ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 1 2010

КРИТИКА И БИБЛИОГРАФИЯ

ОБЗОРЫ

© 2010 г. В. Д. СОЛОВЬЕВ

ТИПОЛОГИЧЕСКИЕ БАЗЫ ДАННЫХ: ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ*

В статье дается обзор выполненных в последние годы зарубежных работ в области применения типологических баз данных. Основное внимание уделяется наиболее интересным содержательным результатам, используемый математический аппарат и программный инструментарий не описываются подробно. Показано, что типологические базы данных обеспечивают новые возможности для решения проблем классификации языков. Классификация языков рассматривается в различных аспектах: генеалогическом, ареальном, эволюционном.

1. ВВЕДЕНИЕ

Одним из новых и многообещающих применений компьютерных технологий в лингвистике являются типологические базы данных (далее, сокращенно ТБД). ТБД содержат описания в формализованном виде грамматических свойств языков и сопровождаются инструментальными средствами интерфейса и статистических расчетов. Развитие этого направления начиналось с небольших баз данных, посвященных весьма ограниченному кругу свойств и содержащих описания небольшого числа языков. Примеры баз данных и общий обзор проблематики применения ТБД можно найти в [Everaert, Musgrave 2009].

Новый этап развития ТБД начался с появления World atlas of language structures (WALS) [Haspelmath et al. 2005] и базы данных «Языки мира». Последняя создана в Институте языкознания РАН на основе одноименной серии монографий. Описание базы данных «Языки мира» можно найти в [Поляков, Соловьев 2006; Виноградов и др. 2003], она доступна в Интернете по адресу http://dblang2008.narod.ru/.

WALS и «Языки мира» можно назвать большими типологическими базами данных, каждая из них содержит более 1 млн бит информации. В WALS описывается более 2500 языков по 142 свойствам (из них 128 грамматических), каждое из которых принимает одно из нескольких значений – от 2 до 9. «Языки мира» содержат описания 315 языков по 3821 свойству. В обеих базах данных свойства охватывают все разделы грамматики.

ТБД исходно создавались как справочники с удобным пользовательским интерфейсом, позволяющим быстро находить интересующую информацию. Однако быстро выяснилось, что ТБД предоставляют принципиально новые возможности изучения грамматик языков с применением математических (в том числе, статистических) и компьютерных методов. Многие явления, которые до сих пор рассматривались лишь на качественном уровне и на основе отдельных примеров, теперь могут изучаться ко-

* Работа выполнена при финансовой поддержке ФАО РФ, проект № 2.2.1.1/6944 «Развитие Российского научно-образовательного центра по лингвистике им. И.А. Бодуэна де Куртенэ».

личественными методами с использованием огромных массивов информации. Важным аспектом подобных исследований является их объективный характер, основанный на применении строгих математических методов. Вот несколько примеров вопросов, на которые появилась возможность дать ответ с помощью ТБД.

1. Насколько однородным является тот или иной языковой ареал? Можно ли считать его языковым союзом? ТБД позволяют применить количественные методы в ареальной лингвистике для оценки степени близости языков.

2. Как происходило распределение языковых признаков в связи с распространением человечества и собственно языковой эволюцией? Пионерские исследования Дж. Николс [Nichols 1992] в этом направлении проводились на очень ограниченной выборке данных. Современные ТБД могут помочь уточнить многие аспекты расселения человечества.

3. Установление (дальнего) языкового родства. Как принято считать, сравнительно-исторический метод позволяет реконструировать историю развития языков не более чем на 8–10 тысяч лет. Есть надежда, что тщательный анализ грамматических свойств (многие из которых, вероятно, более стабильны, чем лексические) позволит выявить сверхдальнее родство.

4. Языковая динамика: какие разделы грамматики языков меняются быстрее? С какой скоростью?

В данной статье дается обзор недавних (2005–2008 гг.) зарубежных работ в этой области. Изложение концентрируется на наиболее интересных содержательных результатах, используемый математический аппарат не описывается подробно. Изложение основных применяемых в этой области математических идей можно найти в [Semple, Steel 2003]. Все публикации условно разобьем на три группы: классификация языков, изучение пространства признаков, общие вопросы организации ТБД. В настоящей статье рассматривается лишь первая группа работ.

Исследования в этом направлении, в свою очередь, можно разбить по главной цели работы на следующие основные группы:

- ареально ориентированные исследования
- генеалогически ориентированные исследования
- исторически ориентированные исследования

2. АРЕАЛЬНО ОРИЕНТИРОВАННЫЕ ИССЛЕДОВАНИЯ

В исследованиях этой группы рассматриваются языки определенного региона на предмет степени их близости (грамматической или, в другой терминологии, типологической). Общая идея изучения близости языков состоит в подсчете расстояния между ними по Хеммингу. Это число признаков, по которым данные языки принимают разные значения (т. е. в одном из языков признак присутствует, а другом отсутствует). Учет большого числа признаков позволяет нивелировать влияние случайностей. В дальнейшем под расстоянием между языками будет подразумеваться расстояние по Хеммингу (хотя существуют и другие способы определения расстояния).

Расстояние между языками характеризует степень, в которой грамматики языков похожи друг на друга. Если расстояние мало, значит грамматики языков схожи. В этом случае можно говорить о типологической близости языков.

Опишем цикл работ Б. Комри (с соавторами), выполненных на материале WALS. База данных WALS охватывает языки всех регионов и семей мира, что позволяет проводить интересные исследования, характеризующие степень языкового разнообразия/ единообразия в различных регионах, сопоставить влияние на степень типологической близости языков генетической близости и языковых контактов. Три работы Б. Комри посвящены трем регионам с принципиально разной степенью близости языков. Это Новая Гвинея, Юго-Восточная Азия и Африка. Как известно, первый из этих регионов характеризуется крайним разнообразием языков, второй, наоборот, может претендовать на статус языкового союза, Африка же в этом аспекте занимает промежуточное положение.

2.1. Языки Новой Гвинеи

В работе [Comrie, Cysouw 2006] отмечается, что хотя давно известно, что в Новой Гвинее распространено огромное число языков и языковых семей, все же степень их типологического разнообразия никогда строго не исследовалась. В работе рассматриваются вопросы: имеют ли языки данного региона какие-то специфические особенности, которые отличают их от языков остального мира, какова степень их типологического разнообразия по сравнению с языками всего мира.

Выбрано 48 языков из 48 традиционно выделяемых языковых семей Новой Гвинеи. В каждой семье выделялся язык из числа наиболее полно описанных в WALS. Для сравнения их с языками всего мира случайным образом выделено 48 семей, и в каждой из них также выбрано по одному наиболее полно описанному языку. Затем к описаниям выбранных 96 языков применен алгоритм NeighborNet. Это очень популярный в настоящее время алгоритм, первоначально разработанный для нужд эволюционной биологии. Он применим как к видам живых организмов, так и к языкам и строит специфическую картинку – сеть, в которой близкие (согласно расстоянию по Хеммингу) объекты располагаются близко друг к другу. 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

