеются четкие процедуры количественной оценки стабильности признаков, выделена группа высокостабильных признаков. Особо хочу подчеркнуть, что все предложенные методы оценки стабильности могут, как и все остальное в этом мире, критиковаться и улучшаться, но, во всяком случае, они есть, опубликованы в открытой печати и допускают проверку.

Третья методологическая проблема — учет заимствований — является очень тяжелой (и для классического сравнительно-исторического метода тоже), исследования в этом направлении планируются. То же самое касается и усовершенствования математических алгоритмов анализа данных.

Как можно оценить текущий статус развиваемого метода? Действительно ли имеет место «факт абсолютной практической неприменимости полученных результатов», как считает Г. С. Старостин?

Проведем такой мысленный эксперимент. Представим себе, что в некоей параллельной Вселенной наука развивалась несколько по иному, чем у нас, и сравнительно-исторический метод был придуман не в Европе, а в экваториальной Африке. После классификации местных языков африканские лингвисты обратили бы свое внимание на слабо описанные языки Европы и Азии. Если бы до начала применения к этим языкам строгих процедур сравнительно-исторического метода (на опыте нашей Вселенной нам известно, что построение генетической классификации индоевропейских языков ведется уже 200 лет усилиями десятков тысяч лингвистов, и она до сих пор не завершена) наши гипотетические лингвисты имели возможность познакомиться с данными по типологической близости языков из приведенной в статье диаграммы 3, то они сэкономили бы массу времени.

Они могли бы высказать гипотезу о родстве хинди, русского, французского, английского, персидского языков (которая в дальнейшем была бы подтверждена строгими методами) или весьма правдоподобную гипотезу об урало-алтайском родстве (которая, как мы знаем, существовала в свое время в нашей исторической лингвистике, хотя сейчас ее мало кто признает), и они бы не тратили время на проверку гипотез, скажем, о русско-абхазской семье или франко-кхмерско-ненецкой. Типологическая диаграмма пропускает небольшое число правильных гипотез (индоевропейская принадлежность армянского и кельтского), но отсекает тысячи неверных. Можно предположить, что лингвисты параллельной Вселенной сочтут этот метод важным первым шагом в установлении родства языков. Возвращаясь в нашу Вселенную, мы обнаруживаем, что в точности такая оценка дается (в том числе и Г. С. Старостиным [Starostin G. 2009]) методу массового сравнения Гринберга. Мне кажется, что статус этих двух методов довольно близок.

Почему в статье метод, основанный на типологической близости, не используется для получения новых результатов, например, не применен к какому-либо сложному случаю, вроде построения генетической классификации папуасских языков? Если бы мы поступили так, то полученный результат было бы трудно оценить, в виду отсутствия

общепризнанной классификации папуасских языков. Применив новый метод к хорошо изученным языкам, мы имеем возможность протестировать его. Вопрос следует поставить не так: что нового дает диаграмма 3, а насколько она близка к надежно установленной классификации?

Наконец, перейдем к последнему и самому принципиальному вопросу: почему делается попытка использовать типологические данные для установления родства языков, почему не ограничиться построением ареально-типологической классификации (как предлагает Г. С. Старостин). Причем ведь такая попытка делается не только мной, и не только Николс. Можно привести целый ряд работ других авторов (обратим внимание, в частности, на публикации С. Вихмана, например, [Wichmann & Saunders 2007]).

Г. С. Старостин выделяет три фактора несколько неожиданной популярности «историко-типологических моделей», и с ним можно в целом согласиться. Однако есть и еще один, на мой взгляд, ключевой момент, коррелирующий с первым из этих факторов, но отличающийся от него.

