

Казанский государственный университет
Институт языкознания РАН
Российский научно-образовательный центр по лингвистике
им. И. А. Бодуэна де Куртенэ

Поляков В.Н., Соловьев В.Д.

Компьютерные модели и методы
в типологии и компаративистике

Казань
Казанский государственный университет
2006

УДК 81:004.9
ББК 81:32.973
П 54

Печатается по решению редакционно-издательского совета
Казанского государственного университета

Поляков В.Н.

П 54 Компьютерные модели и методы в типологии и компаративистике. Монография / В.Н. Поляков, В.Д. Соловьев. – Казань: Казанский государственный университет им. В.И. Ульянова-Ленина, 2006.–208 с.

ISBN 5-98180

В монографии в систематической форме излагаются результаты, полученные применением компьютерных и математических методов к анализу Базы данных «Языки мира». База данных содержит описание 315 языков, в основном, Европы, Северной и Центральной Азии по более, чем 3800 параметрам. Описаны компьютерные методы, которые позволяют оценивать степень близости структуры языков и могут быть применены для уточнения генетической классификации. Построена математическая модель распространения языковых признаков. Монография может представлять интерес как для лингвистов различных специальностей (типология, компаративистика), так и для математиков и специалистов по компьютерным технологиям, интересующихся применением математических методов в гуманитарных науках.

УДК 81:004.9
ББК 81:32.973

ISBN 5-98180

© Казанский государственный университет, 2006
© Поляков В.Н., Соловьев В.Д., 2006

Содержание

Введение	7
Глава 1. ТИПОЛОГИЯ И КЛАССИФИКАЦИЯ ЯЗЫКОВ	11
1.1. Способы классификации языков	11
1.2. Корреляции между различными классификациями	15
1.3. Математические методы	21
1.4. Базы данных	24
1.5. Резюме по главе 1	27
Глава 2. БАЗА ДАННЫХ «ЯЗЫКИ МИРА»	27
2.1. История создания БД	27
2.2. Характеристики контента Базы данных «Языки мира»	27
2.3. Принципы, положенные в основу Базы данных «Языки мира» ИЯ РАН	32
2.3.1. Бинарный принцип	32
2.3.2. Иерархический принцип	32
2.3.3. Принцип парадигмы	36
2.3.4. Формат представления данных: аргументы <i>ЗА</i> и <i>ПРОТИВ</i>	36
2.4. Характеристики модели реферата	38
2.5. Резюме по главе 2	41
Глава 3. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ БАЗЫ ДАННЫХ	43
3.1. Основные задачи, решаемые с помощью ядра БД	43
3.2. Windows-версия ядра БД	43
3.3. Инсталляция БД	43
3.4. Работа с Базой данных «Языки Мира»	43
3.5. Парное сравнение языков на уровне классов и целого реферата	44
3.6. Поиск строк в рефератах	45
3.7. Работа с моделью и рефератом языка	47
3.7.1. Команды главного меню	47
3.7.2. Загрузка реферата в БД из файла	49
3.7.3. Сохранение реферата в файл	49
3.8. Навигация в пределах модели и реферата	49
3.9. Разработка Web-версии БД	49
3.9.1. Просмотр сайта	50
3.9.2. Просмотр реферата языка	51
3.9.3. Сравнение рефератов	52
3.10. Экспертиза БД	52
3.11. Назначение БД	52
3.11.1. Квантитативные исследования	52
3.11.2. Справочные цели	53
3.11.3. Учебные цели	54
3.12. Резюме по главе 3	54

Глава 4.	НОВЫЕ ВОЗМОЖНОСТИ КВАНТИТАТИВНЫХ ИССЛЕДОВАНИЙ	55
4.1.	Расчет мер парной близости языков	55
4.1.1.	Подходы к расчету мер близости	55
4.1.2.	Учет структуры и объема признакового пространства	56
4.1.3.	Методика проверки расчетов	61
4.1.4.	Результаты расчетов мер	62
4.1.5.	Предварительные обсуждения результатов	69
4.2.	Кластеризация	70
4.2.1.	Общие сведения о задаче кластеризации	70
4.2.2.	Кластеризация языков	71
4.2.3.	Кластеризация признаков	73
4.2.4.	Проверка результатов методом Data Splitting	74
4.2.5.	Архитектура программного комплекса	74
4.2.6.	Выявление возможных ошибок в данных	76
4.2.7.	Расчет частот встречаемости признаков в выбранной группе языков	77
4.3.	Интеллектуальный анализ данных	77
4.4.	Обсуждение предварительных результатов кластеризации и интеллектуального анализа данных	83
4.5.	Резюме по главе 4	83
Глава 5.	МОДЕЛЬ СТРУКТУРНОЙ ЭВОЛЮЦИИ ЯЗЫКОВ	85
5.1.	Общие соображения	85
5.2.	Заимствования vs. параллельная эволюция	87
5.3.	Диаграммы распределения частотности признаков	89
5.4.	Анализ математической модели динамики признаков	96
5.5.	Резюме по главе 5	102
Глава 6.	РАЗРАБОТКА НОВЫХ МЕТОДОВ ВЕРИФИКАЦИИ ГЕНЕТИЧЕСКИХ ГИПОТЕЗ	103
6.1.	Расчет квантитативных портретов по языковым семьям и ветвям	103
6.2.	Классификация признаков	103
6.3.	Пример расчета квантитативного портрета по ИЕ-языкам по разделу «Сложное предложение»	105
6.4.	Автоматическое выявление признаков	106
6.5.	Методика проверки генетических гипотез	106
6.6.	Резюме по главе 6	117
Глава 7.	ФЕНОМЕН ТИПОЛОГИЧЕСКОГО СДВИГА	119
7.1.	Природа языковых универсалий	119
7.2.	Природа языковых раритарий	124
7.3.	Феномен типологического сдвига	129
7.4.	Резюме по главе 7	130

ЗАКЛЮЧЕНИЕ		131
Литература		133
Приложение А.	Таблица А.1. Генетическая принадлежность языков БД	136
Приложение Б.	Таблица Б.1. Перечень классифицирующих признаков, представленных в БД	160
Приложение В.	Реферат шведского языка	171
Приложение Г.	Результат сравнения рефератов датского и шведского языков	178
Приложение Д.	Спецкурс «Базы данных для типологических и компаративистских исследований»	183
Приложение Е.	Таблица Е.1. Квантитативная карта признаков раздела «2.5.4. Сложное предложение» для индоевропейских языков	187
Приложение Ж.	Динамика статистических универсалий	190
Приложение З.	Рис. 3.1. Распределение используемых в одном из расчетов 11 мертвых языков по оси времени	202
Приложение И.	Условные обозначения	203
Приложение К.	Предметный указатель	204
Приложение Л.	Именной указатель	205

ВВЕДЕНИЕ

Компаративистика и типология являются традиционными разделами лингвистики, берущими свое начало в древности и активно развиваемые с XIX-го века. Целью исследований в этих разделах является классификация языков и отдельных языковых конструкций, в том числе, на генетической основе, т.е. с установлением общего происхождения языков. В этих направлениях удалось достичь значительных успехов благодаря открытию и систематическому применению сравнительно-исторического метода. Также очень важным явилось введение в лингвистические исследования данных по большому числу редких и исчезающих языков, часто благодаря подвижнической полевой работе.

Накопление огромного числа данных, а по последним оценкам на Земном шаре насчитывается немногим менее 7 тысяч языков, заставляет обратиться к компьютерным методам их хранения и обработки. Основной формой хранения цифровых данных являются базы данных, из которых наиболее удобны реляционные. Различные специализированные базы лингвистических данных используются уже достаточно давно, однако базы данных, содержащие достаточно полные описания большого числа языков, появились только в последние годы в результате кропотливого труда многих исследователей.

Почти одновременно стали доступны для исследователей база данных **WALS** – **The World Atlas of Language Structures**, созданная международным коллективом под руководством М. Хаспельмата, М. Драйера, Д. Гила и Б. Комри, и база данных «Языки мира» – разработка Института языкознания РАН. Следует отметить, что хотя принципы, положенные в основу архитектуры этих баз данных, различны, обе они созданы на основе многочисленных книг и статей, содержащих детальные описания тысяч языков, и, таким образом, аккумулируют результаты исследований нескольких поколений лингвистов. В этом их несомненная ценность.

В данной монографии исследования проводятся на материале базы данных «Языки мира» (далее, сокращенно, БД), и поэтому остановимся вкратце на истории ее создания и основных параметрах.

Разработка концепции БД была начата в 80-е годы практически одновременно с реализацией проекта по созданию энциклопедии «Языки мира», осуществляемого под руководством члена-корреспондента РАН В.Н. Ярцевой. В качестве основного средства формализации была разработана так называемая «модель реферата», представляющая собой универсальную схему описания языка, где языковые явления и категории представлены в жестко заданной форме. БД содержит описание следующих сфер языка: фонетика, морфология, синтаксис.

Основную работу по заполнению БД выполнила Е.И. Ярославцева. Были созданы программные оболочки сначала для DOS, а затем для Windows. В настоящее время в рамках Российского научно-образовательного Центра по лингвистике им. И.А. Бодуэна де Куртенэ создан пилотный вариант Web-версии БД. Он размещен в Интернете по адресу: www.dblang.ru и с 2006 г. доступен в тестовом режиме. В ИЯ РАН под руководством А.А. Кибрика проведены работы по широкомасштабному тестированию БД.

По состоянию на конец 2006 г. БД характеризуется следующими основными параметрами: 315 языков (в основном, Европы, Северной и Центральной Азии), 3821 признак, бинарный формат представления данных.

Бинарность представления означает, что для каждого языка и для каждого признака в БД 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 x 3821, содержащую более миллиона бит информации.

Важным аспектом использования компьютерных баз данных является не только удобство хранения и доступа к данным, но и возможность применения разнообразных математических и статистических методов исследования данных. С увеличением объема информации применение математических методов анализа становится все более насущным.

Видимо, первая попытка использования статистики сделана в хорошо известной монографии Дж. Николс 1992 г. «Linguistic Diversity in Space and Time». Несмотря на ограниченность данных, Дж. Николс удалось получить общую картину распространения языков при расселении человечества по Земному шару. База данных «Языки мира» содержит примерно на два порядка больше информации, что позволяет рассчитывать на получение более детальной и содержательной картины.

В настоящей монографии предложен комплекс математических методов анализа многомерного пространства языковых данных. Конечной целью является разработка таких методов анализа, которые позволили бы на основе всей совокупности данных решать вопросы происхождения языков, идентифицировать близость языков как генетическую или возникшую в результате заимствований, выявлять общие закономерности возникновения и распространения языковых структур.

В первой главе настоящей монографии приводится характеристика основных подходов к классификации языков, перечисляются главные результаты, полученные в недавних работах с применением математических методов, и основные проблемы, стоящие на пути исследователей.

Во второй и третьей главах дается детальное описание структуры Базы данных «Языки мира» и сопутствующего ей инструментария. Описываются методы работы с БД.

В четвертой главе приводится ключевой для развиваемого подхода математический аппарат для расчета мер близости языков. В традиционных типологических исследованиях языки классифицируются на основе небольшого числа параметров. Предложенные меры близости опираются на всю имеющуюся совокупность данных и используют строгий математический аппарат. Это дает новую перспективу в классификации языков. Однако выбор адекватной меры близости сам по себе оказался сложной проблемой. В третьей главе дается описание общего теоретического аппарата и методики выбора подходящей меры близости. Для расчета мер близости и классификации языков используется широкий спектр методов кластеризации, а также ординация с использованием статистического пакета R.

В пятой главе строится математическая модель распространения языковых признаков. В ее основе лежит ряд постулатов, сформулированных на основе анализа БД и хорошо коррелирующих с постулатами эволюции лексики (М. Сводеша). Введено новое понятие ЯП-диаграмм (отражающих зависимость числа признаков от степени их распространенности), оказавшееся мощным инструментом анализа механизмов заимствований.

В шестой главе вводится новая методология анализа генетической близости языков, основанная на частотах характеристических языковых признаков. Ранее, в работах по типологии использовались такие термины, как уникалии, универсалии, в том

числе, статистические. В докторской диссертации Е.И. Ярославцевой введено понятие раритарии. Эти термины уточнены применительно к использованию с БД и для целей данного исследования.

В седьмой главе сопоставляются данные по современным и мертвым языкам. Развитая выше техника позволяет выявить глобальные тенденции языкового развития. В частности, обнаружено явление типологического сдвига, состоящее в повышении частоты высокочастотных признаков и вымывании низкочастотных. В этой же главе демонстрируется применение предложенных методов к проверке гипотез о генетическом родстве языков.