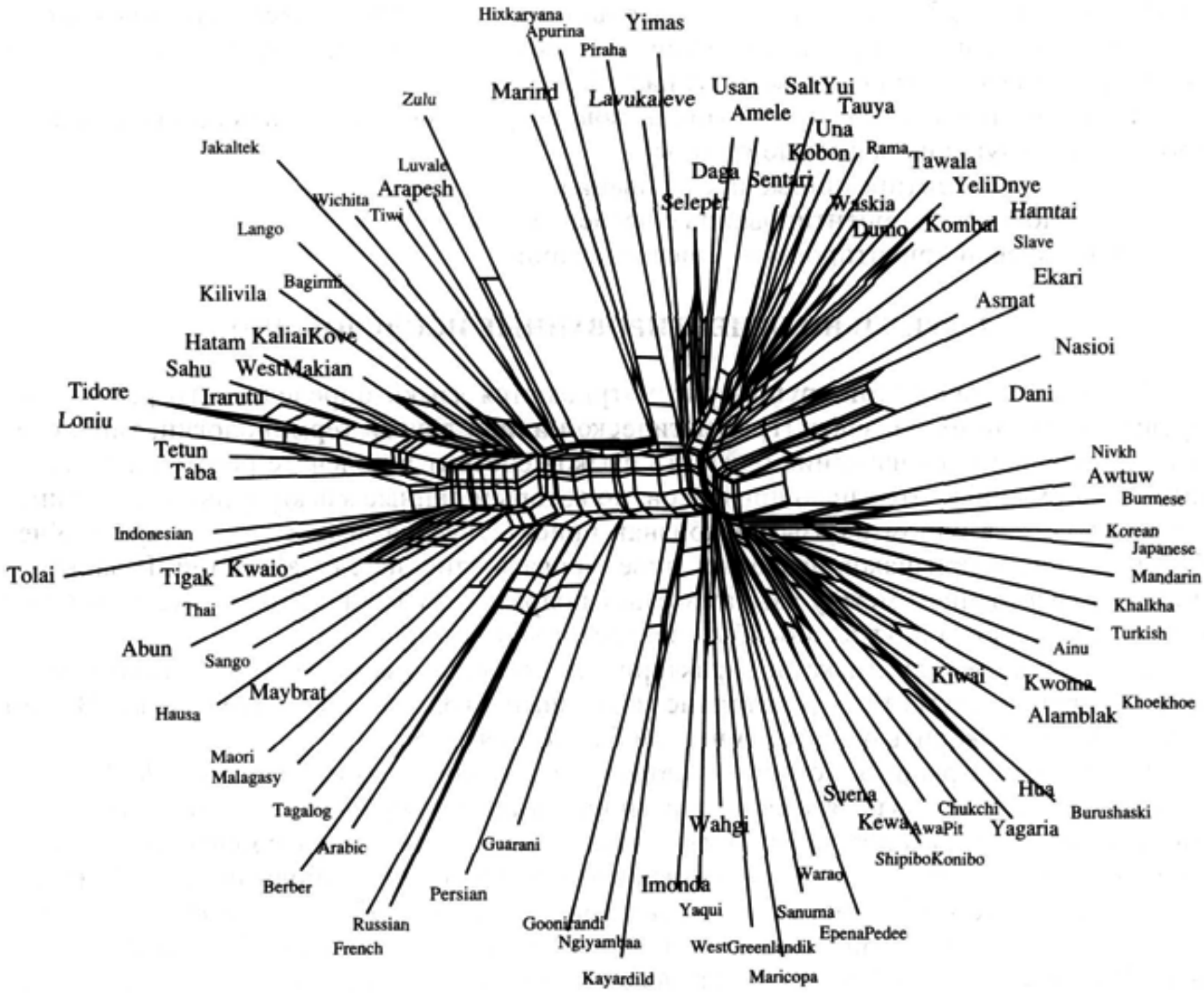

Рис. 1. Языки Новой Гвинеи vs. языки остального мира

австралийских аборигенов Kayardild, языки Hua, Awtuw – к азиатским языкам Burmese и Burushaski соответственно и так далее. Другими словами, типологическое разнообразие языков Новой Гвинеи вполне сопоставимо с типологическим разнообразием языков всего мира. Не будет слишком большим преувеличением сказать, что какой бы язык мы ни выбрали, найдется язык Новой Гвинеи, имеющий схожую грамматику.

Следует отметить, что все расстояния между языками, разумеется, существуют и могут быть предъявлены в числовом виде. Графическая форма используется, скорее, для наглядности представления результатов.

Из других результатов статьи можно выделить следующий. Установлена внутренняя типологическая классификация ново-гвинейских языков на 2 группы: австронезийские и западно-папуасские языки против остальных папуасских языков.

2.2. Языки материковой части Юго-Восточной Азии

В противоположность Новой Гвинее, языки материковой части Юго-Восточной Азии демонстрируют удивительную похожесть, несмотря на то, что принадлежат к 6 разным семьям: австронезийской, сино-тибетской, австроазиатской, тай-кадайской и хмонг-мен.

В работе [Comrie 2007] для анализа выбраны те свойства, которые описаны для достаточно большого числа языков этого региона. В WALS таких оказалось 21. Для каждого из этих свойств воспроизводится соответствующая карта из WALS и рассматриваются значения этого свойства в языках региона. Пример свойства – 'Порядок объекта и глагола' приведен на рис. 2. Практически все языки рассматриваемого региона имеют порядок VO. Аналогичная ситуация и с остальными 20 свойствами.

Это строго поддерживает мнение, что материковая часть Юго-Восточной Азии действительно является внутренне однородной лингвистической областью (языковым союзом).

Из других результатов статьи наиболее интересны следующие. Естественным является вопрос: если сама рассматриваемая область лингвистически однородна, то что происходит на ее границах с остальным миром? Как, в общем-то, и следовало ожидать, выявилась переходная область с промежуточными свойствами. Причем, граница

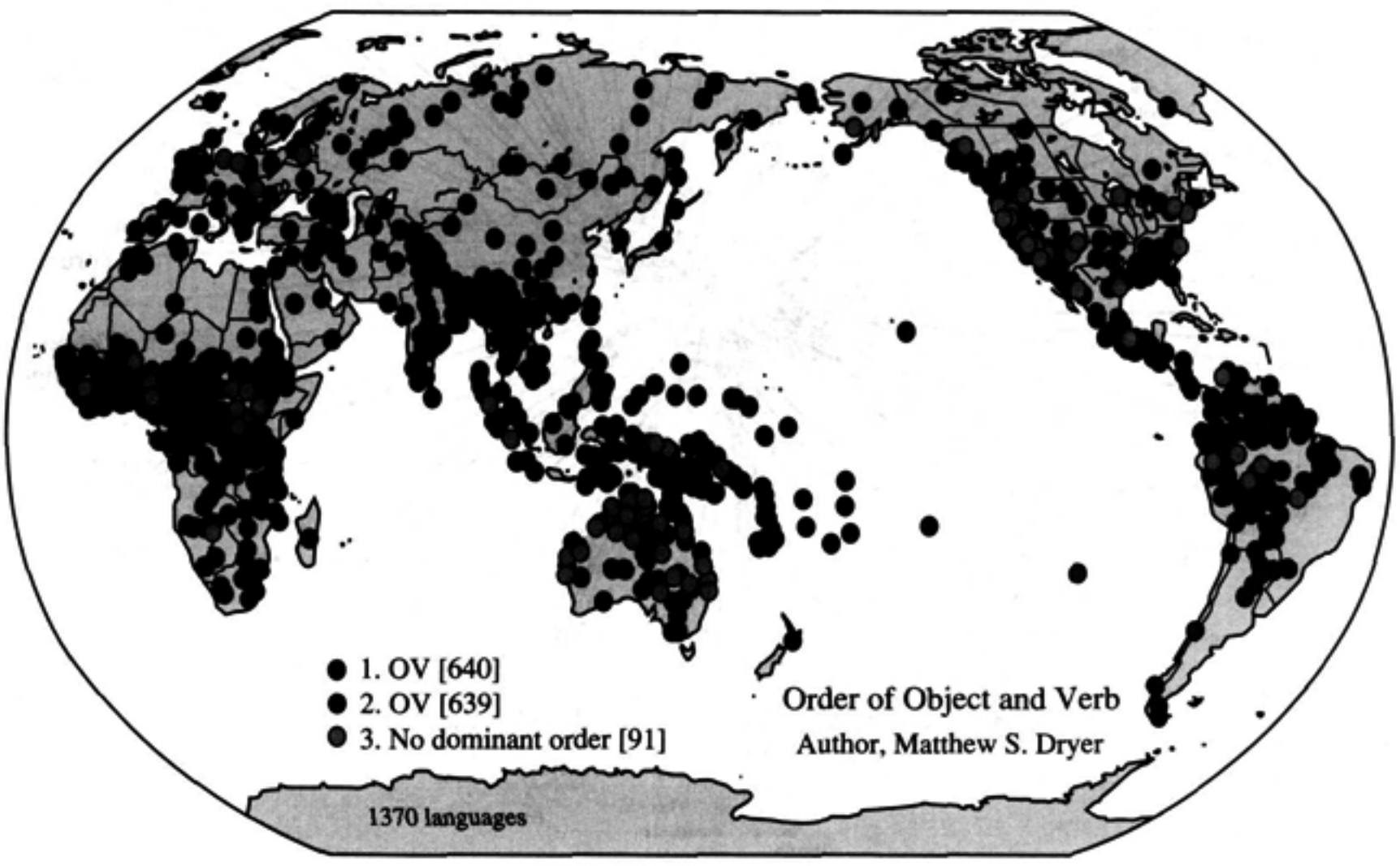

Рис. 2. Карта порядка объекта и глагола

материковой части Юго-Восточной Азии с остальной частью Азии весьма четкая, а с островами – менее четкая.

Еще одно интересное наблюдение: наиболее типичным языком этой области оказался тайский, и это несмотря на то, что тайский язык относительно недавно появился в этом регионе. Понятие «типичный» имеет в данном случае совершенно строгий смысл – тайский язык обладает 19 свойствами из 21 – наибольшим числом свойств из всех языков региона.

2.3. Африканские языки

Области, рассмотренные в предыдущих двух разделах, представляют собой крайние противоположности по степени типологического разнообразия языков. Африка занимает между ними промежуточное положение. В работе [Cysouw, Comrie 2009] рассматриваются следующие вопросы: Образуют ли африканские языки однородную группу, выделяющуюся из всех языков мира? Совпадает ли типологическая близость с генетической классификацией Гринберга?