Широко распространенной является точка зрения, что сравнительно-исторический метод ограничен временным диапазоном 6—8 тыс. лет. Совершенно вне зависимости от того, верно это утверждение или нет, оно, на мой взгляд, крайне выгодно для сравнительно-исторического метода, поскольку содержит пресуппозицию, что уж на меньших-то временных отрезках у этого метода проблем нет. Посмотрим, действительно ли это так. Какова степень достоверности результатов, полученных сравнительно-историческим методом на малой временной глубине, и, главное, какова разрешающая способность метода, т. е. какую долю реальных языковых генетических общностей он способен установить? Рассмотрим тюркские языки — это хорошо описанные языки, к которым сравнительно-исторический метод применяется уже давно, причем ведущими специалистами в этой области. Возраст семьи — всего порядка 2 тыс. лет [СИГТЯ 2002].

В *Лингвистическом энциклопедическом словаре* [*ЛЭС*] сообщается о 6 классификациях тюркских языков Богородицкого, Радлова, Корша, Самойловича, Рамстедта, Баскакова. В [Языки мира. Тюркские] описываются также классификации Бенцинга и Менгеса, Маслова, упомянуты классификации Березина, Ильминского, Аристова, Катанова. Вероятно, большинство из этих классификаций уже устарели, поэтому мы обратимся к современным «каноническим» источникам [Языки мира. Тюркские; Бурлак & Старостин 2001, Gordon 2005] и самым последним публикациям ведущего российского коллектива тюркологов [Wichmann & Saunders 2007; СИГТЯ 2006] и посмотрим, совпадают ли эти классификации.

Халаджский язык. В [Бурлак & Старостин 2001] отнесен к самостоятельной ветви, в [Gordon 2005; СИГТЯ 2002, стр. 4] — к огузским языкам (южным в другой терминологии), в [СИГТЯ 2002, стр. 726] — к карлукским языкам.

Чулымский язык. В [Бурлак & Старостин 2001] отнесен к сибирским, в [СИГТЯ 2002] — к кыргызским, в [Gordon 2005] — кыпчакским (западным), в [Языки мира. Тюркские] — к тюркским(!).

Очень странной является генеалогическая классификация некоторых диалектов, приведенная в книге «Языки мира. Тюркские языки». В [Изидинова 1997] крымскотатарский литературный язык, а также его степной диалект отнесены к кыпчакским языкам, при этом южный диалект — к огузским. Следуя определению родства, приведенному в реплике А. В. Дыбо: «По определению, родственные языки — это, как известно, языки, восходящие к общему языку-предку», приходим к парадоксальной ситуации. С одной стороны, диалект — это «разновидность данного языка» [*ЛЭС*]. Все диалекты языка восходят к общему праязыку, от которого они отделились обычно не более чем несколько сот лет назад. Если исходно крымскотатарский язык являлся кыпчакским, то непонятно, как его южный диалект мог сменить свою генеалогическую принадлежность с кыпчакской на огузскую.

С другой стороны, если указанные диалекты восходят к протокыпчакскому и протоогузскому языкам, разделившимся более 1,5 тыс. лет назад [СИГТЯ 2002], то как они могли настолько сблизиться, что стали вариантами одного языка? А. В. Дыбо не исключает, что «Теоретически возможна, конечно, и такая модель развития, при которой носители нескольких … языков в результате тесных контактов между собой развили некоторый общий язык». Может быть, это как раз такой случай?

Эта же классификация крымскотатарских диалектов приведена и в [Бурлак & Старостин 2001]. Более того, такая ситуация не является чем-то исключительным, из ряда вон выходящим. Урумский язык, согласно [СИГТЯ 2002, стр. 5], обладает чертами как огузских, так и кыпчакских языков. У узбекского языка в [Ходжиев 1997] выделено три наречия: кыпчакское, огузское и карлукское!