Авторы выражают свою благодарность директору ИЯз РАН, члену-корреспонденту РАН В.А. Виноградову за плодотворное сотрудничество в работе над БД, заведующему отделением теоретической и прикладной лингвистики (филфак МГУ), члену-корреспонденту РАН А.Е. Кибрику за предоставленную возможность прочитать спецкурс по тематике БД, декану факультета «Лингвистика» Южно-Уральского университета Т.Н. Хомутовой за такую же возможность в ЮУрГУ, руководителю сектора «Языки мира» А.А. Кибрику за организацию экспертизы БД и плодотворные дискуссии, всем сотрудникам ИЯз РАН – участникам проекта.

Глава 1. ТИПОЛОГИЯ И КЛАССИФИКАЦИЯ ЯЗЫКОВ

1.1. Способы классификации языков

Объект исследований в лингвистике – человеческие языки – в определенном смысле близок к объекту исследований в биологии – живым организмам. Во всяком случае, и языки, и живые организмы являются естественными образованиями, приобретенными известным нам сейчас вид в результате длительной эволюции. И языки, и живые организмы характеризуются, с одной стороны, громадным разнообразием, а с другой стороны, в основе и тех, и других лежит небольшое число фундаментальных структур. Для живых организмов это генетический код, для языков – так называемые, языковые универсалии. В последние десятилетия в результате интенсивных исследований в типологии (Croft 1990) удалось существенно продвинуться в описании и понимании универсалий. Однако наиболее общие вопросы остаются пока без ответа. Например, неясно, существуют ли специальные «языковые» гены, обуславливающие эти универсалии, или их существование является следствием более общих когнитивных законов. Содержательную дискуссию по этому вопросу можно найти в (Givón 2002).

Для лингвистики, как и для биологии, важными являются следующие две задачи: построение классификации изучаемых объектов и описание механизмов их эволюции. Хотя языков значительно меньше (в настоящее время около 6 тысяч), чем различных видов живых организмов, их классификация оказалась более сложной задачей. Если в биологии существует общепризнанная исчерпывающая классификация, в общих чертах созданная еще более 200 лет назад К. Линнеем, то в лингвистике такой единой классификации до сих пор нет.

Можно выделить три подхода к классификации языков: генеалогический, ареальный и типологический.

В рамках генеалогического подхода развит сравнительно-исторический метод, позволяющий реконструировать предков современных языков. Сущность метода состоит в установлении регулярных соответствий между родственными языками и реконструкции на их основании праязыкового состояния. Согласно (Бурлак, Старостин 2001) «Язык А является предком языка В, если а) все фонемы (или их подавляющее большинство) в языке В выводятся из фонем языка А; б) некоторая часть морфем (в том числе грамматических и лексем языка В выводятся из морфем и лексем языка А», причем «некоторая часть» лексики понимается как стандартный 100-словный список Сводеша (Сводеш 1999). В этом направлении удалось добиться значительных успехов, четко описав ряд семей (индоевропейская, уральская и др.), языки каждой из которых восходят к общему предку (к протоиндоевропейскому языку, протоуральскому и т. д.). В итоге строится дерево происхождения языков, аналогичное таксономической классификации биологических видов. Основоположником данной концепции является А. Шлейхер, после работ которого (Schleicher 1871) идея «родословного дерева» получила широкое распространение.

Благодаря красивой идее М. Сводеша (Сводеш 1999), часто удается даже определить время расхождения двух языков. Методика основана на открытом М. Сводешем постоянстве скорости изменения базовой части лексики. Это позволяет определить время, прошедшее с момента разделения языков, просто подсчитав количество в этих языках общих слов в базовой части лексики (100-словный список Сводеша) и применив простую формулу.

Однако сравнительно-исторический метод хорошо применим лишь к верхним уровням дерева, проследить его структуру ближе к корню (праязыку всего человечества) не удастся. Уже при попытке установления родства индоевропейских, уральских и алтайских семей возникли споры. Некоторые лингвисты принимают гипотезу существования ностратической макросемьи и ностратического протоязыка, из которого и произошли протоиндоевропейский, протоуральский, протоалтайский (и некоторые другие). Большинство же считают ее недостаточно обоснованной.

Общепринятой является точка зрения, что реконструкция возможна на глубину не более 10 тыс. лет. Вероятно, такое положение дел объясняется не столько недостаточной разработанностью сравнительно-исторического метода, сколько его принципиальной ограниченностью, по крайней мере, в современной форме. Так, в (Атлас 1998) высказывается мнение, что гипотезы типа ностратической никогда не будут доказаны или отвергнуты на основе только лингвистических данных (остаются надежды на археологические данные и результаты исследований генотипов различных народов. Впрочем, по мнению С.А. Старостина, являющегося подвижником идеи сверхглубокой реконструкции (Бурлак, Старостин 2001), и возможности стандартного сравнительно-исторического метода еще не исчерпаны).

С помощью сравнительно-исторического метода не всегда удается однозначно определить генетическую принадлежность и современных языков. Так, японский язык иногда включается в алтайскую семью (Бурлак, Старостин 2001, с.180), иногда в австронезийскую (Бугаева 2006), иногда выделяют в отдельную группу (изолят). Есть и еще целый ряд изолятов (нивхский, бурушаски и др.), происхождение которых совершенно не ясно.

Более того, сравнительно-исторический метод дает сбои не только на экзотических или протоязыках, но и на детально изученных современных языках. Так, до сих пор нет однозначного ответа на вопрос о существовании протобалтославянского языка. Споры идут с позапрошлого века, в них приняли участие классики лингвистики – И.А. Бодуэн де Куртэнэ, отрицавший существование такого протоязыка (Бодуэн 1903), А.А. Шахматов, наоборот, признававший его существование (Шахматов 1915) и др. Обзор последующей дискуссии по этому вопросу можно найти в (Балалыкина 2001). Неясны причины схожести уральских и алтайских языков (которые раньше рассматривались как одна семья (Лингвистический 1990)) – то ли действительно общность происхождения, то ли массивированные заимствования в результате длительных контактов.

Причина такого положения дел, на наш взгляд, в ряде факторов. Прежде всего, это отсутствие строгих критериев истины в данной области лингвистики. На основе одного и того же набора фактов разные исследователи делают разные выводы, и вопрос либо вообще повисает в воздухе, либо решается мнением большинства.

Далее, недостатком сравнительно-исторического метода является его ограниченность фактически только областью фонетики и лексики (частично, морфологии). Явно недостаточное внимание уделяется синтаксису. Так, в (Бурлак, Старостин 2001) реконструкции синтаксиса уделено менее двух страниц. Авторы пишут: «Реконструкция праязыкового синтаксиса представляет большую трудность, чем реконструкция морфологии. Это связано не только с тем, что морфологические единицы (морфемы и словоформы) сохраняются в языках, а синтаксические единицы (словосочетания и предложения) – нет, но и с тем, что синтаксис многих языков описан гораздо хуже, чем морфология» и, далее, «подход к реконструкции синтаксиса с учетом данных ти-

пологии должен быть очень плодотворен. Такого рода попытки делаются уже сейчас, например, В. Дресслером ((Дресслер 1988), добавлено нами), У. Леманном и др. Но достигнутые ими результаты во многом спорны. Причина этого – в том, что сама синтаксическая типология возникла сравнительно недавно, и, как во всякой молодой науке, в ней гораздо больше спорного, нежели установленного достоверно».

Наконец, в большинстве современных работ рассматривается весьма ограниченный круг данных. Отсутствует методология синтеза и одновременного учета всех аспектов сопоставляемых языков. Применительно к проблеме балтославянского единства на это обращено внимание в (Балалыкина 2001): «Необходимо детальное исследование и последующее сопоставление всех языковых уровней отдельных славянских и балтийских языков».

Другой подход – ареальный – объединяет языки в «языковые союзы» географически близко расположенных языков на основе сходства, возникшего в результате взаимовлияний в пределах ареала. Хорошо известен балканский языковой союз (Атлас 1998, с. 34), изучались и некоторые другие, но систематической классификации всех языков на этой основе никогда не было предложено.

Наконец, классификацией языков занимается и лингвистическая типология, ставящая во главу угла структурные свойства языков. На этом пути возникает сразу целый ряд классификаций. Г.А. Климов еще в 1980 г. писал: «В отличие от единственных в своем роде генеалогической и ареальной классификаций языков мира, в типологии число классификаций уже в настоящее время достаточно внушительно и продолжает увеличиваться» (Климов 1980, с. 14). Проблема в том, что неясно, какие именно свойства следует положить в основу классификации. Многие свойства языков представляются с этой точки зрения важными: степень аналитизма, строй языка, место маркировки (на зависимом слове или управляющем) синтаксических отношений, порядок слов и т.д.

Более того, даже после выбора одного из наборов свойств, не удается построить общепринятой однозначной классификации. Например, по степени аналитизма В. Гумбольдт выделял флективные, агглютинативные и изолирующие языки, а Ф. Шлегель – только аффиксирующие и флективные. Определение строя языка – номинативного, эргативного и т.д. чаще всего не вызывает трудностей. Однако, например, филиппинский язык капампанган Е. Кинан (Keenan, 1985) трактуется как номинативный язык, а М. Митун (Mithun, 1993) – как эргативный.

Неудовлетворительность классификаций основанных на небольшом числе признаков достаточно давно осознана лингвистами. Приведем несколько цитат. Р. Якобсон (Якобсон 1963, с. 97-98): «Не перечень элементов, но система является основой для типологии... Лингвистическая типология языков, основанная на произвольно выбранных признаках, не может дать удовлетворительных результатов». Г.А. Климов (Климов 1980, с. 18): «Широко признанная в современном языкознании неудовлетворительность традиционной типологической (так называемой морфологической) классификации объясняется, в частности, тем, что она оперирует не целостными языковыми типами, а в лучшем случае так называемыми “типами в языке”».

Следствием того, что не удастся создать единую классификацию языков, являются признание существующих классификаций взаимодополняющими друг друга и изучение корреляций между ними.

Приведем обзор генетических гипотез, которые могут быть проверены с помощью БД. Составлено по материалам (Языки мира 1993...2006).

I. Гипотезы верхнего уровня

К этому уровню отнесем гипотезы наиболее широкого объединения языковых семей в макросемьи. Для охваченного БД региона, приблизительно соответствующего Евразии (Европа + Северная и Центральная Азия), есть три гипотезы этого уровня.

Евроазиатская гипотеза объединяет три основные семьи этого региона – индоевропейскую, уральскую и алтайскую в единую евроазиатскую макросемью. Эта гипотеза является более или менее признанной, хотя, по мнению таких авторитетных типологов, как Б.Комри, вряд ли когда-нибудь будет доказана.

Ностратическая гипотеза добавляет к ним также дравидийские, картвельские и афроазиатские языки. Дж. Гринберг присоединяет к ним еще и эскимосско-алеутскую семью.

Дене-кавказская (синокавказская) гипотеза объединяет языки Северного Кавказа (абхазско-адыгейская и нахско-дагестанская семьи), енисейские, сино-тибетские языки и северо-американскую семью на-дене. На имеющемся в БД материале можно проверить близость енисейских языков и двух северо-кавказских семей.

II. Гипотезы среднего уровня

Отнесем к ним предположения о родстве двух языковых групп. В первую очередь нуждаются в проверке гипотезы о родстве уральских и алтайских языков, абхазско-адыгейских и нахско-дагестанских языков, а также уральских и эскимосско-алеутских языков.

III. Гипотезы нижнего уровня

К этой группе отнесем гипотезы о принадлежности (близости) отдельных языков к тем или иным семьям. Наиболее сложными являются вопросы о происхождении языков-изолятов: баскского, нивхского, юкагирского, бурушаски. Вполне возможно, что это осколки древних языковых семей.

Считается, что баскский язык наиболее близок к кавказским языкам: картвельским и/или северо-кавказским.

Бурушаски, возможно, наиболее изолированный из всех языков региона. Все же, и у него имеются общие черты с северо-кавказскими, енисейскими и тибетскими.

Юкагирский иногда включают в уральские языки, иногда в палеосибирские.

Наиболее сложная ситуация с нивхским языком. В разных работах он сближается с уральскими, алтайскими, кавказскими, чукотско-камчатскими языками, а также с юкагирским, китайским и языками индейцев Северной Америки.