Аналогично подходу, примененному при изучении языков Новой Гвинеи, выбраны представители всех языковых семей и представлены на схеме (рис. 3) с помощью алгоритма NeighborNet. Жирным шрифтом выделены африканские языки. Легко видеть, что они не образуют компактную группу, а перемежаются другими языками. Таким образом, ответ на первый вопрос отрицательный. Правда, по сравнению с языками Новой Гвинеи, африканские языки все же не так разнообразны и располагаются (кроме Khoekhoe) в одном секторе, в котором находятся еще только языки Юго-Восточной Азии.

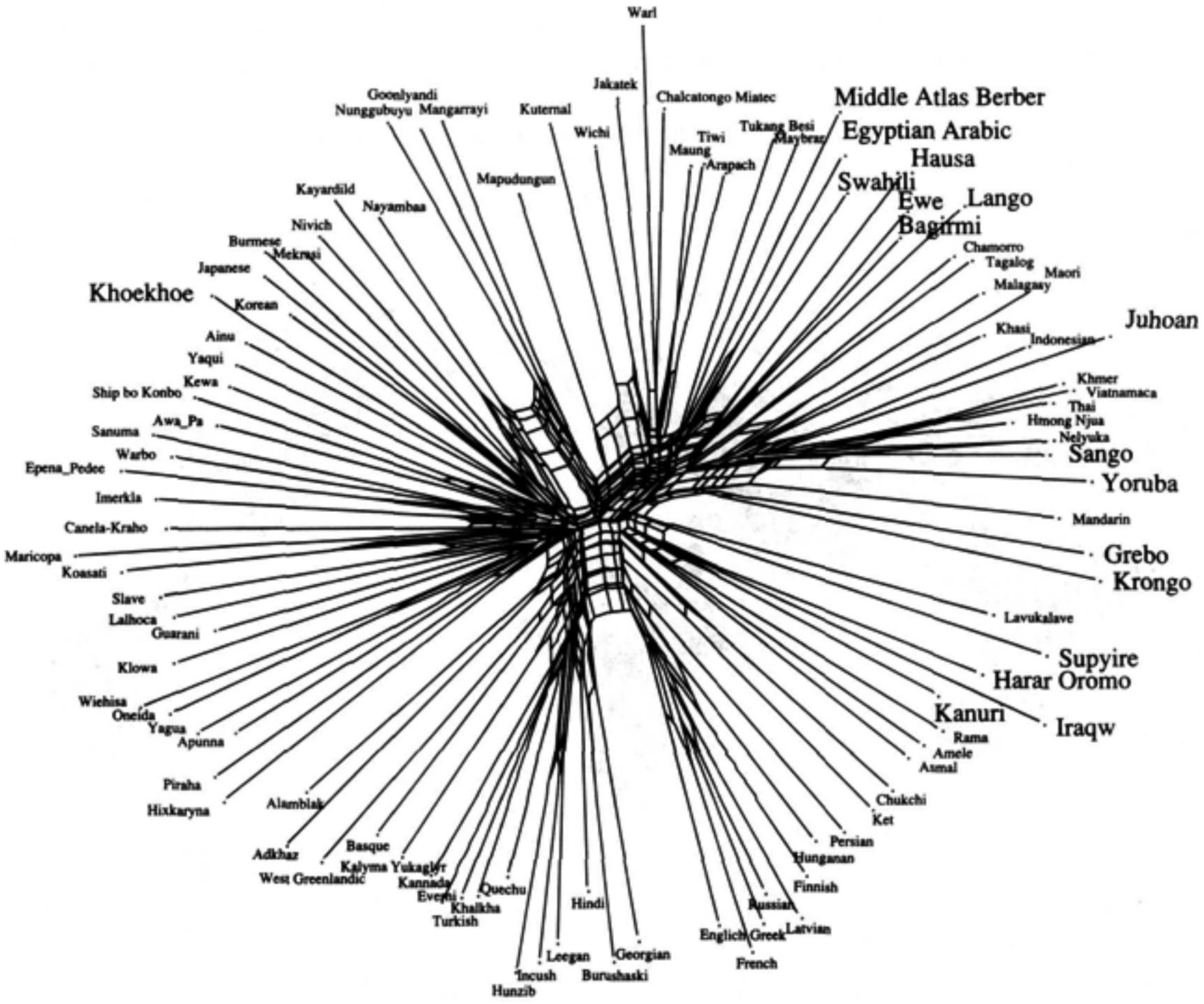

Рис. 3. Африканские языки vs. языки остального мира

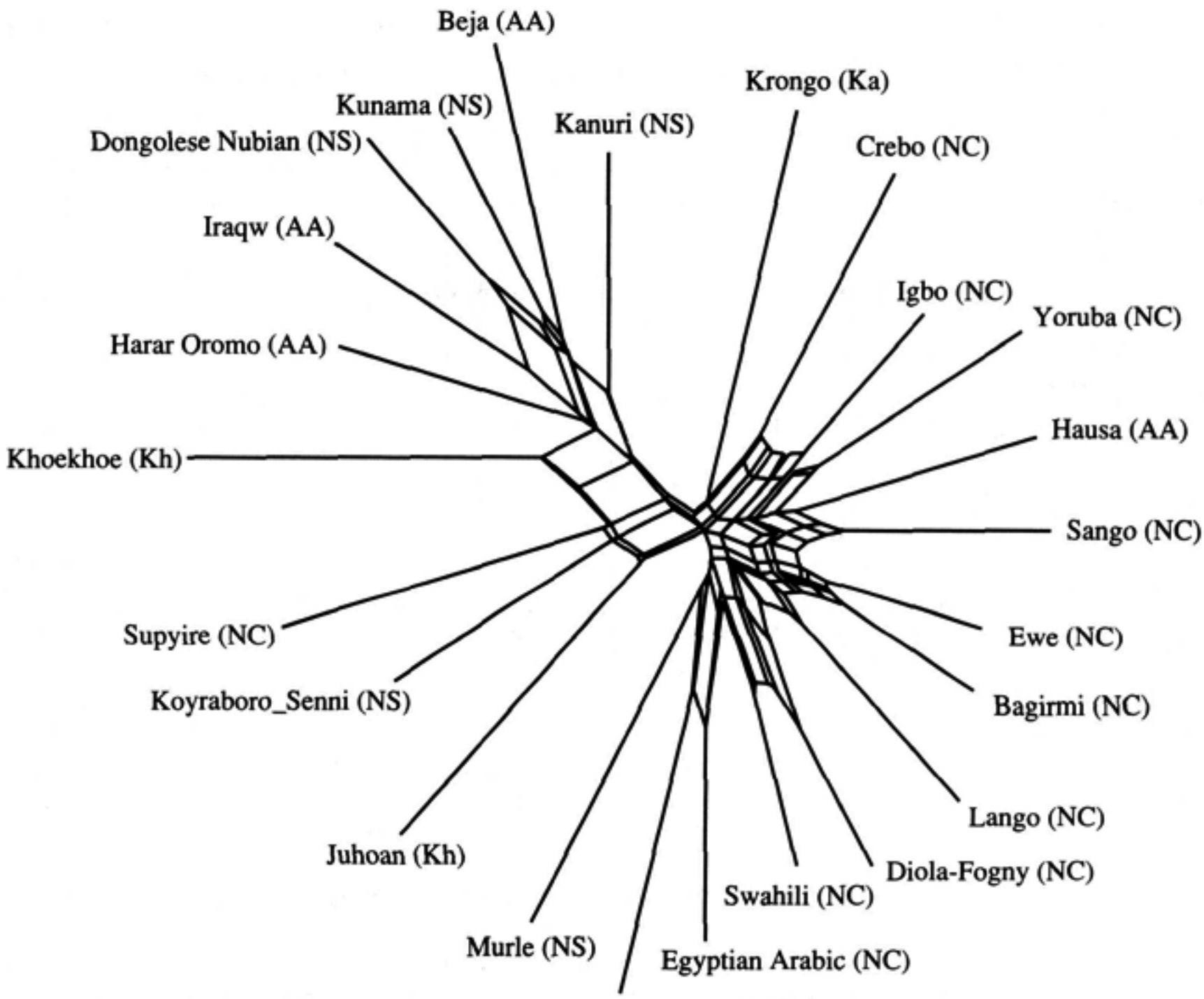

Рис. 4. Типологическая близость африканских языков

На рис. 4. представлена типологическая близость африканских языков. В скобках около языка указана его генетическая принадлежность по Гринбергу: AA означает афразийскую семью, NS – нило-сахарскую, NC – нигеро-конголезскую, Kh – койсанскую, Ka – кадугли. Здесь тоже легко видеть, что типологически близкими оказываются языки из разных семей, и, наоборот, языки из одной семьи могут располагаться далеко друг от друга, т. е. значительно типологически отличаться. Ответ на второй вопрос также оказывается отрицательным.