Представляется, что сравнительно-историческому методу не хватает фундаментальной монографии (на русском языке), в которой подобные неяс-

ные моменты нашли бы ясное теоретическое объяснение. В англоязычной литературе дело обстоит несколько лучшим образом. Несколько серьезных монографий и обзорных статьи выпустил, пожалуй, ведущий на Западе защитник сравнительно-исторического метода в его наиболее строгой форме Л. Кемпбелл. Его последняя монография [Campbell & Poser 2008], рецензировавшаяся в журнале Вопросы языкового родства [Starostin G. 2009], однако, вопреки желанию автора, наносит серьезный удар по всему сравнительно-историческому методу.

В какой-то мере, подводя итоги развития этого метода, монография заставляет читателя поставить вопрос: каковы конечные цели и задачи метода, что с его помощью можно сделать? Что может дать сравнительно-исторический метод: полное описание дерева эволюции языков семей со всеми разветвлениями (пусть и на ограниченной временной глубине) или только сами семьи?

В монографии Кемпбелла задача сравнительно-исторического метода, фактически, сводится к ответу на вопрос — родственны два заданных языка или нет, т. е. результатом будет описание только семей (неструктурированного множества языков, входящих в семью), но не структуры дерева эволюции семьи. Если это так, то в исторической лингвистике обнаруживается большая лакуна, которую займут другие методы, в том числе и «историко-типологический».

Если же цель сравнительно-исторического метода состоит все же в описании полного дерева эволюции, то давайте посмотрим, в какой мере эта цель достигнута. В [Бурлак & Старостин 2001] тюркские языки делятся на 7 равноправных ветвей, никак не объединяемых в более крупные кластеры. Но могло ли быть так, чтобы народ, являвшийся носителем прототюркского языка, вдруг одномоментно распался на 7 народов, разбежавшихся в разные стороны? Здравый смысл подсказывает, что это абсурд. Даже одновременное деление протоязыка на 3 языка-потомка представляется достаточно маловероятным. Неясно, что может заставить народ разделиться одномоментно на несколько частей, двинувшихся в разные стороны. Разумеется, в истории имели место серьезные социально-политические изменения, которые приводили к распаду языковых общностей. Например, падение Римской империи привело к региональной дифференциации латыни и, в итоге, к появлению романских языков. Однако такого рода события являются далеко не одномоментными. Римская империя разделилась на Западную Римскую империю и Восточную Римскую империю в 395 г., а Восточная Римская империя распалась затем только почти через сто лет — в 476 г. [Неронова 1983]. Романские же языки окончательно обособились лишь в IX веке [ЛЭС; Неронова 1983]. Неизвестны строго доказанные случаи одновременного распада народа на несколько частей с формированием разных языков. В случае распада прототюркского логично предположить, что мы просто не можем определить точный порядок отделения ветвей.

Если семья состоит из n языков, то полное дерево эволюции будет состоять из $n - 1$ внутренней (т. е. не концевой) вершины, отражающих все генетические подгруппы семьи (промежуточные протоязыки и случаи их деления на 2 потомка). Деревья именно с таким характером ветвления строят большинство современных филогенетических алгоритмов [Semple & Steel 2003].

В [Бурлак & Старостин 2001] перечислено 36 тюркских языков с 8 внутренними вершинами (вершина всего дерева — прототюркский и семь ветвей). Таким образом, в этой классификации идентифицировано лишь 23% вершин из полного дерева. В [Gordon 2005] классифицировано 40 тюркских языков в дереве с 12 внутренними вершинами, т. е. идентифицировано лишь 32% от всех вершин.

В рамках сравнительно-исторического метода построено и полное дерево эволюции тюркских языков. В [СИГТЯ 2002] приведено дерево, включающее 41 тюркский язык с бинарным ветвлением вершин за исключением одного-единственного случая тернарного ветвления (это исключение не в счет). К сожалению, первые же сравнения его с данными других публикаций показывают, что оно вряд ли претендует на статус надежно установленного. Например, тувинский и тофаларский в этом дереве отнесены к одной ветви с якутским, а в [Бурлак & Старостин 2001] они входят в группу сибирских языков, в которой нет якутского. На этом дереве башкирский находится на одной ветви с ногайским, а татарский на другой ветви, а в [Gordon 2005], наоборот, башкирский с татарским находятся на одной ветви, а ногайский на другой.