К этой же группе отнесем и случаи, когда некоторые языки традиционно относятся к той или иной языковой семье, но они там занимают явно периферийное положение и их принадлежность к этой семье оспаривается.

Наиболее хорошо известный пример – корейский и японский языки, включаемые в алтайскую семью. Они имеют существенную специфику, и если принадлежность к этой семье корейского языка является достаточно признанной, то генетические корни японского языка, имеющего общие черты не только с алтайскими, но и с австронезийскими языками, менее ясны.

Ительменский язык традиционно относят к чукотско-камчатской семье. Однако ряд исследователей (А.П. Володин, Д. Уорт) считают его генетически изолированным.

Чувашский язык, хотя и относят к тюркским, но, возможно, он представляет собой «отсутствующее звено», соединяющее тюркские и монгольские языки.

Г.Г. Сильницким путем статистического анализа на сравнительно небольшом количестве признаков получены следующие корреляции, которые могут быть подтверждены или опровергнуты при учете большего числа признаков.

1. Литовский язык (входит в балто-славянскую группу) ближе к германским, чем к славянским.
2. Болгарский язык находится на периферии славянских, сближаясь с индийскими.
3. Иранские языки занимают центральное положение среди индоевропейских, романские – периферийное.
4. Афразийские языки (хамито-семитские) объединяются с кавказскими.
5. Эскимосский, нивхский и тюркские – с уральскими.
6. Алтайские (кроме тюркских) с австронезийскими.

1.2. Корреляции между различными классификациями

Наиболее последовательной и впечатляющей попыткой установления корреляций между различными классификациями языков является работа Дж. Николс (Nichols 1992). Мы приведем здесь некоторые результаты Дж. Николс, релевантные данной работе, т. е. относящиеся к языкам региона Европы, Северной и Центральной Азии (сокращенно, LENCA) и некоторым сопредельным, включенным в базу данных «Языки мира». В рамках генеалогического подхода в работе Дж. Николс рассмотрены следующие языки и семьи этой области (табл. 1.1.).

Т а б л и ц а 1.1

Языки и семьи, представленные в исследовании Дж. Николс

Семья	Ветвь	Язык
Афразийская	Чадские	Хауса
	Кушитские	Оромо
	Омотские	Дизи
	Семитские	Амхарский
	Семитские	Аккадский
Индоевропейская	Армянские	Армянский
	Балто-славянские	Русский
	Германские	Английский
	Романские	Французский
	Анатолийские	Хетский
	Индоиранские	Бенгальский
Дравидийская	Дравидийские	Кота
Картвельская	Картвельские	Грузинский
Нахско-дагестанская	Нахские	Чеченский
Западнокавказская	Западнокавказские	Абхазский
Уральская	Угорские	Венгерский
	Финские-пермские	Коми-зырянский
	Самодийские	Юрок
Урало-юкагирская? Изолят?	Урало-юкагирские? Изолят?	Юкагирский
Хуррито-урартская	Хуррито-урартские	Хурритский

Семья	Ветвь	Язык
Монгольская	Монгольские	Монгольский
Тунгусская	Тунгусские	Нанайский
Тюркская	Тюркские	Тувинский
Енисейская	Енисейские	Кетский
Чукотско-Камчатская	Чукотско-Камчатские	Чукотский
Сино-тибетская	Тибето-бирманские	Гурунг
Австронезийская	Чамские	Ачехский
Австроазиатская	Австроазиатские	Тэмиар
Изоляты		Эламский
		Шумерский
		Баскский
		Айну
		Нивхский
		Корейский
		Японский
		Бурушаски
		Нахали

В рамках ареального подхода Дж. Николс предложила интересную концепцию, введя два типа зон: протяженные (spread) и замкнутые (residual) (Nichols 1992, с.16-21). К первому типу отнесены Европа и Древний Ближний Восток (Аккадский, Эламский, Хеттский, Хурритский, Шумерский). Ко второму типу отнесены Кавказ (Абхазский, Армянский, Чеченский, Грузинский) и Север азиатского побережья (Айну, Чукотский, Нивхский, Японский, Корейский, Нанайский, Юкагирский). Языки последней области в какой-то мере соответствуют палеоазиатским, но в оригинале используется термин North Asia Coast.

Кроме того, она выделила на основе географических и культурных критериев следующие большие области: Африка (к югу от Сахары и Сахель), Древний Ближний Восток (Месопотамия и Анатолия), Северная Евразия (Европа, Кавказ, северная часть Азии), а также макрозоны: Новый Свет, Старый Свет, Океания.

Дж. Николс рассматривает 4 основных типологических (структурных) свойства языков: порядок слов, строй языка, место маркировки синтаксических отношений и морфологическая сложность. Порядок слов понимается традиционным образом, выделяются три случая: глагол в начале предложения (V...), в середине (...V...), в конце (...V). Строй языка может быть аккузативным, эргативным, активным, нейтральным, трипартильным, иерархическим. В нижеследующих таблицах А будет обозначать языки аккузативного строя совместно с нейтральными, Е – эргативные языки, S – активные совместно с иерархическими. Трипартильные языки встречаются слишком редко.

Место маркировки синтаксического отношения может быть либо на главном члене отношения (Н), либо на зависимом (D), либо на обоих (2/S или просто 2). Морфологическая сложность определяется в какой-то мере условным образом в форме целого числа от 0 до 15, после чего все языки делятся на три класса: с низкой (L), средней (M) и высокой (H) морфологической сложностью. Для удобства отсылок и сопоставлений сохранены обозначения книги (Nichols 1992).

На основе данных 174 языков (из них 26 относятся к интересующей нас области LENCА) Дж. Николс выявляет закономерности сочетания различных структурных

свойств между собой и встречаемости их в выделенных зонах. Нижерасположенные таблицы являются сокращением таблиц, приведенных в (Nichols 1992) на указанных страницах. Названия семей, областей и зон переведены дословно, исправлены бесспорные арифметические ошибки. При подсчете статистики по области LENCA просуммированы данные строк «Ближний Восток», «Европа и Кавказ» и «Северная Азия» из соответствующих таблиц. В таблицах 1.2-1.5 число в клетке означает число языков с данным свойством.

Т а б л и ц а 1.2

Строй языка (стр. 187)

Область	A	E	S
Африка	19	0	0
Древний Ближний Восток	2	2	1
Европа и Кавказ	6	3	1
Северная Азия	9	1	1
Южная и Юго-восточная Азия	6	3	1
Всего в мире	112 (67%)	32 (19%)	24 (14%)
Всего в Старом Свете	42 (76%)	9 (16%)	4 (7%)
Всего в LENCA	17 (65%)	6 (23%)	3 (12%)
Зона	A	E	S
Кавказ	2	2	1
Север азиатского побережья	6	1	0
Всего в мире в residual зонах	32 (78%)	5 (12%)	4 (10%)
Древний Ближний Восток	2	2	1
Европа	5	1	0
Всего в мире в spread зонах	18 (46%)	13 (33%)	(21%)

Большинство языков мира являются аккумулятивными. Этот тип широко представлен на всех континентах. Языки активного строя распространены по всему Новому Свету, но редко встречаются за его пределами. Эргативные языки распределены по нескольким небольшим кластерам, один из которых – Кавказ и Древний Ближний Восток – находится на территории LENCA. Распределение по типам в области LENCA близко к распределению во всем мире в целом.

Порядок слов (стр. 94)

Область	V...	...V...	...V	Сплит	Свободный
Африка	1	7	9	2	0
Древний Ближний Восток	0	0	5	0	0
Европа и Кавказ	1	4	3	2	0
Северная Азия	0	0	9	2	0
Южная и Юго- восточная Азия	0	4	5	0	1
Всего в мире	21 (14%)	29 (20%)	78 (53%)	11 (8%)	7 (5%)
Всего в Старом Свете	2 (4%)	15 (27%)	31 (56%)	6 (11%)	1 (2%)
Всего в LENCA	1 (4%)	4 (15%)	17 (65%)	4 (15%)	0 (0%)
Зона	V...	...V...	...V	Сплит	Свободный
Кавказ	0	1	3	1	0
Север азиатского побережья	0	0	6	1	0
Всего в мире в residual зонах	1	9	20	4	4
Европа	1	3	1	1	0
Древний Ближний Восток	0	0	5	0	0
Всего в мире в spread зонах	9	6	16	3	0

Наиболее частотным в мире является порядок слов с глаголом в конце предложения. Распределение по типам в области LENCA близко к распределению в Старом Свете.

Место маркировки (стр. 186)

Область	Место маркировки		
	D	2/S	H
Африка	12	4	3
Древний Ближний Восток	2	3	0
Европа и Кавказ	6	3	1
Северная Азия	6	3	2
Южная и Юго-восточная Азия	8	1	1
Всего в мире	59 (34%)	52 (30%)	61 (35%)
Всего в Старом Свете	34 (62%)	14 (25%)	7 (13%)
Всего в LENCA	14 (54%)	9 (35%)	3 (12%)

Зона	D	2/S	H
Кавказ	3	1	1
Север азиатского побережья	4	1	2
Всего в мире в residual зонах	18	15	10
Древний Ближний Восток	2	3	0
Европа	3	3	0
Всего в мире в spread зонах	14	13	16

Хотя в мире все типы встречаются практически одинаково часто, существует четкая дифференциация распределения по континентам. В Старом Свете (и в меньшей степени в LENCA-области) маркировка на зависимом явно преобладает. Маркировка на голове синтаксического отношения преобладает в Новом Свете, двойная маркировка – в Австралии.

Т а б л и ц а 1.5

Морфологическая сложность (стр. 89)

Область	Сложность			Среднее значение
	L (0-5)	M (6-9)	H (11-15)	
Африка	6	8	5	7,9
Древний Ближний Восток	0	1	4	12,2
Европа и Кавказ	1	3	6	10,0
Северная Азия	2	5	4	9,2
Южная и Юго-восточная Азия	3	5	2	7,2
Всего в мире	34	96	42	8,0
Всего в Старом Свете	12	22	21	8,8
Всего в LENCA	3	9	14	-
Зона	L (0-5)	M (6-9)	H (11-15)	Среднее значение
Кавказ	0	2	2	9,8
Север азиатского побережья	2	4	1	7,9
Всего в мире в residual зонах	3	25	14	9,2
Древний Ближний Восток	0	1	4	12,2
Европа	1	1	4	10,2
Всего в мире в spread зонах	8	22	13	8,3

В Новом Свете средняя морфологическая сложность языков меньше, чем в других областях. Новый Свет также представляет собой в этом аспекте более однородную область с меньшим разбросом значений сложности. В Старом Свете разброс значений больше. В целом в мире преобладают языки средней сложности, в Старом Свете – средней и высокой практически поровну, в LENCА преобладают языки высокой сложности.

Т а б л и ц а 1.6

Свойства большинства языков в группе (стр. 171)

Группа	Число языков	Порядок слов		Строй		Место маркировки		Сложность	
		тип	%	тип	%	тип	%	тип	%
Афро-азиатская	4	...V	75	A	100	D	75	H	100
Индоевропейская	5	...V...	60	A	80	D	80	H	60
Урало-юкагирская	4	...V	75	A	100	-		-	
Австронезийская	6	V...	83	A	50	-		M	100
Древний Ближний Восток	5	...V	100	-		2	60	H	80
Кавказ	5	...V	80	-		D	60	H	60
Европа	5	...V...	60	A	80	2	60	H	60
Внутренняя Сибирь	6	V...	100	A	83	-		-	
Север Азиатского побережья	5	V...	100	A	80	D	60	-	

Примечание. ‘-’ означает отсутствие преобладающего типа.

Эта таблица представляет вышерассмотренные данные в несколько ином ракурсе, а также содержит данные по некоторым семьям.

Суммируя по определенным формулам вышеприведенные и другие, не упомянутые здесь данные, Дж. Николс выявляет, какие области наиболее близки друг другу по всему спектру свойств. Старый Свет оказался обособленной зоной, хотя Южная и Юго-восточная Азия имеют некоторую схожесть с Океанией, а в опущенной здесь строчке Австралия упомянута как типологически близкая Европа. Регион LENCА явно выделяется из Старого Света, хотя опять-таки Южная и Юго-восточная Азия имеют определенное сходство с Европой и Северной Азией. Внутри региона LENCА наибольшей близостью характеризуются Северная Азия с Кавказом и Ближним Востоком. Европа, являясь периферией, имеет сходство с ними обоими. В принципе, эти результаты естественны и ожидаемы, т.к. практически совпадают с географической близостью областей.