Автор специально подчеркивает, что отсюда нельзя сделать вывод о неверности классификации Гринберга. Языки из разных семей могут оказаться типологически близкими благодаря заимствованиям. Таким образом, схожесть грамматик может иметь разную природу – либо свидетельствовать о родстве языков, либо являться следствием заимствований. На данном этапе исследований четко различить эти две причины довольно сложно и приходится говорить только о типологической (грамматической) близости языков.

2.4. Мальтийский язык

В контексте сопоставления влияния на близость языков двух факторов – родства и заимствований – интересной представляется работа по мальтийскому языку [Comrie 2009], генетически афразийскому, но находившемуся длительное время в контакте с романскими языками. Для сравнения с мальтийским языком (ниже в таблице 1 обозначен буквой M) выбраны египетский (E) и испанский (S).

Таблица 1

Сравнение мальтийского языка с египетским и испанским

Разделы WALS	S = M = E	S = M ≠ E	S ≠ M = E	S = E ≠ M	S ≠ M ≠ E	Всего
Все категории	60	14	36	6	4	120
Фонология	101	3	4	1	1	19
Морфология	2	3	3	0	0	8
Категории существительных	11	0	11	2	0	24
Синтаксис существительных	3	0	1	1	1	6
Глагольные категории	10	0	3	2	0	15
Порядок слов	8	3	2	0	1	14
Простые предложения	8	4	9	0	1	22
Сложные предложения	1	0	2	0	0	3
Лексикон	7	1	1	0	0	9

Из 120 рассматривавшихся признаков (исключены признаки трех категорий: не описанные для данных языков, не грамматические, а также логически выводимые из других, т. е. не являющиеся независимыми) большая часть – 60 – совпадает для всех трех языков. Для 36 признаков их значения в мальтийском и египетском совпадают, но отличаются от испанского. Для 14 признаков, наоборот, их значения совпадают в мальтийском и испанском языках, но отличаются в египетском. Таким образом, генетическое родство оказывает заметно большее влияние, чем заимствования.

Соответствующие данные получены также и по различным разделам грамматики. Оказалось, что (в данном случае) разные разделы грамматики ведут себя совершенно по-разному. Так, свойства существительных совершенно не заимствовались, в то время как, например, порядок слов в мальтийском языке ближе к испанскому, чем к египетскому. Представляется интересным провести аналогичный анализ для других контактных ситуаций.

2.5. Баскский язык

Баскский язык считается изолятом, его генетические корни не установлены. Высказывалось много гипотез о его родственных связях, из которых наиболее часто упоминаемой является гипотеза о родстве с кавказскими языками. В то же время баскский язык давно находится в контакте с испанским и французским языками, что могло оказать влияние на его грамматический строй. Интересные данные по типологической близости баскского языка получены в [Comrie 2008].

На таблице 2 приведены выбранные для сравнения языки из различных семей. Воспроизведены названия и сокращения, принятые в WALS. В таблице 3 – расстояния между баскским и выбранными языками.

Расстояния между баскским и испанским и между баскским и французским языками весьма велики, т. е. заимствования из этих языков не оказали существенного влияния на грамматический строй баскского языка. Ближе всего баскский к турецкому и хинди. Этот результат совершенно неожиданный и труднообъяснимый. Требуются до-

Таблица 2

Языки, сравниваемые с баскским

Язык	Сокращение	Ветвь семьи/семья
Баскский	bsq	Баскский/Баскский
Французский	fre	Италийские/Индоевропейские
Испанский	spa	Италийские/Индоевропейские
Ирландский	iri	Кельтские/Индоевропейские
Русский	rus	Славянские/Индоевропейские
Хинди	hin	Индийские/Индоевропейские
Финский	fin	Финские/Уральские
Турецкий	tur	Тюркские/Алтайские
Абхазский	abk	Северо-западные кавказские/ Северо-западные кавказские
Грузинский	geo	Картвельские/картвельские
Лезгинский	lez	Лезгинские/Нахско-дагестанские
Арабский (египетский)	aeg	Семитские/Афразийские
Берберский (Малый Атлас)	bma	Берберские/Афразийские
Бурушаски	bur	Бурушаски/Бурушаски
Чукотский	chk	Чукотско-камчатские/ Чукотско-камчатские
Западногренландский	grw	Эскимосско-алеутские/ Эскимосско-алеутские
Навахо	nav	Атабасканские/На-дене

полнительные исследования, в том числе с применением других ТБД. Следующие по степени близости к баскскому языку – грузинский, абхазский, лезгинский, бурушаски и язык гренландских эскимосов. Широкое представительство здесь кавказских языков ожидаемо и в какой-то мере подтверждает гипотезу об их родстве с баскским.

Из других расстояний, приведенных в таблице, можно отметить, что наиболее близкими оказались французский и испанский, что и следовало ожидать, учитывая, что это единственные два языка в выборке из одной ветви семьи. Таким образом, полученные по данным WALS расстояния частично ожидаемые, частично неожиданные.

3. ГЕНЕАЛОГИЧЕСКИ ОРИЕНТИРОВАННЫЕ ИССЛЕДОВАНИЯ

К этой категории отнесены работы, целью которых является установление родства языков. Рассмотрим несколько наиболее характерных работ.

В статье [Dunn et al. 2005] рассмотрен вопрос о родстве папуасских языков. Различные папуасские языки содержат очень мало слов, для которых может быть установлено их общее происхождение традиционным сравнительно-историческим методом. В результате приходится постулировать существование около полусотни генетически несвязанных семей, часть из которых содержат всего по 1–2 языка. Складывается впечатление, что достичь существенного прогресса в понимании эволюции папуасских языков традиционными методами не удастся. Нужно привлекать новые данные и новые методы.

Таблица 3

Расстояния между баскским и другими языками

	bsq	fre	spa	iri	rus	hin	fin	tur	abk	geo	lez	aeg	bma	bur	chk	grw	nav
bsq	0	54	50	57	50	43	52	42	45	44	45	55	60	44	48	44	50
fre	54	0	26	43	31	45	44	52	62	57	61	46	59	60	68	70	63
spa	50	26	0	44	35	48	47	56	58	53	62	46	52	56	59	62	62
iri	57	43	44	0	45	54	47	65	67	69	70	44	45	66	64	67	73
rus	50	31	35	45	0	40	36	48	61	48	56	48	52	55	58	66	65
hin	43	45	48	54	40	0	43	42	54	40	42	49	60	39	49	60	57
fin	52	44	47	47	36	43	0	43	62	55	54	56	59	60	55	54	67
tur	42	52	56	65	48	42	43	0	52	48	43	60	65	47	55	54	62
abk	45	62	58	67	61	54	62	52	0	55	47	55	57	48	62	55	50
geo	44	57	53	69	48	40	55	48	55	0	46	60	64	40	46	56	61
lez	45	61	62	70	56	42	54	43	47	46	0	69	69	41	49	51	66
aeg	55	46	46	44	48	49	56	60	55	60	69	0	36	56	62	64	66
bma	60	59	52	45	52	60	59	65	57	64	69	36	0	68	62	60	63
bur	44	60	56	66	55	39	60	47	48	40	41	56	68	0	51	53	58
chk	48	68	59	64	58	49	55	55	62	46	49	62	62	51	0	48	60
grw	44	70	62	67	66	60	54	54	55	56	51	64	60	53	48	0	45
nav	50	63	62	73	65	57	67	62	50	61	66	66	63	58	60	45	0

Для продвижения в этом направлении в [Dunn et al. 2005] используется грамматика папуасских языков, представленная в форме базы данных. Анализ осуществляется с помощью алгоритма Maximum parsimony. Этот алгоритм первоначально был разработан для нужд эволюционной биологии. Он позволяет восстановить дерево эволюции (как видов живых организмов, так и языков) исходя из предположения экономности эволюции. Алгоритм строит такое дерево эволюции, в котором число мутаций (изменений значений признаков от предка к потомку) будет минимально. Результат представлен на рисунке 5.

Примененная версия алгоритма строит некорневое дерево. Она дает информацию (как графическую, так и с некоторыми числовыми параметрами, поясненными в статье [Dunn et al. 2005]) о порядке отщепления языков и ветвей, т. е. о родстве папуасских языков.

Авторы статьи обсуждают вопрос – в какой мере результаты, полученные таким путем, являются достоверными?

Фактически здесь речь идет о совершенно новой методологии исследований в области исторической лингвистики. Достоверность результатов опирается на два основных момента: достоверность используемой базы данных и применимость для установления родства алгоритма Maximum parsimony.

Первый из них представляется не столь принципиальным. База данных по грамматике папуасских языков доступна через Интернет (www.sciencemag.org/cgi/content/full/309/5743/2072/DC1) и может быть тщательно протестирована. Ее корректность зависит от правильности описания грамматики папуасских языков. В случае обнаружения ошибок и корректировки базы данных результаты можно перепроверить, просто пересчитав заново.