Если перейти к более высокому уровню, то мы обнаружим, что в дереве из [СИГТЯ 2002] от прототюркского первым отделяется чувашский язык, затем сибирская ветвь и только потом обособляются кыпчакские, карлукские и огузские ветви. В другой же монографии [СИГТЯ 2006], написанной практически тем же коллективом авторов, что и [СИГТЯ 2002], сначала от прототюркского также отделяется чувашский, но далее реконструируется иная картина эволюции. На стр. 771 приводится

дерево, в котором карлукско-кыпчакские языки располагаются на одной ветви с сибирскими, а огузские располагаются отдельно.

Таким образом, хотя с помощью сравнительно-исторического метода сгенерировано много гипотез о структуре дерева эволюции тюркских языков, на роль надежно установленных может претендовать лишь 6—7 группировок языков, т. е. в любом случае не более четверти от общего числа.

Может быть, тюркские языки недостаточно изучены? Давайте посмотрим, как обстоят дела с самым изученным семейством — индоевропейскими языками. В [Бурлак & Старостин 2001] в классификацию включено порядка 190 языков, в дереве 36 внутренних вершин — это около 20%. В генеалогической классификации [Gordon 2005] присутствует 439 индоевропейских языков и 91 внутренняя вершина — опять примерно 20%.

Итак, разрешающая способность сравнительно-исторического метода находится где- то в районе 20%, т. е. он позволяет идентифицировать (даже на малых временных глубинах) только пятую часть языковых генетических общностей. Если исследования будут продолжать такими темпами, то для построения полного дерева индоевропейских языков потребуется еще лет 800. Впрочем, складывается впечатление, что темпы получения новых результатов здесь замедлились — основные генетические группы индоевропейских языков Мейе знал еще сто лет назад [Meillet 1903].

Вероятно, водораздел возможностей сравнительно-исторического метода проходит не между «до» и «после» рубежа в 6—8 тыс. лет назад, а между четко выделяющимися по каким-то историческим причинам генетическими группами (восточно-славянские, тюркские, ностратические и т. д.) и нечеткими группами (сибирские?, романо-кельтские?, словенско-сербохорватские? и т. д.), плохо поддающимися сравнительно-историческому методу, независимо от их возраста.

Это еще один фактор, стимулирующий развитие новых методов, прежде всего, с привлечением новых данных. Невозможно спорить, что генетическая близость является одной из причин типологической близости и, следовательно, диаграммы типологической близости содержат генетический сигнал (как сейчас принято говорить). Вопрос состоит в том, как эту информацию оттуда извлечь и как использовать.

С пуском Большого адронного коллайдера в ЦЕРНе у физиков появилась надежда на построение, как ее называют, Теории Великого объединения (включающей все известные фундаментальные физические силы). Если такая теория будет создана, то она явится обобщением классических работ Гелл-Манна по квантовой хромодинамике и слабому взаимодействию. Может быть, и в лингвистике пришла пора Великого объединения всех накопленных данных для совершения прорыва в увлекательнейшей области эволюции языка?

Литература

Бурлак & Старостин 2001 — Бурлак С. А., Старостин С. А. *Введение в лингвистическую компаративистику.* М.: УРСС, 2001.

Володин 1997 — Володин А. П. Палеоазиатские языки. В сб. *Языки мира. Палеоазиатские языки*. М.: Индрик, 1997. С. 8—11.

Грунтов и др. 2006 — Грунтов И. А., Дыбо А. В., Кормушин И. В., Решетников К. Ю. Палеокультура, прародина а внешние связи языков алтайской семьи в древности. В сб. *Этнокультурное взаимодействие в Евразии*. Кн. 2. М.: Наука, 2006. с.30—40.