Дж. Николс связывает это с путями распространения человечества по Земному шару. В частности, некоторые черты сходства языков Австралии и Европы она объясняет тем, что выходцами из Старого Света первой была заселена Австралия. Учитывая, что, согласно (Nichols 1992), заселение Австралии началось 50 тыс. лет назад, сохранившиеся общими черты должны быть предельно стабильны.

Интегральная корреляция между областями (стр. 223)

Область	Наиболее близкая область
Африка	Южная и Юго-восточная Азия
Кавказ и Ближний Восток	Северная Азия
Европа	Кавказ и Ближний Восток, Северная Азия
Северная Азия	Кавказ и Ближний Восток
Южная и Юго-восточная Азия	Африка, Европа, Северная Азия, Океания

Между собой основные структурные свойства связаны следующим образом. Маркировка на голове коррелирует с низкой сложностью, на зависимом – с высокой. Эргативный строй коррелирует с маркировкой на зависимом, активный и иерархический – на голове. Начальная позиция глагола или свободный порядок благоприятствуют маркировке на зависимом. Эргативный строй ассоциируется с высокой сложностью (следствие из первых двух).

Для подтверждения статистической достоверности выводов применялся χ^2 -критерий. Таким образом, Дж. Николс описала глобальные закономерности распространения языковых свойств, связав их с путями миграции человечества. Следует отметить, что было изучено распределение лишь очень ограниченного набора свойств, причем среди них нет ни одного фонетического.

1.3. Математические методы

Уже в описанной в предыдущем разделе работе Дж. Николс для выявления закономерностей и подтверждения их валидности использовались методы математической статистики, хотя и в простейшей форме. Следующим важным шагом в развитии методологии применения статистических методов в типологии является работа Г.Г. Сильницкого (Сильницкий 2004). Им широко применяется целый спектр математических методов: корреляционный, факторный, кластерный и дискриминантный анализ. Расчеты велись с помощью пакета Statistika 4.3.

Для анализа им выбраны 78 языков, которые относятся к следующим генеалогическим типам:

Индоевропейские языки.

Славянские языки (8): русский, болгарский, сербохорватский, македонский, польский, чешский, словацкий, старославянский.

Германские языки (8): английский, немецкий, голландский, норвежский, шведский, древнеанглийский, древненемецкий, древнеисландский.

Итало-романские языки (7): французский, итальянский, испанский, португальский, румынский, старофранцузский, латынь.

Индоиранские языки (8): санскрит, ведийский, хинди, персидский, среднеперсидский, хорезмийский, белуджский, талышский.

Балтийские языки (1): литовский.

Другие (5): древнегреческий, новогреческий, армянский, хеттский, лидийский.

Неиндоевропейские языки.

Дравидийские языки (1): малаялам.

Хамито-семитские (9): арабский, иврит, эфиопский, сирийский, ассирийский, древнеегипетский, аккадский, сомали, хауса.

Уральские языки (2): финский, венгерский.

Алтайские языки (5): турецкий, узбекский, монгольский, японский, корейский.

Сино-тибетские, тайские языки (4): китайский, тибетский, бирманский, тайский.

Австроазиатские языки (2): вьетнамский, кхмерский.

Австронезийские языки (3): индонезийский, яванский, тагальский.

Палеоазиатские языки (2): эскимосский, нивхский.

Кавказские языки (3): кабардино-черкесский, грузинский, лезгинский.

Африканские языки (5): суахили, зулу, бабунго, нкоре-кига, мупун.

Австралийские языки (1): йидин.

Другие (4): баскский, лимбу, амеле, хиксариана.

Для классификации изучаемых языков Г.Г. Сильницкий рассмотрел 45 признаков. Признаки выбирались из тех соображений, чтобы они были представлены во всех или большинстве рассматриваемых языков и являлись общепринятыми, однозначно трактуемыми.

Грамматические.

Номинативные признаки: артикль, род существительного, число существительного, падеж существительного, морфологическая неизменяемость существительного, род прилагательного, число прилагательного, падеж прилагательного, морфологическая неизменяемость прилагательного, род местоимения, число местоимения, падеж местоимения, двойственное число, род глагола, число глагола (последние два отнесены к этой группе, т. к. они категориально совпадают с номинативными).

Глагольные признаки: лицо глагола, категория залога, категория наклонения, категория сослагательного наклонения, временные формы сослагательного наклонения, категория времени, категория будущего времени, будущее-в-прошедшем, категория перфекта, категория длительных времен, категория вида, инфинитив, причастие, деепричастие, герундий/супин, морфологическая неизменяемость глагола.

Фонетические.

Вокальные признаки: количество гласных, процент гласных в фонетической системе, наличие долгих гласных, наличие дифтонгов, количество носовых гласных, наличие тона, фиксированное ударение.

Консонантные признаки: количество согласных, количество сонорных неназализованных согласных, количество носовых согласных, количество аффрикат, наличие удвоения согласных, наличие палатализации согласных, количество полугласных.

В (Сильницкий 2004) приведено большое количество результатов (в форме таблиц) по классификации языков и признаков с помощью различных математических методов. Воспроизведем здесь некоторые итоговые классификации.

Полученная классификация индоевропейских языков довольно хорошо коррелирует с общепринятой генетической. В основном группа К1 состоит из славянских и иранских языков, в К2 – из германских, в К3 – из романских, в К4 – из индийских.

Итоговая классификация индоевропейских языков (стр. 134)

К1	К2	К3	К4
Русский	Английский	Французский	Болгарский
Польский	Немецкий	Испанский	Древнеанглийский
Чешский	Голландский	Итальянский	Санскрит
Словацкий	Норвежский	Португальский	Ведийский
Сербохорватский	Шведский	Румынский	Хинди
Македонский	Древне немецкий	Белуджский	
Старославянский	Древнеисландский	Хеттский	
Новогреческий	Старофранцузский		
Армянский	Латынь		
Персидский	Древнегреческий		
Среднеперсидский	Литовский		
Талышский	Лидийский		
Хорезмийский			

Численные данные показывают, что из индоевропейских языков наиболее противопоставлены друг другу славянские и германские языки, особенно по фонетическим признакам. Вообще Г.Г. Сильницкий считает, что фонетические свойства являются более фундаментальными, менее подверженными социальным и иным внешним влияниям.

Итоговая классификация неиндоевропейских языков (стр. 162)

К1	К2	К3	К4
Арабский	Венгерский	Китайский	Суахили
Иврит	Финский	Бирманский	Бабунго
Эфиопский	Турецкий	Тибетский	Мупун
Аккадский	Японский	Индонезийский	Амеле
Сирийский	Эскимосский	Яванский	Нкоре-кига
Древнеегипетский	Нивхский	Тагальский	Лимбу
Ассирийский	Малаялам	Тайский	Зулу
Черкасский	Сомали	Вьетнамский	Йидин
Лезгинский		Кхмерский	
Грузинский		Монгольский	
Баскский		Корейский	
Узбекский		Хауса	
		Хиксариана	

Эта классификация также в значительной степени коррелирует с ареально-генетической. Класс К1 состоит в основном из семитских и кавказских языков, К2 – из уральских и палеоазиатских, К3 – из языков Южной и Юго-восточной Азии, К4 – африканские. Интересно, что алтайские языки оказались распределены по трем группам.

Сравнивая классификации, полученные с помощью различных математических методов и на разных наборах признаков (фонетических, грамматических и полном), Г.Г. Сильницкий выделил набор из 11 «блоков» – наиболее стабильных групп языков.

Б1: русский, польский, чешский, словацкий, сербохорватский, старославянский, армянский.

Б2: английский, немецкий, норвежский, шведский, древнегреческий.

Б3: французский, португальский, испанский, румынский.

Б4: санскрит, ведийский.

Б5: иврит, эфиопский, аккадский, сирийский.

Б6: китайский, бирманский.

Б7: индонезийский, яванский.

Б8: тайский, кхмерский.

Б9: нивхский, эскимосский.

Б10: суахили, амеле, йидин.

Б11: бабунго, мупун.

Вообще же корреляционный анализ показал наличие сильных положительных корреляций между всеми языками, что, по Г.Г. Сильницкому, является объективным подтверждением фундаментального глоттогонического единства языка.

1.4. Базы данных

С развитием вычислительной техники появились новые возможности для представления результатов исследований научной общественности – базы данных. Наиболее известной базой данных в компаративистике является STARLING С.А. Старостина (Бурлак, Старостин 2001), в типологии – WAL (Haspelmath 2005).

STARLING – этимологическая база данных, содержащая информацию о лексическом составе семей на нескольких уровнях, связанных друг с другом ссылками. На уровне отдельных языков даны стандартные 100-словные списки ядра лексики. На уровне групп языков или ветвей это этимологические словари соответствующей группы. На уровне всей семьи представлены реконструированные корни протоязыка этой семьи. Связи позволяют проследить отражения любого корня в отдельных языках семьи.

Кроме данных, STARLING содержит и набор алгоритмов работы с данными. Эти алгоритмы позволяют решать определенные задачи компаративистики: моделирование процесса установления фонетических соответствий между языками, моделирование фонетических изменений, лингвостатистические подсчеты, автоматическое построение генеалогического дерева для семьи, поиск слов по приблизительному звучанию и значению.

STARLING является учебно-справочной базой данных. В настоящее время она содержит данные по большинству семей Евразийского континента. Система открыта для расширения.

WALS – The World Atlas of Language Structures – недавно заверченный крупный международный проект, выполненный под руководством М. Хаспельмата, Б. Комри и др. (Haspelmath 2005). WALS включает компьютерную базу данных и бумажное издание. Он оперирует со значительно большим числом языков и признаков, по сравне-

нию с работами Дж. Николс и Г.Г. Сильницкого: 2560 языков и 140 признаков. Каждый признак может принимать несколько значений, в среднем, примерно, 5.

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 (в том числе, в работе Дж. Николс), но впервые она была реализована столь масштабно. Во-вторых, база данных снабжена чрезвычайно удобным интерфейсом, позволяющим легко ориентироваться в огромном массиве информации – WALS включает десятки тысяч справочных статей по различным признакам и языкам. Признаки классифицируются по тематике и по алфавиту, языки – по семьям и регионам. Поисковые средства позволяют находить нужную информацию по комбинации признаков, генерировать новые карты с заданными свойствами.

1.5. Резюме по главе 1

1. Языки могут классифицироваться по различным основаниям: по общности происхождения (генетическая классификация), территориальной близости зон распространения (ареальная), сходству структурных свойств (типологическая).
2. Между этими классификациями существуют значимые корреляции, однако нет жестких зависимостей.
3. В генетической классификации языков достигнуты значительные успехи применением сравнительно-исторического метода, для основных языковых семей реконструированы протоязыки.
4. Вероятно, сравнительно-исторический метод в классической форме практически исчерпал свои возможности: реконструкция на глубину, большую 10 тыс. лет, считается невозможной, многие тонкие вопросы остаются без ответа.
5. С введением в научный оборот все новых лингвистических данных и созданием обширных баз данных появились принципиально новые возможности исследований с применением математических и компьютерных методов.

Глава 2. БАЗА ДАННЫХ «ЯЗЫКИ МИРА»

2.1. История создания БД

В 80-е годы в Институте языкознания РАН были начаты работы по созданию базы данных (БД) «Языки мира». В качестве источника информации используется одноименное энциклопедическое издание (Языки мира 1993...2006). Работы были инициированы членом-корреспондентом АН В.Н. Ярцевой и велись в отделе прикладного языкознания под руководством А.И. Новикова. В разработке концепции и структуры БД принимали участие: А.К. Зотова, Н.К. Рябцева, Н. Рогова, О.И. Романова – анализ рефератов, В.А. Виноградов, М.А. Журина, Я.И. Тестелец, Е.И. Ярославцева – авторы модели, Ю.П. Скокан, А.И. Новиков, Н.Н. Нестерова – компьютерная формализация модели.

Первая версия БД реализована программистом Ю.П. Скоканом в СУБД Clipper (MS DOS). На БД получено регистрационное свидетельство ФГУП НТИЦ «Информрегистр» № 7706 от 26 ноября 2001 г. Был сделан ряд публикаций¹. В 2005 году защищена докторская диссертация Е.И. Ярославцевой на тему «Компьютерная база данных “Языки Мира” и ее возможные применения».