Принципиально новой является идея установления родства с помощью специальных математических алгоритмов, типа Maximum parsimony. Эти алгоритмы имеют под собой серьезную теоретическую и эмпирическую основу. Теоретическая база, изложенная в монографиях [Semple, Steel 2003; Гасфилд 2003], к настоящему времени оформилась в самостоятельный раздел науки – математическую филогенетику. Ее результаты широко применяются в эволюционной биологии, а, с недавнего времени, и в лингвистике. Первые публикации появились около 10 лет назад [Warnow 1997], но особенно интенсивно стали применяться в последние годы [Nakhlen et al. 2005].

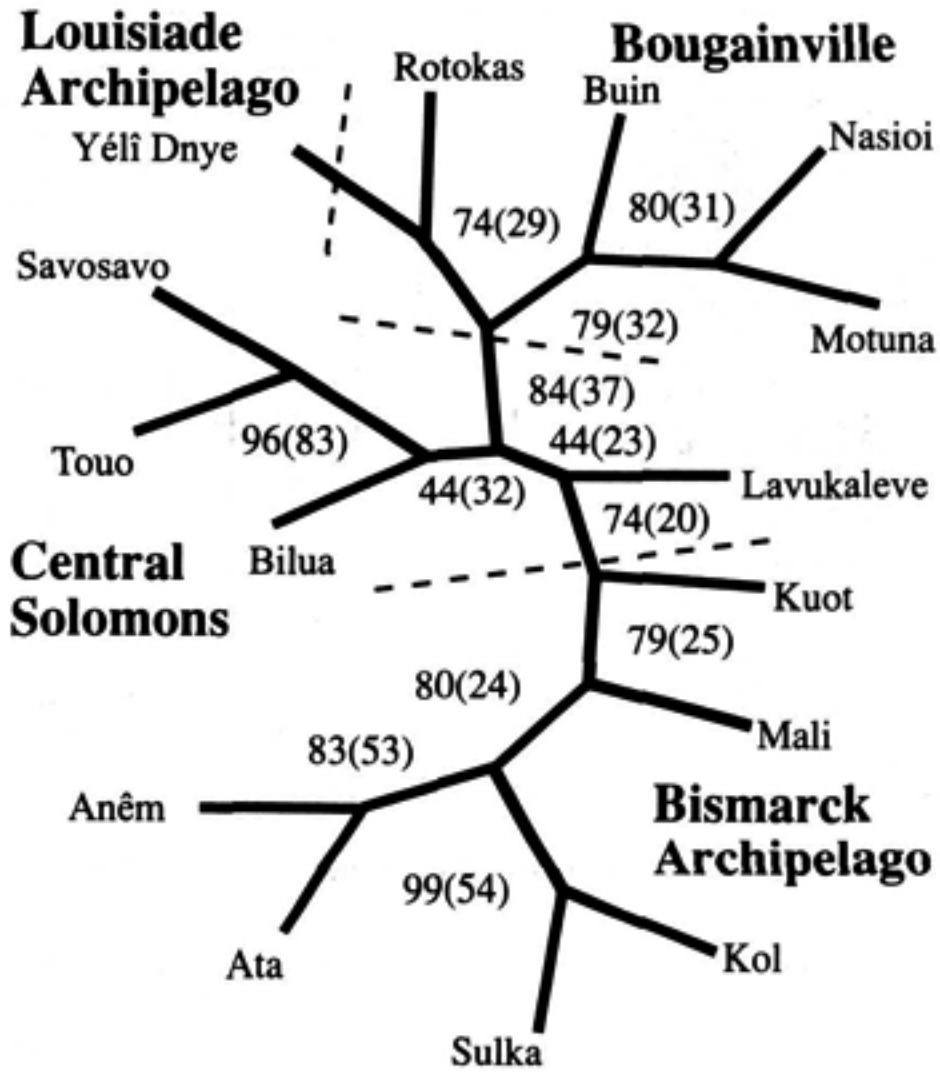

Рис. 5. Дерево эволюции папуасских языков Меланезийских островов (Science, № 309, 2005)

Авторы рассматриваемой работы для демонстрации корректности применения алгоритма Maximum parsimony приводят следующий пример. Они применяют его к некоторой группе австронезийских языков (для них также имеется типологическая база данных), классификация которых традиционным сравнительно-историческим методом уже установлена. Оказалось, что дерево, построенное алгоритмом по ТБД, очень близко к установленному классическими методами. Таким образом, есть основание полагать, что и в случае папуасских языков полученный результат, если и не является абсолютно точным, то, по меньшей мере, дает реалистичную картину их эволюции.

Maximum parsimony не является единственным алгоритмом, разработанным в математической филогенетике. Кроме того, существуют различные ТБД, а эти алгоритмы могут применять не только к грамматическим, но и к лексическим базам данных. Возникает закономерный вопрос – какие алгоритмы и применительно к каким базам данных дают лучшие результаты. В этой задаче есть и другие параметры, которые могут варьироваться – способ кодирования признаков (бинарный или не бинарный), мера близости языков и др. В начале развития любого научного направления методологические исследования особенно важны.

Поставленные вопросы исследовались в [Saunders 2005] на материале австронезийских языков и в [Wichmann, Saunders 2007] на материале языков американских индейцев.

В работе [Saunders 2005] показано, что лучшие результаты достигаются при совместном использовании лексических и грамматических (с бинарным кодированием) данных, и при применении алгоритма Bayes, основанного на статистическом методе Байеса.

Полученное алгоритмическим путем дерево приведено на рис. 6 справа. Можно отметить его близость к ‘известному’ дереву (рис. 6, слева). Однако есть одно важное отличие. В ‘известном’ дереве есть много ситуаций множественного ветвления, которые возникают в тех случаях, когда точную последовательность отщеплений в ходе эволюции различных ветвей традиционными методами установить не удается. Математический же алгоритм всегда «доводит дело до конца», предлагая полную эволюционную картину в