Дыбо 1978 — Дыбо В. А. Ностратическая гипотеза (итоги и проблемы) // *Известия АН СССР*, Сер. литературы и языка, т. 37, №5, 1978.

Изидинова 1997 — Изидинова С. Р. Крымскотатарский язык // *Языки мира. Тюркские языки*. М.: Издательство «Индрик», 1997. С. 298—308.

Кузнецов 1954 — Кузнецов П. С. *Морфологическая классификация языков*. М., 1954.

ЛЭС — Лингвистический энциклопедический словарь. М.: «Советская энциклопедия», 1990.

Маслова 2004 — Маслова Е. Динамика типологических распределений и стабильность языковых типов // Вопросы языкознания. № 5. 2004. С. 3—16.

Неронова 1983 — Неронова В. Д. Вторжение варваров и крушение Римской империи // *История древнего мира. Упадок древних обществ*. М.: Гл. ред. восточной литературы, 1983. С. 239—273.

Николаева & Хелимский 1997 — Николаева И. А., Хелимский Е. А. Юкагирский язык. В сб. *Языки мира. Палеоазиатские языки*. М.: Индрик, 1997. 155—168.

Поляков & Соловьев 2006 — Поляков В. Н., Соловьев В. Д. *Компьютерные модели и методы в типологии и компаративистике*. Казань: КГУ, 2006.

СИГТЯ 2002 — *Сравнительно-историческая грамматика тюркских языков. Региональные реконструкции* / Ред. Поцелуевский Е. А.. М.: Наука, 2002.

СИГТЯ 2006 — *Сравнительно-историческая грамматика тюркских языков. Пратюркский язык-основа* / Ред. Поцелуевский Е. А.. М.: Наука, 2006.

Соловьев & Фасхутдинов 2009 — Соловьев В. Д., Фасхутдинов Р. Ф. Методика оценки стабильности грамматических свойств // *Известия РАН. Серия литературы и языка*. Т. 68. № 4. 2009.

Старостин 1989 — Старостин С. А. Сравнительно-историческое языкознание и лексикостатистика // *Лингвистическая реконструкция и древнейшая история Востока*. Ч. 1. М., 1989. С.3—39.

Старостин 2007 — Старостин С. А. Гипотеза о генетических связях сино-тибетских языков с енисейскими и северокавказскими языками // С. А. Старостин. *Труды по языкознанию*. М.: Языки славянских культур, 2007. С. 265—282.

Трубецкой 1987 — Трубецкой Н. С. Мысли об индоевропейской проблеме // Трубецкой Н. С. *Избранные труды по филологии*. М., 1987.

Ходжиев 1997 — Ходжиев А. П. Узбекский язык // *Языки мира. Тюркские языки*. М.: Издательство «Индрик», 1997. С. 426—436.

Эдельман 1997 — Эдельман Д. И. Бурушаски язык // *Языки мира. Палеоазиатские языки*. М.: Индрик, 1997. С. 204—220.

Языки мира. Тюркские — *Языки мира: Тюркские языки*. М.: «Индрик», 1997.

ASJP — Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the World's languages: A description of the method and preliminary results. *STUF* — Language Typology and Universals 61.4. 285—308.

Bryant et al. 2005 — Bryant, D., Filimon, F. and Gray, R. Untangling our past: Languages, Trees, Splits and Networks // *The Evolution of Cultural Diversity: Phylogenetic Approaches*. Editors: R. Mace, C. Holden, S. Shennan. Publisher: UCL Press, 2005, pp. 69—85.

Campbell & Poser 2008 — Lyle Campbell & William J. Poser. *Language Classification: History and Method*. Cambridge: Cambridge University Press. 2008.

Comrie & Cysouw 2006 — Comrie B., Cysouw M. New Guinea through the eyes of WALS. Language and Linguistics in Melanesia. http://email.eva.mpg.de/~cysouw/publications.html. 2006.