В 2002 году создана Windows-версия БД «Языки мира» (руководитель проекта – В.Н. Поляков, программист – В. Логунов). В 2005 году представлен первый вариант Web-версии, и в марте 2006 г. база была опубликована в сети Интернет по адресу www.dblang.ru (руководитель проекта – В.Н. Поляков, программисты – Е. Гончаров, Т. Щербинин). Разработана учебная программа спецкурса «БД “Языки мира” и новые возможности типологических и компаративных исследований» (авторы: В.Н. Поляков, В.Д. Соловьев), который был прочитан на ОТиПЛе (филфак МГУ) и на факультете «Лингвистика» ЮУрГУ в 2006 году.

Работа была неоднократно поддержана грантами РГНФ, частично финансировалась из средств МГЛУ, в 2006 году вошла в программу финансирования Российского научно-образовательного центра по лингвистике имени И.А. Бодуэна де Куртенэ при КГУ (руководитель НОЦа – В.Д. Соловьев, госконтракт № 02.438.11.7015).

2.2. Характеристики контента Базы данных «Языки мира»

По состоянию на 01.12.06 БД «Языки Мира» характеризуется следующими количественными параметрами.

- Количество языков: 315 языков Евразии.
- Количество признаков: 3821.
- Содержит описание следующих сфер языка: фонетика, морфология, синтаксис.
- Формат представления данных: бинарный.

В БД «Языки мира» ИЯ РАН представлены следующие языковые семьи и языковые сообщества: австроазиатская, австронезийская, алтайская, афразийская, индоевропейская, кавказская², палеоазиатская³, синотибетская, уральская, хуррито-

¹ (Виноградов, Новиков, Ярославцева 2003), (Виноградов, Новиков, Ярославцева, Поляков, Логунов 2003), (Жури-ская, Новиков, Ярославцева 1986), (Новиков, Ярославцева 1985), (Поляков, Савельев, Соловьев 2006-1...2), (Поляков, Соловьев 2006-1...7), (Поляков, Соловьев, Ахтямов 2006), (Соловьев, Поляков 2006), (Ярославцева 1999), (Ярославцева 2001), (Ярцева 1977), (Novikov, Yaroslavtseva 1986).

² Не является семьей, а представляет собой ареальное сообщество языков.

³ То же самое.

урартская. По вполне понятным причинам, БД «Языки мира» содержит описание языков, распространенных в основном на территории Евразии. Однако, широкий охват представленных в БД языковых семей вполне оправдывает название «Языки мира». Исторически так сложилось, что общепринятая классификация языков одновременно включает две системы оснований: генетическую и ареальную. Самые крупные сообщества языков, объединенных по генетической близости, принято называть макросемьями. Далее по мере дробления выделяют семьи, ветви, группы и подгруппы языков. Кратко охарактеризуем представленные в БД семьи.

Австроазиатская⁴. Семья языков, распространенных в Юго-Восточной и Южной Азии. Включает ветви языков: семанг-сакай (аслианскую), вьетмыонгскую, монкхмер, палаунг-ва, никобарского языка, языка кхаси, мунда, нагали языков. В БД представлена двумя языками: вьетнамским и кхмерским. Оба языка относятся к монкхмерской ветви. Отличительной особенностью австроазиатских языков принято считать их слабую морфологизированность (Сильницкий 2004, с. 147)⁵. Фонетические признаки (Сильницкий 2004, с. 154): тон, носовые согласные (не менее трех), отсутствие палатизации, аффрикатов, носовых гласных, ограничения на неносовые сонорные (не более трех) и полугласные (не более одного).

Австронезийская. Семья языков, традиционно подразделявшаяся на 4 группы: индонезийские, полинезийские, меланезийские и микронезийские языки. Ядро семьи австронезийских языков называется малайско-полинезийскими языками. В современных классификациях австронезийские языки Тайваня противопоставляются остальным австронезийским языкам, среди которых выделяется океанийская (восточноавстронезийская) ветвь. Распространены на территории Индонезии, Папуа-Новой Гвинеи, островах Тихого океана. В БД имеется всего один представитель этой семьи: язык ротума, который относится к восточно-океанийской группе восточно-малайско-полинезийской ветви. Для языков австронезийской семьи характерны следующие грамматические категории (Сильницкий 2004, с. 147): вид, нет четкого выделения прилагательного, слабая морфологизированность. В области фонетики выявлены характерные черты (Сильницкий 2004, с. 154): консонантность, носовые согласные (не менее трех), отсутствие дифтонгов, носовых гласных, палатизации, ограничения в аффрикатах (не более трех), неносовых сонорных (не более трех).

Алтайская. Макросемья языков, объединяющая 3 языковые семьи: тюркскую, монгольскую, тунгусо-маньчжурскую⁶. Часто к алтайским языкам относят также японский и корейский (Бурлак, Старостин 2001). О происхождении и родстве алтайских языков существуют разные гипотезы. Ареал существования алтайских языков очень широк: от территории современной Турции на западе до территории современной Монголии на востоке. Алтайская макросемья является второй после индоевропейской по числу языков, представленных в Базе Данных. Включает описания 76 языков. Для языков алтайской макросемьи характерны (Сильницкий 2004, с. 147): неизменяемость прилагательного, дееприча-

⁴ Здесь и далее характеристики языковых семей даются по (БЭС 1997).

⁵ Необходимо отметить, что релевантные типологические признаки по методике Г.Г. Сильницкого выявляются статистическими методами и характерны для всей совокупности рассматриваемых языков. При этом в каждом конкретном языке они могут отсутствовать.

⁶ Иногда к алтайской макросемье относят и уральские языки.

стие, послелогои, временные формы причастия, суффиксация, каузативные аффиксы, отсутствие категории рода, слабая морфологизированность языка.

Тюркская. Семья близкородственных языков. Входит в алтайскую макросемью языков. В свою очередь делится на восемь групп языков (булгарские, древнетюркские, карлукские, кыпчакские, огузские, тоба, хакасско-алтайские, якутские). В БД представлено описание 54 тюркских языков. В (Сильницкий 2004, с. 154) выделяется набор релевантных фонетических признаков для тюркских языков: консонантность, фиксированное ударение, отсутствие долгих гласных, дифтонгов, носовых гласных, ограничение носовых сонорных (не более трех) и полугласных (не более одного).

Монгольская. Семья языков монгольских народов, образовавшаяся в XIV-XVI вв. из диалектов раннего монгольского языка, единого для всех племен. Включает монгольский, бурятский, калмыцкий языки; бесписьменные — дунсянский, монгорский, баоаньский языки и др. В БД представлена 11 языками.

Тунгусо-маньчжурская. Семья родственных языков, включаемая в макросемью алтайских языков. Представлена 10 языками, которые в свою очередь разделяют на три группы (амурские, сибирские и южные).

Афразийская (семито-хамитские, афро-азиатские). Макросемья языков, распространенных в Северной Африке, Западной Азии и на о. Мальта. Делятся на 5 (или 6) основных ветвей: семитскую, египетскую, берберо-ливийскую (берберо-канарскую), чадскую, кушитскую и омотскую (иногда рассматриваемую как ветвь кушитской). В БД описаны 9 языков из этой семьи: 7 семитских, 1 берберо-канарский и 1 кушитский. Согласно (Сильницкий 2004, с. 147), семитские языки имеют следующие характерные грамматические признаки: двойственное число, два времени (прошедшее – настоящее/будущее), род, масдар, отсутствие временных форм причастия. Выделены также следующие релевантные фонетические признаки (Сильницкий 2004, с. 154): консонантность (исключения: сирийский, ассирийский), долгие гласные, отсутствие носовых гласных, палатализации, ограничение в носовых согласных (не более двух), ограничение в аффрикатах (не более двух).

Индоевропейская (ИЕ). Одна из наиболее крупных языковых семей. В БД «Языки мира» представлено 143 языка ИЕ-семьи⁷, которые объединены в ветви: анатолийские, армянские, иранские, дардские, индоарийские, иранские, италийские, кельтские, нуристанские, романские, славянские. Ареал распространения ИЕ-языков охватывает значительную часть Евразии. Согласно (Сильницкий 2004, с. 106), общими релевантными грамматическими признаками⁸ ИЕ-языков являются: артикль, перфект, будущее в прошедшем, высокая степень морфологизированности языка. При этом для **славянских** языков характерны: род (прилагательного, существительного, глагола), вид, деепричастие, сильная морфологизированность языка. Для **романских**: артикль, перфект, продолженные времена, сослагательное наклонение, будущее-в-

⁷ Часть языков, например, греческая ветвь, пока не охвачены энциклопедией, поэтому не представлены в БД. Другие, такие как балтийские, описаны, но пока еще не введены в БД.

⁸ Необходимо отметить, что релевантные типологические признаки по методике Г.Г. Сильницкого выявляются статистическими методами и характерны для всей совокупности рассматриваемых языков. При этом в каждом конкретном языке они могут отсутствовать.

прошедшем, герундий, сильная морфологизированность языка. **Германские** языки характеризуют: артикль, перфект, будущее-в-прошедшем, сильная морфологизированность языка. **Индоиранские** языки: двойственное число, средний залог, отсутствие вида и будущего времени.

Релевантные фонетические признаки ИЕ-языков (Сильницкий 2004, с. 121) для **славянских** языков: консонантность, палатализация, отсутствие дифтонгов, долгих гласных. **Германские** языки характеризуют: вокальность, дифтонги, долгие гласные, отсутствие палатализации, ограничения аффрикатов (не более двух), отсутствие носовых гласных. Для **романских** языков: отсутствие палатализации, ограничения на аффрикаты (не более двух) и сонорные носовые (не более трех). Для **индийских** языков характерны: консонантность, долгие гласные, носовые согласные (не менее трех), полугласные (не менее трех), сонорные носовые (более трех), тон, придыхание, отсутствие носовых гласных, палатализации, аффрикатов. **Иранские** языки характеризуют: консонантность, долгие гласные, ограничения сонорных носовых (не более трех), носовых согласных (не более двух), отсутствие дифтонгов, палатализации, носовых гласных.

Интересна позиция Н.С.Трубецкого по родству индоевропейских языков. Он считает, «что индоевропейскими стали языки, которые приобрели следующие шесть признаков: 1. консонантизм начала слова не беднее консонантизма середины и конца слова; 2. приставки; 3. аблаут; 4. чередование согласных в грамматических формах; 5. аккузативность (не-эргативность); 6. отсутствие сингармонизма. Приводится по (Бурлак, Старостин 2001).

Кавказская (иберийско-кавказские языки). Включает несколько групп: картвельские, абхазско-адыгские и нахско-дагестанские языки. Нахские и дагестанские языки иногда рассматривают как две различные группы. Древнеписьменную традицию, начиная с V в., имеет грузинский язык; возможно, в V-VIII вв. имел письменность удинский язык. Остальные языки — младописьменные или бесписьменные. Это сообщество языков не представляет собой генетически однородную семью. Генетически четко различаются южно-кавказские (картвельские) и северо-кавказские (абхазо-адыгские и нахско-дагестанские) языки. Последние в (Бурлак, Старостин 2001) относятся к сино-кавказской макро-семье. В БД представлено 38 кавказских языков, в том числе: 4 картвельских, 5 абхазо-адыгских и 29 нахско-дагестанских. Г.Г. Сильницкий выделил следующие релевантные признаки в группе кавказских языков⁹ в фонетике: консонантность, аффрикаты, отсутствие дифтонгов, палатализации, ограничение на носовые согласные (не более двух), в грамматике: сильная морфологизированность языка, масдар, каузативные морфемы, эргативность, версия, отсутствие рода.

Палеоазиатские языки (палеосибирские). Условное название генетически различных языков малочисленных народностей Северо-Восточной Сибири. Традиционно выделяются 4 генетические общности палеоазиатских языков: чукотско-камчатская; эскимосско-алеутская; юкагиро-чуванская; енисейская и изолированные айнский и нивхский языки. Внешние генетические связи палеоазиатских языков остаются во многом спорными. В БД представлено 19 палеоазиатских языков, среди которых 2 изолята (айнский, нивхский). Г.Г.Сильницкий выделил релевантные призна-

⁹ В (Сильницкий 2004) в группу кавказских языков были включены: кабардино-черкесский, лезгинский и грузинский, т.е. языки из различных генетических сообществ (картвельская и нахско-дагестанская группы).

ки для группы палеоазиатских языков¹⁰. В фонетике это: консонантность, носовые согласные (не менее трех), отсутствие дифтонгов, носовых гласных, аффрикатов, ограничение носовых сонорных (не более трех), полугласных (не более одного). В грамматике: деепричастие, нет четкого выделения прилагательных, отсутствие рода, каузативные морфемы, слабая морфологизированность языка.