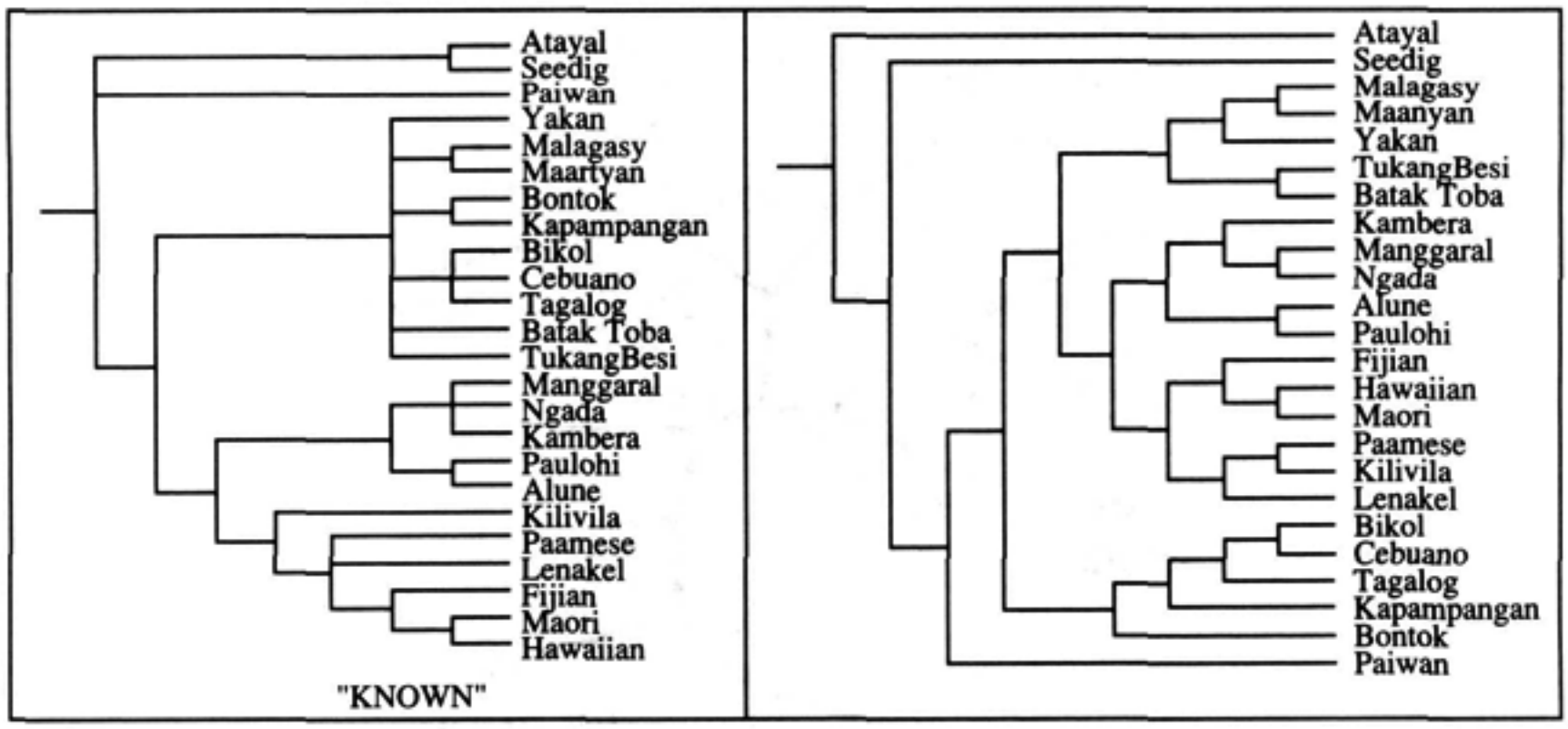

Рис. 6. Классификации австронезийских языков

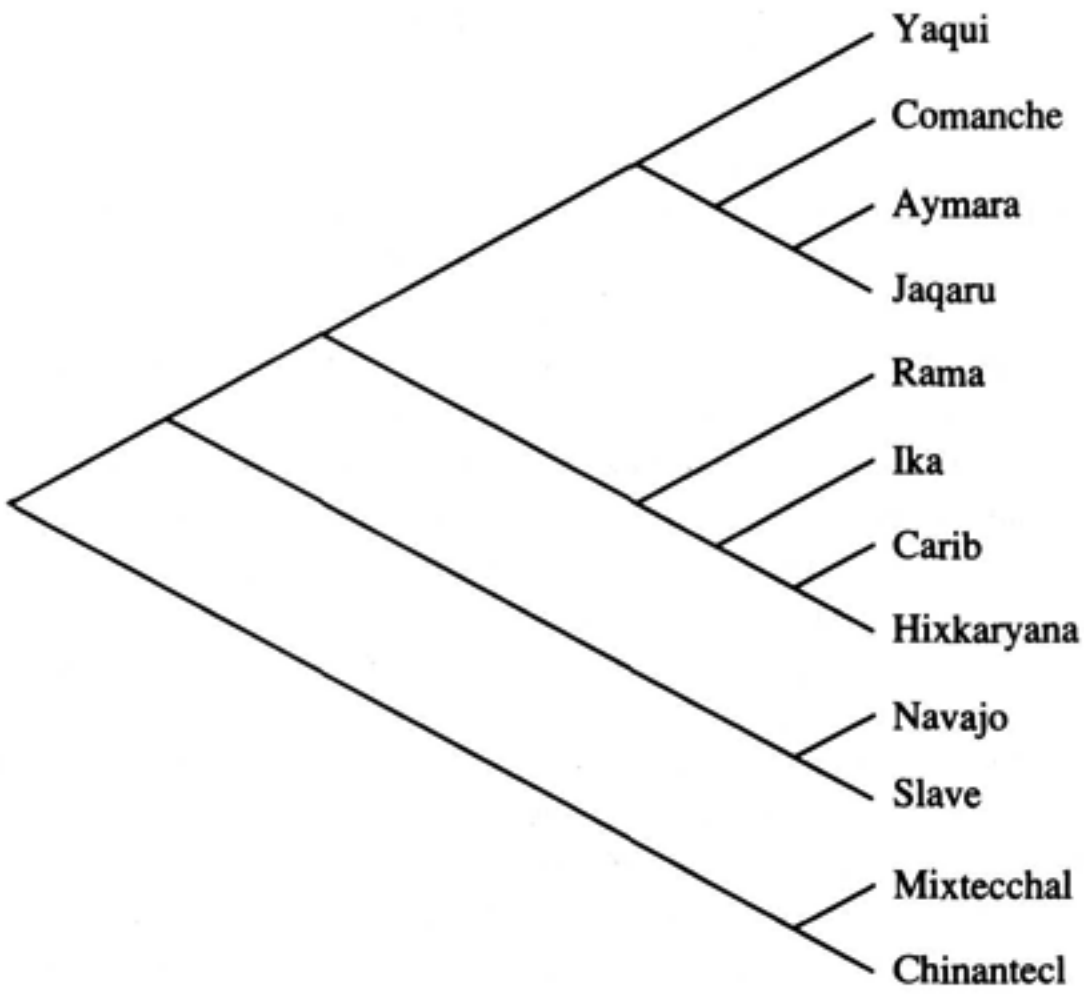

Рис. 7. Классификация индейских языков

виде дерева только с бинарным ветвлением. Можно считать, что таким образом выдвигаются некие гипотезы, которые затем могут проверяться традиционными методами.

В работе [Wichmann, Saunders 2007] рассмотрен следующий набор языков, состоящий из 6 пар языков различных семей. Athapaskan семья: Slave, Navajo; Chibchan семья: Rama, Ika; Aymaran семья: Jagaru, Aymara; Uto-Aztecan семья: Comanche, Yagui; Otomanguean семья: Chalcatongo Mixtec, Lealao; Chinantec Carib семья: Carib, Hixkaryana. Использована база данных WALS. Полученная в итоге картина представлена на рис. 7.

Только для 4 пар языков алгоритм верно определил их родство. Установлено, что лучшие результаты дают алгоритмы Bayes, Neighbor-joining, Neighbor Net. Вероятная причина ошибки алгоритма кроется в недостаточности данных по этим языкам в WALS. Впрочем, применение других математических алгоритмов может компенсировать недостаточность данных – в [Соловьев 2007] нами показано, что алгоритм UPGMA со спе-

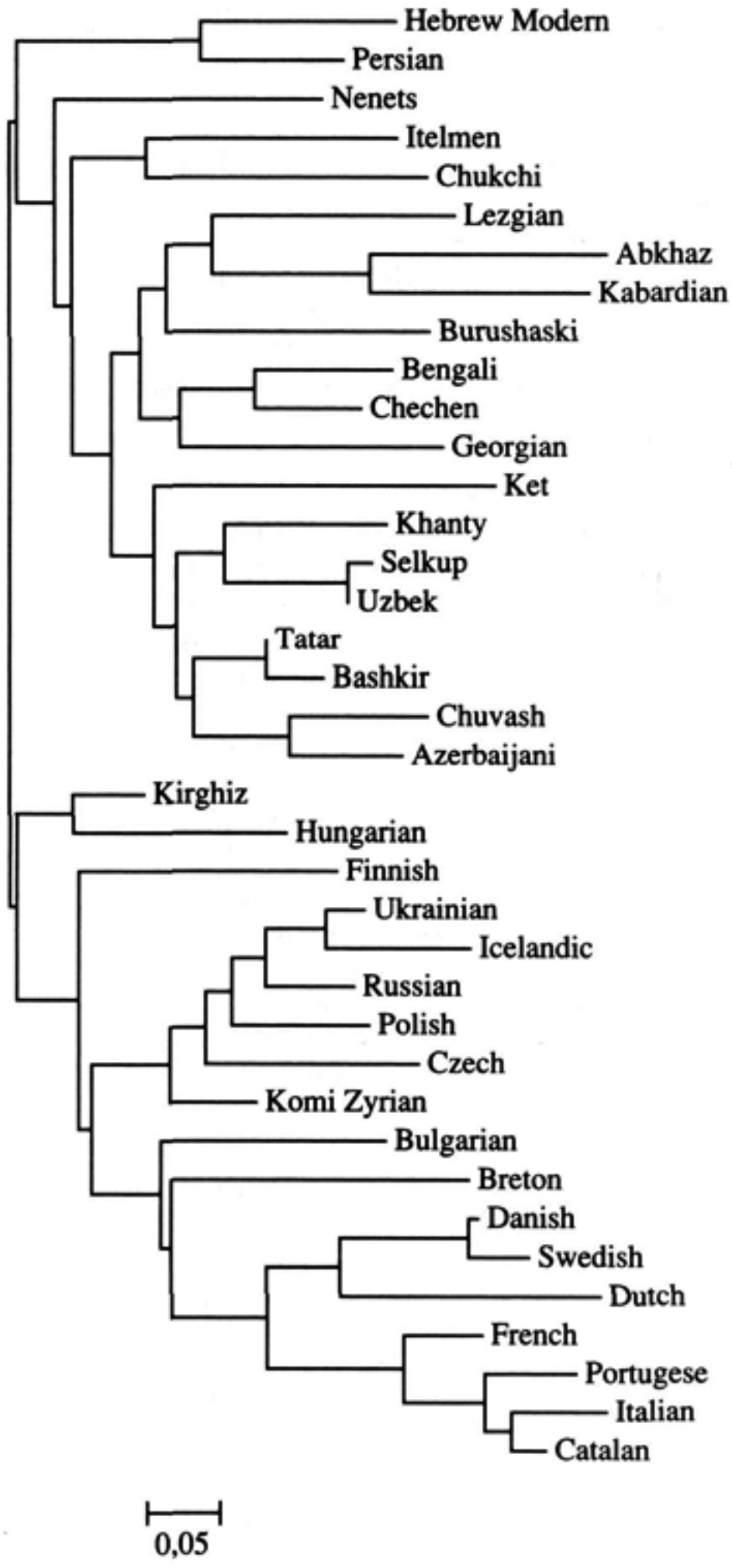

Рис. 8. Эволюционное дерево для 38 языков, построенное на основе WALS

циальной λ-метрикой дает разбиение выбранных 12 индейских языков на пары в точном соответствии с их родством.