Comrie 1989 — Comrie B. *Language Universals and Linguistic Typology*. Chicago: The University of Chicago Press. 1989.

Cysouw & Comrie 2009 — Cysouw M., Comrie B. How varied typologically are the languages of Africa? // Rudie Botha & Chris Knight. (eds.). *The Cradle of Language*. Oxford, 2009.

Everaert & Musgrave 2009 — Everaert M., Musgrave S. (eds.). *The Use of Databases in Cross-Linguistic Studies.* Berlin, 2009.

Gordon 2005 — *Ethnologue: Languages of the World*. Gordon R. G. Jr. (ed.). SIL International: Dallas, 2005, http://www.ethnologue.com.

Gray & Atkinson 2003 — Gray R., Atkinson Q. Language-tree divergence timer support the Anatolian theory of Indo-European origin // *Nature*. V. 426. 2003. p. 435—439.

Greenhill et al. 2010 — Greenhill S. J., Atkinson Q. K., Meade A., Gray R. D. The shape and tempo of language evolution // *Proceedings of the Royal Society: Biological Sciences* 2010, Apr. 7, http://simon.net.nz/files/2010/04/Greenhill_et_al2010-preprint.pdf.

Meillet 1903 — Meillet F. *Introduction à l'étude comparative des langues indoeuropéennes*. Paris. 1903.

Nichols 1992 — Nichols J. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press. 1992.

Nichols 2007 — Nichols J. Typology in the service of classification. http://aalc07.psu.edu/papers/jn_typol_class3.pdf. Stanford, 2007.

Polyakov et al. 2009 — Polyakov V., Solovyev V., Wichmann S., Belyaev O. Using WALS and Jazyki mira // *Language Typology*. V. 13. 2009. P. 135—165.

R-Project — R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2006. http://www.R-project.org.

Semple & Steel 2003 — Semple C., Steel M.: *Phylogenetics*. New York. Oxford University Press, 2003.

Starostin 2007 — S. STAROSTIN. Indo-European among other language families problems of dating, contacts and genetic relationships // С. А.СТАРОСТИН. *Труды по языкознанию*. М.: Языки славянских культур. 2007. С. 806—820.

Starostin G. 2009 — G. STAROSTIN. Book review: Lyle Campbell & William J. Poser. Language Classification: History and Method, 2008 // *Вопросы языкового родства*. № 2009, с. 158—174.

WALS — *The World Atlas of Language Structures* / HASPELMATH M., DRYER M., GIL D., COMRIE B. (Eds.) Oxford, 2005.

Warnow 1997 — WARNOW T. Mathematical approaches to comparative linguistics // *Proc. Natl. Acad. Sci. USA.* 1997, v. 94.

Wichmann & Holman 2009 — WICHMANN S., HOLMAN E. W. *Temporal stability of linguistic typological features*. Lincom: Muenchen. 2009.

Wichmann & Saunders 2007 — WICHMANN S., SAUNDERS A. How to use typological database in historical linguistic research // *Diachronica*, 2007, v. 24, №2.

Known typological classifications of languages, as a rule, are based upon one or just a few select features. With the appearance of large typological databases such as The World Atlas of Language Structures or “Languages of the World”, it has become possible to build classifications that simultaneously account for hundreds, if not thousands, of features. In building these classifications, one can employ various mathematical algorithms, thus raising their level of objectivity. The present paper discusses the first results achieved in this field. Some of the most interesting and least expected discoveries include the typological differentiation between the languages of Africa and Eurasia as well as typological proximity between Chukchee-Kamchatkan and Sino-Caucasian languages. The paper offers several possible variants for the classification of the languages of Northern Africa, Europe, North and Central Asia. It also discusses the possibility of using this methodics for the establishment of distant relationship between languages, based on two considerations: (a) simultaneous matches between hundreds of features cannot be coincidental; (b) at least some grammatical features are quite stable and capable of preserving information on language stages that go back very deep in time.