Синотибетская. Семья языков, распространенных в Китае, Мьянме, Непале, Бутане и на северо-востоке Индии. Общепринятая генетическая классификация отсутствует. В (Бурлак, Старостин 2001) отнесена к синокавказской макросемье. Выделяют 2 ветви: китайский и тибето-бирманский. В БД включено два языка этой семьи: бирманский и древнекитайский. В грамматике данной семьи языков выделены релевантные признаки: неизменяемость существительного и глагола, вид, слабая морфологизированность языка. В фонетике синотибетские языки характеризуют (Сильницкий 2004, с. 154): вокальность (искл. бирманский), дифтонги, носовые согласные (не менее трех), отсутствие носовых гласных, палатализации, ограничение носовых сонорных (не более трех), аффрикативность, тон.

Уральская¹¹. Крупная семья языков, включающая финно-угорскую (угро-финскую) и самодийскую ветви. В БД содержится 22 уральских языка (16 финно-угорских и 6 самодийских). Основным ареалом распространения уральских языков является север Европы и Урал. Уральские языки характеризуются отсутствием вида (Сильницкий 2004, с. 147). В фонетике уральскую семью характеризуют (Сильницкий 2004, с. 154): вокальность, долгие гласные, носовые согласные (не менее трех), отсутствие носовых гласных, ограничение носовых сонорных (не более трех).

Хуррито-урартская. Включает два мертвых языка: хурритский и урартский. Эти мертвые языки были распространены в VI-IX вв. до нашей эры на территории Армении.

Языки-изоляты. БД содержит описание 5 языков изолятов: айнский, нивхский, бурушаски, шумерский, эламский.

Мертвые языки. Уникальной особенностью БД «Языки мира» является крупный корпус описаний мертвых языков, который насчитывает 54 реферата. Аналогов в мире для такого подробного и системного описания мертвых языков просто не существует.

В таблице А.1 (Приложение А) приведены сравнительные данные по традиционной генетической принадлежности языков из БД «Языки мира»¹² и генетической классификации из книги (Бурлак, Старостин 2001).

¹⁰ Группа палеоазиатских языков в (Сильницкий 2004) включает два языка: эскимосский и нивхский.

¹¹ См. сноску 6.

¹² Составлено на основе материалов, предоставленных Сектором «Языки мира» ИЯ РАН.

2.3. Принципы, положенные в основу Базы данных «Языки мира» ИЯ РАН

2.3.1. Бинарный принцип

БД представляет собой таблицу, состоящую из строк и столбцов. В каждой строке такой таблицы записан один определенный языковой факт, а столбцам поставлены в соответствие названия языков, по отношению к которым эти факты являются релевантными. Наличие такой взаимной релевантности фиксируется в виде определенной пометки на пересечении соответствующих столбцов и строк. На языке программистов такой тип представления данных называется переменной булевого (или логического) типа, когда переменная, кодирующая признак, может принимать два значения: ИСТИНА или ЛОЖЬ. Значение ИСТИНА соответствует факту наличия данного признака в данном языке, значение ЛОЖЬ кодирует его отсутствие¹³. Множество языковых фактов, относящихся к некоторому конкретному языку, составляет формализованный **реферат** описания данного языка.

Содержимое всех строк, взятых без пометок, отдельно, как перечень языковых явлений и категорий, составляет так называемую **модель реферата (МР)**. В отличие от рефератов, которые создаются в процессе функционирования БД, модель реферата, вернее, ее базовый компонент создается предварительно на одном из первых этапов ее построения. Также модель постоянно пополняется по мере ввода в БД новых языков и накопления новых фактов. В (Ярославцева 2005) для наименования всей совокупности категорий модели реферата вводится термин *грамматикон*.

«Грамматикон – универсальный, конкретно-языковой и частные – наборы грамматических категорий и явлений в модели языка, в рефератах конкретных языков и в отдельных классах модели».

Также вводятся термины для частей реферата.

«Фонематикон, ономастикон, вербатикон, просодикон, нумерикон, птофикон, дейктикон, фонотактикон, фонотипикон, морфотипикон, партикон, парадигматикон, слово-форматикон, дериватикон, сентенсикон, комплексикон, графикон и т.п. – названия частных таксонов грамматикона».

2.3.2. Иерархический принцип

В структурном отношении МР представляет собой классификационную схему в виде иерархического дерева. Первому уровню такой иерархии соответствуют наименования классов языковых явлений. Они практически идентичны тем классам, которые содержатся в схеме статьи, используемой в энциклопедии «Языки мира».

Внутри классов выделяются возможные аспекты рассмотрения данного класса языковых явлений. Совокупность аспектов – это дальнейшая градация содержания, осуществляемая на уровне каждого раздела.

Классы и аспекты – это универсалии, априорно задаваемые в модели как наименования явлений, общих либо для большинства описываемых языков, либо для некоторой группы языков. Они соответствуют подтемам и субподтемам описания некоторого языка. Элементы, находящиеся на более низких уровнях иерархии, подаспекты и характеристики соответствуют более конкретным языковым явлениям, специфичным для одного или нескольких описываемых языков. Классы, аспекты и подаспекты, т.е.

¹³ Строго говоря, двухзначной логики для описания признакового пространства не всегда достаточно. Есть языки, и их довольно много, которые изучены не так подробно, как хотелось бы, и в этом случае требуется ввести еще третье состояние – «НЕ ИЗВЕСТНО». Частично выходом из такой ситуации явилось введение специально признака «О», которым кодируется факт отсутствия достоверной информации по разделам реферата.

нетерминальные признаки, принято считать классифицирующими признаками, которые соответствуют межвидовым отличиям между языковыми сообществами. Напротив, характеристики (или терминальные признаки) являются фактографическими признаками и иллюстрируют внутривидовые отличия между родственными языками внутри языковых сообществ.

Для иллюстрации соотношения структурных элементов в модели приведем в качестве примера фрагмент класса 2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ (табл. 2.1).

Т а б л и ц а 2.1.

Фрагмент класса 2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ

Номер в БД	Наименование и уровень в иерархии признака¹⁴
3773	2.5.4.СЛОЖНОЕ ПРЕДЛОЖЕНИЕ
3774	.О
3775	..линейный порядок компонентов
3776	..главное предшествует придаточному
3777	..главное предшествует целевому
3778	..не фиксирован
3779	..придаточное предшествует главному
3780	..целевое предшествует главному
3781	..особенности подчиненного компонента
3782	..особое оформление именных групп
3783	..особое оформление подлежащего
3784	..особый порядок слов
3785	..оформление сказуемого
3786	...зависимые личные формы
3787	...квотатив
3788	...нефинитные формы
3789	...оборот с релятивной формой глагола
3790	...особые глагольные показатели
3791	...особый порядок слов
3792	...таксисные конструкции
3793	...финитные формы
3794	...абсолютные обороты
3795	..'тематическое' придаточное
3796	..особенности построения дискурса
3797	..обвиатив
3798	..переключение референции
3799	..структура относительного предложения
3800	..главное предшествует придаточному
3801	..относительно-местоименные аффиксы в глаголе
3802	..отсутствие союзного слова или его аналога
3803	..придаточное предшествует главному

¹⁴ Здесь и далее количество точек перед названием признака маркирует уровень в иерархии.

Номер в БД	Наименование и уровень в иерархии признака ¹⁴
3804	..союзное слово в главном
3805	..союзное слово в придаточном
3806	..сказуемое придаточного следует за союзом
3807	.тип построения
3808	..сериализация
3809	..сочинение/подчинение
3810	..только подчинение
3811	..только сочинение
3812	.тип связи
3813	..союзная/бессоюзная
3814	..преобладает бессоюзная
3815	..преобладает союзная
3816	..союзы
3817	...знаменательные слова
3818	...отсутствие союзов как грамматического разряда
3819	...присоединяемые служебные элементы
3820	...самостоятельные служебные элементы
3821	...союзные формы глагола

В таблице Б.1 (Приложение Б) приведен перечень классифицирующих признаков, представленных в БД «Языки мира».

Реферат имеет структуру, идентичную модели, поскольку является результатом включения в него из модели тех языковых фактов (строк), которые присущи данному конкретному языку.

2.3.3. Принцип парадигмы

Представление многих фактографических признаков в Базе Данных выполнено с использованием принципа парадигмы. Это означает, что в поле наименования признака записываются через косую черту (слэш) все его возможные значения. Таким образом характеристикой выступает не отдельное значение, а вся парадигма значений. В таблице 2.2 представлен пример описания признака «конкретная (или преобладающая) структура слога» из раздела 2.1.4.

Т а б л и ц а 2.2.

Пример признака «конкретная (или преобладающая) структура слога» из раздела «2.1.4. СЛОГ»

Номер в БД	Частота признака	Наименование и уровень в иерархии признака
802	247	.конкретная (или преобладающая) структура слога
803	1	..C/V/CVC/CVV
804	1	..CC/CCC/CCVCCCVCCC
805	3	..CV
806	1	..CVC

Номер в БД	Частота признака	Наименование и уровень в иерархии признака
807	1	..CCCVCC
808	4	..CV/VC
809	7	..CV/CVC
810	1	..CV/CCV
811	7	..CV/VC/CVC
812	2	..CV/VC/VCC
813	12	..CV/CVC/CVCC
814	3	..CV/CVC/CCV
815	2	..CV/CVC/CCVC
816	2	..CV/CVC/CCV/CCVC
817	1	..CV/VC/CVC/CVCC
818	1	..CV/VC/CCVC/CVCC
819	2	..CV/CCV/CVC/VC
820	2	..CV/CVC/CCVC/CVCC
821	1	..CV/CVC/CCV/CCCV
822	2	..CV/CVC/VC/VCC/CVCC
823	2	..CV/VC/CVC/CCV/CCVC
824	1	..CV/CCV/CCCV/CVC/CCVC/CVCCCC
825	1	..CV/CCV/CCCV/CVC/VCC/VC/CCVCC
826	1	..CV/VC/CVC/VCC/CCV/CVCC
827	1	..CV/VC/CVC/CCVC/CVCC
828	3	..CV/CVC/VC/VCC/CVCC/CCVC
829	1	..CV/CVC/CCV/CCVC/CCCV
830	1	..CV/VC/CVC/VCC/CCVC/CCVCC/CVCC
831	1	..CVC/CV/CVCC/CCV/CCVC/CCVCC
832	2	..CV/CVC/CVCC/CVCCC/CVCCCC
833	1	..CV/VC/CVC/CCV/CCVC/CCCV/CCCV
834	0	..CCV/VC/VCC
835	1	..CV/CVC/CVCC/CVCCC/CCV/CCVC/CCVCC/CCVCCC/ см.F4
836	1	..CV/CVC/CCV/CCVC/CCCV/CVCC/CCVCC/CCCV/CVCC
837	2	..CV/CVCC/CVC/CCVC/VC/V/VCC
838	1	..V/CV/VC
839	3	..V/CV/CVC
840	2	..V/CV/CVC/CVCC
841	38	..V/CV/VC/CVC
842	1	..VC/CVC/VCC/CVCC
843	2	..VC/CVC/CV/VCC/CVCC
844	15	..V/VC/CV/CVC/CVCC
845	1	..V/CV/CCV/CCCV/VC
846	16	..V/CV/VC/CVC/VCC/CVCC
847	1	..V/CV/VC/CVC/CCVC/CVCC
848	1	..V/CV/VC/CVC/CVCC/CCVC/CVCV
849	4	..V/VC/CV/CVC/(C)VC(C)
850	1	..V/CV/CCV/CCCV/CCCV/CCVC
851	1	..V/VC/CCV/CCVC/CCCV/CVC/CVCC
852	9	..V/CV/CCV/VC/CVC/CCVC
853	1	..V/VC/CV/CCV/CCVC/CCCV/CCVC