В [Wichmann, Saunders 2007] поставлена важная проблема правильного выбора набора признаков. Получены первые результаты, указывающие на то, что выбор стабильных признаков способствует получению лучших результатов. Ситуация здесь в какой-то мере аналогична глоттохронологии, в которой сопоставляются не все слова языков, а только наиболее устойчивое ядро лексики – 200 или даже меньше слов [Бурлак, Старостин 2001].

Методологический характер носит и работа [Polyakov et al. 2009], в которой проведено сопоставление баз данных WALS и «Языки мира». Для сравнение выбраны

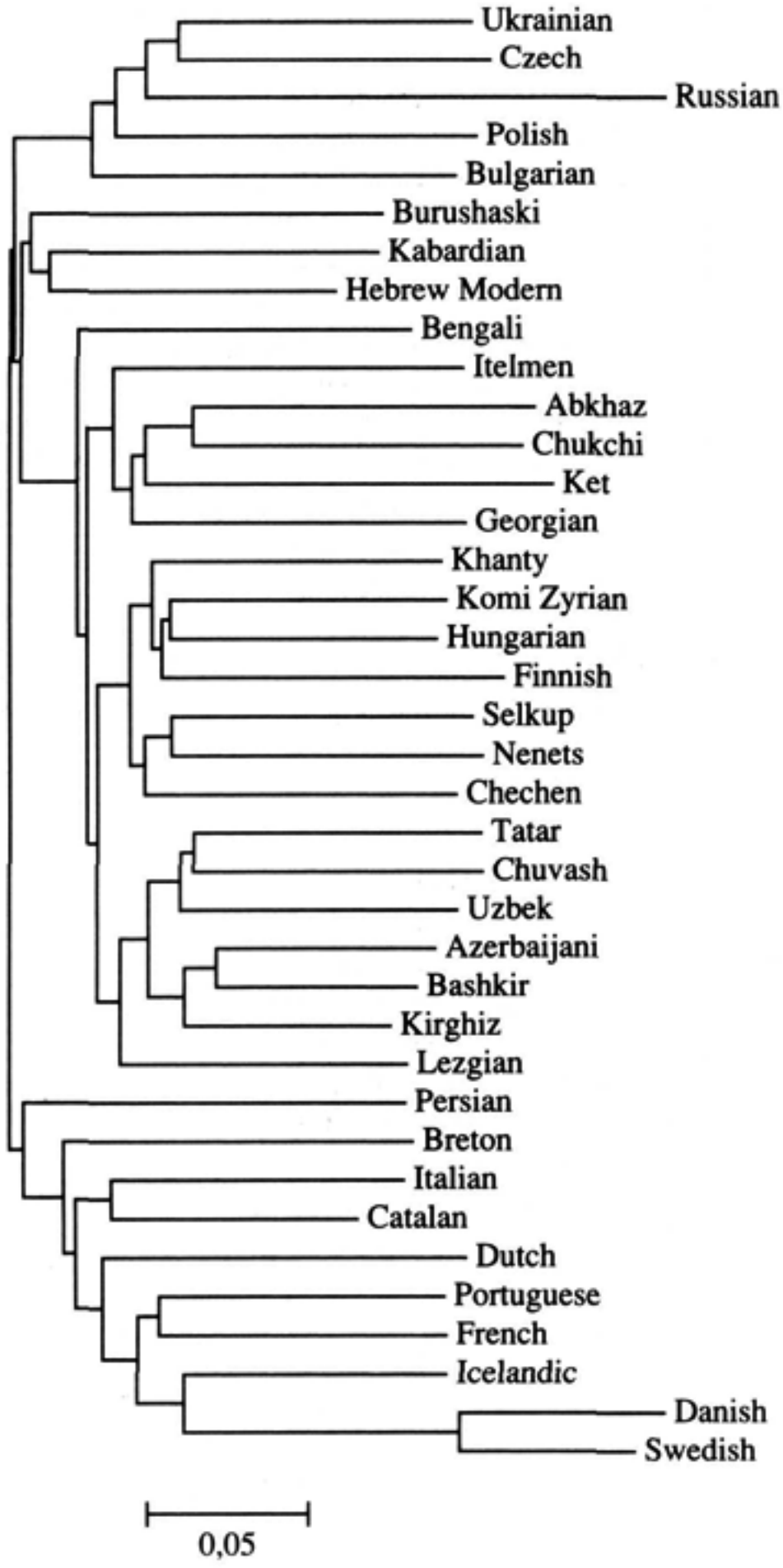

Рис. 9. Эволюционное дерево для 38 языков, построенное на основе БД «Языки мира»

38 языков, представленных в базе «Языки мира» и достаточно полно описанных WALS. С помощью алгоритма Neighbor-joining построены деревья, приведенные на рис. 8 и 9.

Дерево на основе WALS слишком сильно зашумлено и показывает результаты, далекие от реальной близости языков. Так, украинский оказывается наиболее близок к исландскому, венгерский к киргизскому и т. д. В дереве на основе БД «Языки мира» правильно выделено несколько групп языков (славянские, тюркские), результаты близки к действительным. В то же время в полученном дереве генетический сигнал смешивается с ареальным (перемешаны контактирующие группы германских и романских языков). Требуются дальнейшие кропотливые исследования, которые позволили бы разделить влияние родства и ареальных контактов.

4. ИСТОРИЧЕСКИЕ ИССЛЕДОВАНИЯ

К этой группе отнесены работы, целью которых являются исследования на основе лингвистических данных путей миграции при расселении человечества по планете.

В [Wichmann et al. 2009] изучается вопрос – сколько было заселений Америки? Единой общепризнанной точки зрения по этому вопросу нет. Существует несколько гипотез, число волн заселений Америки колеблется от 1 до 4 [Лимборская и др. 2002]. Последние данные генетических исследований как будто бы подтверждают версию с одним заселением.

Авторы [Wichmann et al. 2009] надеются ответить на этот вопрос, анализируя описания автохтонных языков Америки из WALS. Они применяют два метода, первый из которых оценивает степень типологического разнообразия языков, второй – степень стабильности характерных для этих языков признаков. Рассматриваются все автохтонные языки Америки, кроме эскимосских и на-дене, появившихся на американском континенте определенно позже.

В таблице 4 приведены 24 свойства, которые встречаются в Новом Свете значительно чаще, чем в Старом, причем распространены во всей Америке. После названия свойства указан номер соответствующего признака в WALS (f) и номер значения признака (v). Далее указан процент языков в обоих полушариях, в которых встречается это свойство.

Таблица 4

Характерные для языков Нового света типологические черты

Свойство WALS	Номер свойства/ значения	Частота в Старом Свете	Частота в Новом Свете
Интенсификаторы и рефлексивные местоимения различаются	47/2	16	81
Начальные вопросительные группы	93/1	9	64
Союзы и универсальные кванторы формально различаются	56/1	30	88
Отсутствие редукции при релятивизации субъектов	122/2	9	48
Эпистемическая возможность реализуется глагольными аффиксами	75/2	27	67
Отсутствие доминирующего порядка дополнения, обстоятельства и глагола	84/6	10	45
Реципрокная конструкция идентична рефлексивной	106/4	14	46
Порядок ‘глагол-подлежащее’	82/2	3	29
Контраст между назализованными и неназализованными гласными	10/1	9	37
Глагольное кодирование предикативных прилагательных	118/1	28	59
Малое число качественно различных гласных	2/1	8	34

Таблица 4 (окончание)

Свойство WALS	Номер свойства/ значения	Частота в Старом Свете	Частота в Новом Свете
Маркирование ситуационной и эпистемической модальностей не перекрывается	76/3	43	72
Местоименные подлежащие выражаются аффиксами на глаголе	101/2	42	69
Глагол 'давать' допускает конструкцию со вторичным объектом	105/3	12	35
Отсутствие контраста в произношении взрывных и фрикативных	4/1	27	33
Множественное число существительных выражается аналитическими показателями	33/7	3	21
Падежи выражаются послеложными клитиками	51/6	6	25
Эвиденциальность кодируется отдельной частицей	78/4	5	23
Местоименные подлежащие выражаются субъектными клитиками с вариативным расположением	101/3	0	12
Отсутствие передних огубленных гласных	11/1	86	98
Да/нет-вопросы требуют особого вопросительного морфологического оформления глагола	116/2	15	36
Только инклюзивное 1 л. мн.ч. имеет особое маркирование в глагольной морфологии	40/4	0	10

Широкое распространение этих редких для остального мира свойств можно объяснить либо эффектом основателя (это будет означать одно заселение Америки), либо заимствованиями. В последнем случае о числе заселений ничего сказать нельзя, но при этом придется постулировать возможность заимствований на гигантских расстояниях американского континента.