Номер в БД	Частота признака	Наименование и уровень в иерархии признака
854	5	..V/VC/CV/CVC/CVCC/CCVC/CCV
855	2	..V/CV/VC/CVC/VCC/CVCC/CCV/CCVC/CVCCC
856	5	..V/VC/CV/VCC/CCV/CVCC
857	20	..V/VC/VCC/CV/CVC/CVCC
858	2	..V/VC/CV/VCC/CVCC/CVC/CCV
859	2	..V/CV/CVC/CCV/CCVC/CCVCC/CVCC/VCC
860	2	..V/VC/VCC/CV/CVC/CVCC/CCVC/CCVCC
861	3	..V/VC/CV/CVC/VCC/CVCC/CCV/CCVC
862	2	..V/CV/CCV/CCCV/CVC/VCC/CCVC/CCCV
863	1	..V/CV/VC/VCC/CVC/CCVCCC/CCCVCC
864	2	..V/CV/CCV/CCCV/CCCV/CVC/CCVCC/CCCVCC
865	1	..V/CV/VC/CVC/VCC/CCV/CCVC/CVCC
866	2	..V/CV/CVC/VC/VCC/VCCC/CVCC/CVCCC
867	1	..V/VC/CV/CVC/VCC/CVCC/VCCC/CCV/CCVC/CVCC/ см.F4
868	2	..V/VC/CV/CVC/VCC/CVCC/CCV/CCVC/CCVCC
869	1	..V/CV/VC/CVC/CVCC/CVCCC/CCVC/CCCV/CCVCC/ см.F4
870	1	..V/VV/VC/VVC/VCC/CV/CVC/CVV/CVVC/CVCC/ см.F4
871	1	..V/CV/VC/CVC/CCV/VCC/CCVC/CCVCC/CVCC/CVCC см.F4
872	1	..V/CV/VC/CVC/VCC/CVCC/VCCC/CVCCC/CCV/CCVC см.F4
873	1	..V/VC/CV/CVC/CVCC/VCC/CCV/CVCCC/CCVC/CCVCC
874	1	..VC/CVC/VCC/CVCC/CC(C)VC/CC(C)VCC/V/CV/CC(C)V
875	4	..V/CV/CCV/CCCV/VCC/VCC/CVC/CCVC/CCVCC/CCCV
876	1	..VC/CV/CCV/VCC/VCC/CVC/CCVC/CVCC/CCVCC
877	1	..VC/CV/CVC/CCV
878	1	..V/CV/CVC/CCVC
879	1	..V/VC/VCC/VCCC/CV/CCV/CCCV
880	3	..V/CV/CVC/VCC/CCV/CCVC/CVCC/CCVCC/CCCV/VCCC см.F4
881	1	..V/VC/CV/CVC/VCC/CVCC/CCV/CCVC/CCCV/CCVCC
882	0	..CV/CVC/VVC/VCC/VCCC/CCV/CCVC/CCVCCC/ см.F4
883	1	..V/CV/VC/CVC/CCV/CVCCC/VCCC
884	1	..V/CV/VC/CVC/CVCC/CCVC/VCC
885	2	..V/VC/CV/CVC/CVCC/CCVC/VCC/CCV/CCVCC/CCCV
886	1	..V/VC/CVCVC/VCC/CVCC/CVCCC
887	1	..V/VC/CVC/CVCC
888	1	..V/C/VC/CV

Предполагается, что такой подход к представлению фактографических данных более достоверно выявит сходство и различие языков.

2.3.4. Формат представления данных: аргументы ЗА и ПРОТИВ

Как уже говорилось, перед тем, как сделать БД доступной для заинтересованной части научного лингвистического сообщества, в Институте Языкознания РАН была проведена самая тщательная экспертиза как формата представления, так и выбороч-

ной части содержания БД. Экспертами было высказано несколько критических замечаний, которые и обсуждаются в этом разделе¹⁵.

Описанная методика кодирования данных вызвала критические замечания экспертов, которые ориентируются на преимущественное использование БД в справочных целях. В частности, авторам высказывались нарекания в излишней «артефактности» БД. Действительно, некоторые следствия из принятых решений выглядят не совсем убедительно. Например, для описания значений признаков, имеющих целочисленный тип, приходится искусственно переводить эти значения в бинарный формат. В таблице 2.3 представлен пример такого описания для признака «..число монофтонгов» из раздела «2.1.1. ФОНЕМНЫЙ СОСТАВ». Также бинарная структура не позволяет кодировать плавные градации некоторых признаков. Например, признаки «..агглютинативные языки с элементами флексии» и «..флективные языки с элементами агглютинации» в разделе 2.3.0. МОР-ФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА явно несут в себе элементы нечеткости, степень которой пока невозможно определить и задать в рамках предложенного формата БД.

Т а б л и ц а 2.3

Пример описания для признака «..число монофтонгов» из раздела «2.1.1. ФОНЕМНЫЙ СОСТАВ»

Номер в БД	Частота признака	Наименование и уровень в иерархии признака
83	283	..число монофтонгов
84	0	...один
85	3	...два
86	2	...три
87	3	...четыре
88	32	...пять
89	27	...шесть
90	26	...семь
91	36	...восемь
92	28	...девять
93	27	...десять
94	14	...одиннадцать
95	14	...двенадцать
96	12	...тринадцать
97	14	...четырнадцать
98	5	...пятнадцать
99	12	...шестнадцать
100	5	...семнадцать
101	7	...восемнадцать
102	4	...девятнадцать
103	5	...двадцать
104	2	...двадцать один

¹⁵ Мы сейчас не касаемся проблемы ошибок в данных, о которой речь пойдет в разделе «Экспертиза содержания БД».

Номер в БД	Частота признака	Наименование и уровень в иерархии признака
105	2	...двадцать два
106	1	...двадцать четыре
107	1	...двадцать пять
108	1	...двадцать семь

В качестве другого примера наведенных явлений, т.е. «артефактов», приводится иерархическая структура БД. Ниже, в таблице 2.5 приводятся данные по разделам реферата и в том числе максимальное количество уровней в каждом из разделов. Видно, что глубина иерархической структуры колеблется от 2 до 7 уровней.

Аргументом против парадигматического принципа представления обычно высказывается то, что если в каком-то языке встретится неполная парадигма, то БД уже не позволит выявить сходство языков по этому параметру.

Приведем теперь аргументы ЗА принятый формат.

Ответом на критику бинарного формата представления данных служит то, что такой формат является весьма удачным с позиций квантитативных исследований, так как позволяет привести весьма разнородную грамматическую информацию к унифицированному виду. Кроме того, как показали многочисленные эксперименты, простой подсчет совпадающих бинарных признаков уже дает вполне адекватную меру близости языков.

Решение проблемы градуального описания признаков лежит в плоскости дальнейшей формализации градаций в бинарном формате и пополнении БД новыми признаками, расшифровывающими и уточняющими введенные ранее нечеткие характеристики. Полностью исключить градуальность на современном уровне развития лингвистики не представляется возможным. На этом же принципе построена и база данных WALS.

Против довода, касающегося «артефактности» бинарного представления целочисленных данных и парадигм можно возразить, что структура БД позволяет использовать фильтры и веса при квантитативных расчетах, которые помогают снизить вклад так называемых «артефактов» или устранить их влияние вовсе. Тем не менее, эти замечания экспертов планируется учесть, и, в частности, провести более детальное исследование вклада парадигматических и целочисленных признаков в результаты расчетов.

2.4. Характеристики модели реферата

По целому ряду причин описание разных языков выполнено с разной степенью подробности. Основная причина – это отсутствие полных сведений о той или иной части описания языка в самой энциклопедии, что, как правило, характерно для не полностью изученных, а также мертвых языков. Для некоторых языков изначально было сделано лишь краткое описание.

В таблице 2.4 представлено распределение языков по количеству признаков, для которых зафиксировано значение «Истина», что означает «Присутствует».

Анализ таблицы 2.4 показывает, что для полноценных статистических исследований полностью пригодна группа III языков БД (201-400) в количестве 258 единиц. О группах II и IV можно сказать, что для них можно выполнять частные сравнительные и типологические исследования на выборочном множестве признаков. При этом

в группе II необходимо делать поправку на недостаток информации, а в группе IV – на ее избыток. Причина «переописания» языков в группе IV по всей видимости связана с большим количеством доступной информации, и в некоторых случаях может внести статистическую погрешность. О языках группы I можно сказать, что информации в БД о них нет совсем или ее слишком мало для того, чтобы делать какие-то основательные выводы.

Т а б л и ц а 2.4

Распределение языков по числу признаков

Группа	Число признаков «Истина»	Число языков	В примечании в скобках за названием языка дается число признаков.
I	До 100	29	Чар-аймаков диал. гр.(40), пиктский (47), парфянский (нез.) (53), саргулямский (53), ванджский (56), трегами (57), кудрари (59), гуннов (86)
II	101-200	21	Сонкорско-тюркский (105), аланский (106), печенежский (111), скифский (126), лепонтийский (131), курдшули (134), башкарди диал. гр. (135), хорасано-тюркский (139), фуйнойских кыргызов (141), давани диалект (145), кумзари (нез.) (163), урумский (171), балканских тюрков (174), шотландский (177), ларская диал. гр. (179), тирахи (182), шумашти (182), болгарский (190), галльский (191), кельтиберский (196), мамлюкско-кыпчакский (197)
III	201-400	258	Все остальные языки
IV	свыше 400	8	Сингальский (401), эвенкийский (402), белорусский (405), латинский (409), чешский (нез.) (409), норвежский (415), немецкий (423), русский (536)

Рассмотрим структуру реферата на примере описания шведского языка (см. Приложение В)¹⁶. Реферат представляет совокупность бинарных признаков, выстроенных в виде иерархии. В реферате отражаются только те признаки, относительно которых есть достоверная информация, что они присутствуют. Отсутствие информации в БД кодируется специальным признаком **О** и обычно используется для маркирования самых верхних уровней иерархии. Пример такого маркера можно найти в реферате для признака 1.1.1 ВАРИАНТЫ НАЗВАНИЯ. Первая часть реферата (от пункта 0.0.0 до пункта 1.6.0 включительно и пункт 2.6.0 ЛЕКСИЧЕСКИЕ ЗАИМСТВОВАНИЯ) представляет собой обычное текстовое поле и используется в нашем исследовании только в справочных целях¹⁷. Она получила название в БД – «Индивидуальная часть». Вторая часть реферата (разделы 2.1.1. – 2.5.4.) представлена в бинарном виде и доступна для дальнейших манипуляций в исследовательских целях. Структура реферата строго соответствует структуре описания языка в энциклопедическом издании.

¹⁶ Описание шведского языка взято из (Языки мира: Германские языки. Кельтские языки, 1999), формализация в виде реферата и ввод в БД выполнены Е.И. Ярославцевой, экспертиза реферата выполнена в 2006 г. Д.Б. Никуличевой, корректировка по результатам экспертизы выполнена В.Н.Поляковым.

¹⁷ Нумерация разделов реферата строго соответствует нумерации разделов текстовой статьи в энциклопедии.

Количество точек перед названием признака маркирует уровень иерархии в системе языковых характеристик. Чем глубже уровень иерархии (больше точек), тем более частной является данная характеристика языка. Таким образом, в единой модели БД объединены и классифицирующие, и фактографические элементы описания языка. Можно спорить об удобстве такой формы описания, но у нее есть одно несомненное достоинство, эта структура инвариантна к описаниям языков любой сложности. Другое следствие – простота конвертирования бинарной структуры в числовое значение путем простого подсчета числа признаков со значением «Истина». Этот принцип лежит в основе расчета мер близости языков.

Как уже было указано, в БД содержится свыше 3800 признаков, в то время как для каждого языка фиксируется лишь небольшое количество от всей модели: 10-15 %. Этот факт еще раз подтверждает высокую степень языкового разнообразия. Было отмечено также, что, как правило, при заполнении описания для нового языка в модель добавляется небольшое количество новых признаков (в основном фактографического характера).

В таблице 2.5 приведены данные о количестве признаков в каждом разделе второй (бинарной) части общей модели.

Т а б л и ц а 2.5

Распределение признаков по разделам реферата

Раздел реферата	Общее число признаков	Число фактографических признаков	Максимальное число уровней в иерархии признаков
2.1.1. ФОНЕМНЫЙ СОСТАВ	437	371	7
2.1.2. ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ	131	105	4
2.1.3. ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ	167	145	4
2.1.4. СЛОГ	164	140	4
2.2.1. ФОНОЛОГИЧЕСК. СТРУКТУРА	131	107	5
2.2.2. ФОНОЛОГИЧЕСКИЕ ПРОТИВОПОСТАВЛЕНИЯ МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ	29	23	2
2.2.3. МОРФОЛОГИЧЕСКИ ОБУСЛОВЛЕННЫЕ ЧЕРЕДОВАНИЯ	22	12	4
2.3.0. МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА	58	40	4
2.3.1. КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ	7	4	2
2.3.2. ИМЕННЫЕ КЛАССИФИКАЦИИ	101	84	5
2.3.3. ЧИСЛО	111	99	3
2.3.4. ПАДЕЖНЫЕ ЗНАЧЕНИЯ	406	385	4
2.3.5. ГЛАГОЛЬНЫЕ КАТЕГОРИИ	763	721	4
2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ	308	255	4
2.3.7. ЧАСТИ РЕЧИ	73	71	4
2.4.0. ПАРАДИГМЫ	507	450	5
2.5.1. СТРУКТУРА СЛОВОФОРМЫ	184	166	3
2.5.2. СЛОВООБРАЗОВАНИЕ	40	32	3
2.5.3. ПРОСТОЕ ПРЕДЛОЖЕНИЕ	140	114	4
2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ	49	39	3

Как можно видеть из таблицы 2.5, большая часть признаков в модели описания языков является фактографической. Отношение фактографических признаков к общему числу по всей базе составляет 87,8%. По разделам это соотношение лежит в диапазоне от 54,5 до 97,2%.