Авторы используют эффект пространственной автокорреляции, описанный ранее в [Holman et al. 2007]. Суть его состоит в том, что чем ближе языки расположены друг к другу географически, тем они в целом ближе и типологически. Построенные на основе данных WALS эмпирические кривые автокорреляции демонстрируют значительно большую степень типологической близости родственных языков (из одной семьи), по сравнению с неродственными на всех географических расстояниях.

Если бы языки американских индейцев происходили из одного источника, то кривая автокорреляции показывала бы большую близость языков, чем кривая, построенная для языков Старого Света (разделившихся значительно раньше). Авторы [Wichmann et al. 2009] нашли, что это действительно так, но разница не столь значительна, чтобы иметь доказательный характер.

Другой подход состоит в анализе стабильности выделенных выше 24 характерных для языков Америки свойств. Если эти свойства сохранились от общего предка всех

американских автохтонных языков, то они должны быть чрезвычайно стабильными (заселение американского континента происходило никак не менее 12 тыс. лет назад, а по многим оценкам и 30–40 тыс. лет назад). В то же время из предпринятого ранее в [Wichmann, Kamholz 2008] анализа стабильности свойств из WALS следует, что доля стабильных свойств в табл. 4 примерно такая же, как в полном списке свойств WALS. Это можно рассматривать как некоторый аргумент в пользу нескольких волн миграции. В целом авторы [Wichmann et al. 2009] делают заключение, что данных WALS недостаточно, чтобы сделать определенный вывод о числе заселений Америки.

5. ЗАКЛЮЧЕНИЕ

Большие типологические базы данных, такие как WALS и «Языки мира», являются не просто справочниками, но и могут быть эффективно использованы в типологических, ареальных и компаративистских исследованиях. Хотя эта область является очень молодой и методология использования ТБД не до конца отработана, все же приведенные в статье результаты указывают на перспективность данного направления и возможность получения новых, неожиданных результатов. Появляются и новые вопросы, которые могут быть поставлены только в контексте ТБД.

Одним из возможных применений ТБД является решение спорных вопросов родства языков. Интересна позиция в этом вопросе Дж. Николс. Ее современные взгляды [Nichols 2007] кратко можно резюмировать следующим образом:

– ограничения сравнительно-исторического метода вряд ли удастся преодолеть;

– типология может помочь проникнуть значительно глубже в прошлое, чем сравнительно-исторический метод;

– для получения лучших результатов с помощью типологии требуется лучшее понимание таких аспектов, как стабильность, независимость признаков, скорость изменений.

На основе анализа результатов первого этапа исследований можно сделать следующие общие выводы по применению ТБД в генеалогических исследованиях:

1. Для математической обработки данных целесообразно применять филогенетические алгоритмы.

2. В получаемых эволюционных деревьях отражаются и генетическая, и ареальная близость.

3. Лучшие результаты получаются при комбинировании различных (грамматических, лексических) данных.

4. В целом вопрос о выборе баз данных, алгоритмов, мер близости, наборов признаков и т. д. остается открытым.

Перспективным является также применение ТБД в ареальных и исторических исследованиях. Резюмируя можно отметить, что:

- ТБД дают новую перспективу в изучении взаимосвязей между языками в синхронном и диахронном аспектах,
- применение филогенетических алгоритмов и других математических методов впервые позволяет дать строгие ответы на многие вопросы.

Исследования находятся в самом начале, и очень многое еще предстоит сделать.

СПИСОК ЛИТЕРАТУРЫ

Бурлак, Старостин 2001 – *С.А. Бурлак, С.А. Старостин.* Введение в лингвистическую компаративистику. М., 2001.

Виноградов и др. 2003 – *В.А. Виноградов, А.И. Новиков, Е.И. Ярославцева.* База данных «Языки мира» как инструмент лингвистических исследований // ВЯ. 2003. № 3.

Гасфилд 2003 – *Д. Гасфилд.* Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология. СПб., 2003.

Лимборская и др. 2002 – *С.А. Лимборская, Э.К. Хуснутдинова, Е.В. Балановская.* Этногеномика и геногеография народов Восточной Европы. М., 2002.
Поляков, Соловьев 2006 – *В.Н. Поляков, В.Д. Соловьев.* Компьютерные модели и методы в типологии и компаративистике. Казань, 2006.
Соловьев 2007 – *В.Д. Соловьев.* Задачи и методы лингвистической филогенетики. Труды конф. «Знания. Онтологии. Теории». Новосибирск, 2007.
Comrie 2007 – *B. Comrie.* Areal typology of mainland Southeast Asia: what we learn from the WALS maps // Pranee Kullavanijaya (ed.). Trends in Thai linguistics. Bangkok, 2007.
Comrie 2008 – *B. Comrie.* Basque, Romance, and areal typology: What do we learn from the World atlas of language structures? // H.-J. Döhla, R. Montero Muñoz, F. Báez de Aguilar González (eds.). Lenguas en diálogo: El iberromance y su diversidad lingüística y literaria. Madrid; Frankfurt-am-Main, 2008.
Comrie 2009 – *B. Comrie.* Maltese and the World atlas of language structures // B. Comrie. et al. (eds.). Introducing Maltese linguistics. Papers from the 1st International Conference on Maltese linguistics. Amsterdam; Philadelphia, 2009.
Comrie, Cysouw 2006 – *B. Comrie, M. Cysouw.* New Guinea through the eyes of WALS. Language and linguistics in Melanesia // http://email.eva.mpg.de/~cysouw/publications.html. 2006.
Cysouw, Comrie 2009 – *M. Cysouw, B. Comrie.* How varied typologically are the languages of Africa? // R. Botha, Ch. Knight (eds.). The cradle of language. V. 2. Oxford, 2009.
Dunn et al. 2005 – *M. Dunn, A. Terrill, G. Reesink, R. Foley, S. Levinson.* Structural phylogenetics and the reconstruction of ancient language history // Science. V. 309. № 5743. 2005.
Everaert, Musgrave 2009 – *M. Everaert, S. Musgrave* (eds.). The use of databases in cross-linguistic studies. Berlin, 2009.
Haspelmath et al. 2005 – *M. Haspelmath, M. Dryer, D. Gil, B. Comrie* (eds.). The world atlas of language structures. Oxford, 2005.
Holman et al. 2007 – *E. Holman, Ch. Schulze, D. Stauffer, S. Wichmann.* On the relation between structural diversity and geographical distance among languages: observations and computer simulations // Linguistic typology. V. 11. № 2. 2007.
Nakhlen et al. 2005 – *L. Nakhlen, D. Ringe, T. Warnow.* Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages // Language. V. 81. 2005.
Nichols 1992 – *J. Nichols.* Linguistic diversity in space and time. Chicago; London, 1992.
Nichols 2007 – *J. Nichols.* Typology in the service of classification. http://aalc07.psu.edu/papers/jn_typol_class3.pdf. Stanford, 2007.
Polyakov et al. 2009 – *V. Polyakov, D. Solovyev, S. Wichmann, O. Belyaev.* Using WALS and Jazyki Mira // Linguistic typology. V. 13. № 1. 2009.
Saunders 2005 – *A. Saunders.* Linguistic phylogenetics of the Austronesian family. B.A. Thesis. Swarthmore College, 2005.
Semple, Steel 2003 – *C. Semple, M. Steel.* Phylogenetics. New York, 2003.
Warnow 1997 – *T. Warnow.* Mathematical approaches to comparative linguistics // Proceedings of the National Academy of Scienes. 1997. V. 94.
Wichmann, Saunders 2007 – *S. Wichmann, A. Saunders.* How to use typological database in historical linguistic research // Diachronica. V. 24. № 2. 2007.
Wichmann et al. 2009 – *S. Wichmann, E. Holman, D. Stauffer, C. Brown.* Similarities among languages of the Americas: An exploration of the WALS evidence. http://email.eva.mpg.de/ ~wichmann/SantaBarbWichmannRevSubmit.pdf. 2009.
Wichmann, Kamholz 2008 – *S. Wichmann, D. Kamholz.* A stability metric for typological features // STUF. Language typology and universals. V. 61. № 3. 2008.