Также очевидно, что вклад разделов в итоговую меру близости не одинаков по причине разного объема представленного в разделе материала. Самым влиятельным с количественных позиций окажется раздел 2.3.5. ГЛАГОЛЬНЫЕ КАТЕГОРИИ, самым невлиятельным – раздел 2.3.1. КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ.

2.5. Резюме по главе 2

1. База данных «Языки мира» содержит детальные описания почти всех языков Европы, Северной и Центральной Азии и является одной из двух крупнейших типологических баз данных в мире.
2. Архитектура БД ориентирована на поддержку математических методов исследований.
3. В области типологии и компаративистики языков существует ряд нерешенных проблем и спорных вопросов, для разрешения которых база данных и связанные с ней математические модели и методы потенциально могли бы быть применены.

Глава 3. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ БАЗЫ ДАННЫХ

3.1. Основные задачи, решаемые с помощью ядра БД

Программное обеспечение БД «Языки мира» ИЯ РАН состоит из ядра и инструментальных средств.

Программное обеспечение ядра БД «Языки мира» решает следующие задачи:

- 1) формирование и ведение модели и рефератов базы данных;
- 2) поиск информации;
- 3) парное сопоставление рефератов.

Программная реализация функций формирования и ведения БД позволяет осуществлять постоянное ее расширение как по горизонтали (ввод новых языков), так и по вертикали (ввод новых строк в модель реферата). Средства поиска и навигации позволяют найти нужный язык в списке языков, перейти к любому разделу реферата или модели, найти признак по строке или подстроке. Кроме того, в системе предусмотрен обратный поиск, который позволяет найти и сохранить в виде файла список языков, обладающих одним или совокупностью признаков. Принятая форма представления информации в БД позволяет осуществлять построчное сопоставление рефератов между собой и вычислять количественные показатели, характеризующие степень близости языков на структурном (грамматическом) уровне.

3.2. Windows-версия ядра БД

Windows-версия БД представляет собой 32-разрядное приложение, написанное на языке Delphi Pascal (версия 7). В качестве СУБД используется Borland DataBase Engine (BDE). Рабочая среда: Windows 95/98/2000/NT/XP. Объем инсталляционного варианта: 17,4 МБ. Объем программы вместе с БД (версия от 01.12.2006): 18,8 МБ.

3.3. Инсталляция БД

БД поставляется в комплекте с инсталляционным диском. Инсталляция проводится запуском файла Setup.exe. Во время инсталляции происходит запись программы на диск, установка ядра СУБД BDE, регистрация программы в главном меню системы Windows. Инсталляция не требует от пользователя специальных знаний в области программирования и компьютерных технологий.

3.4. Работа с базой данных «Языки Мира»

После запуска файла Langworld появляется окно программы, показанное на рис. 3.1.

В главном меню доступны функции **выбора группы, сравнения языков, поиск строк** в рефератах и поиск по индексу. Рассмотрим эти функции.

3.5. Парное сравнение языков на уровне классов и целого реферата

Для сравнения двух языков необходимо выбрать в меню пункт **Язык**, затем пункт **Сравнить**. Для сравнения выбирается в соответствующем списке название группы и название для исходного языка и языка, с которым необходимо сравнить данный. Далее необходимо нажать кнопку **Сравнить**.

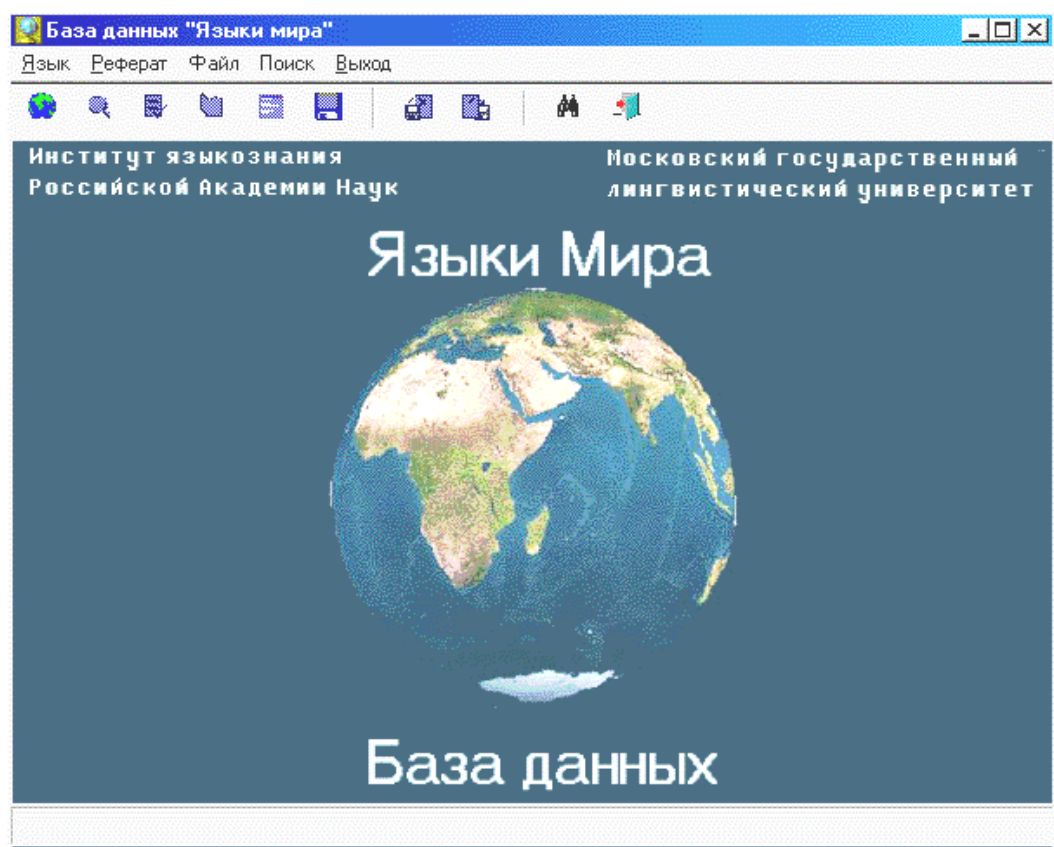


Рис. 3.1. Главное окно программы Langworld

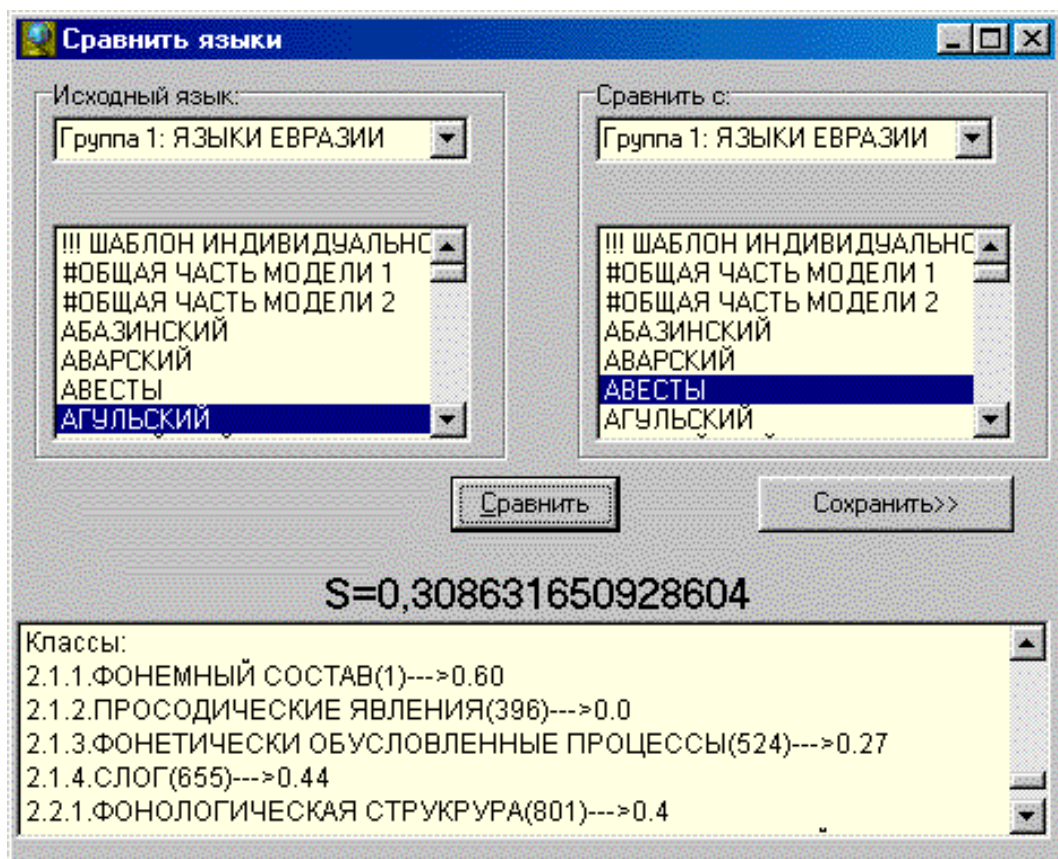


Рис. 3.2. Окно функции «Сравнить языки»

Внизу отображаются протокол сравнения по классам и итоговая величина S , которая вычисляется по формуле¹⁸:

$$S = \frac{\sum_{i=1}^m Skl[i]}{m} \quad (3.1)$$

где m – количество классов, а $Sk1 [i]$ – результат сравнения по классу, полученный следующим образом:

$$\begin{aligned} Skl [i] &= (F + S) / 2 \\ F &= \frac{\frac{Vs\ 1}{Vs} + \frac{Vs\ 1}{Vs\ 2}}{2} \\ S &= \frac{\frac{Kl\ 1}{Kl} + \frac{Kl\ 1}{Kl\ 2}}{2} \end{aligned} \quad (3.2.)$$

- Vs – вес всех строк класса исходного реферата;
- $Vs1$ – вес всех совпавших строк класса;
- $Vs2$ – вес всех строк класса второго реферата;
- $K1$ – количество всех строк класса исходного реферата;
- $K11$ – количество всех совпавших строк класса;
- $K12$ – количество всех строк класса второго реферата.

В формуле 3.2 под весом понимается уровень признака в иерархии.

Результаты сравнения можно сохранить в файл, нажав кнопку **Сохранить**. Пример сравнения двух рефератов приведен в Приложении Г.

3.6. Поиск строк в рефератах

Для поиска строк модели в рефератах необходимо в главном меню выбрать пункт **Поиск → Поиск строк**. Далее необходимо ввести искомые строки и установить между ними соотношение (на рис. 3.3: «Не» – ищутся все рефераты в которых данная строка не встречается, «Или» – ищется любая из строк, «И» – ищутся обе строки). Далее нужно нажать кнопку **Поиск** и через некоторое время появится результат. Список найденных языков можно сохранить в файл или продолжить с ним работу, введя новые значения строк и нажав кнопку **Искать в найденном**. Для русских символов надо учитывать регистр.

Описанную поисковую систему рекомендуется использовать для комбинированного поиска, зная уже значения строк модели. Посмотреть же значения строк модели в алфавитном порядке или в порядке, заданном структурой реферата, и найти одну из этих строк в рефератах можно с помощью поиска по индексу. Для этого из главного меню необходимо выбрать пункт **Поиск → Поиск по индексу**. Данный поиск удобен своей быстротой. Если пользователю необходимо найти конкретную строку, которая принадлежит конкретному классу, очень удобно использовать поиск по индексу. Для поиска необходимо выбрать интересующий вас тип представления индексов (в алфавитном порядке или предметном), выбрать строку и нажать кнопку **Поиск**. Получен-

¹⁸ В книге представлены результаты расчетов также и по другим формулам (см. главу 4).

ные результаты можно распечатать, сохранить в файл или продолжить с ними работу, выбрав другую строку и нажав **Искать в найденном** (рис. 3.4).

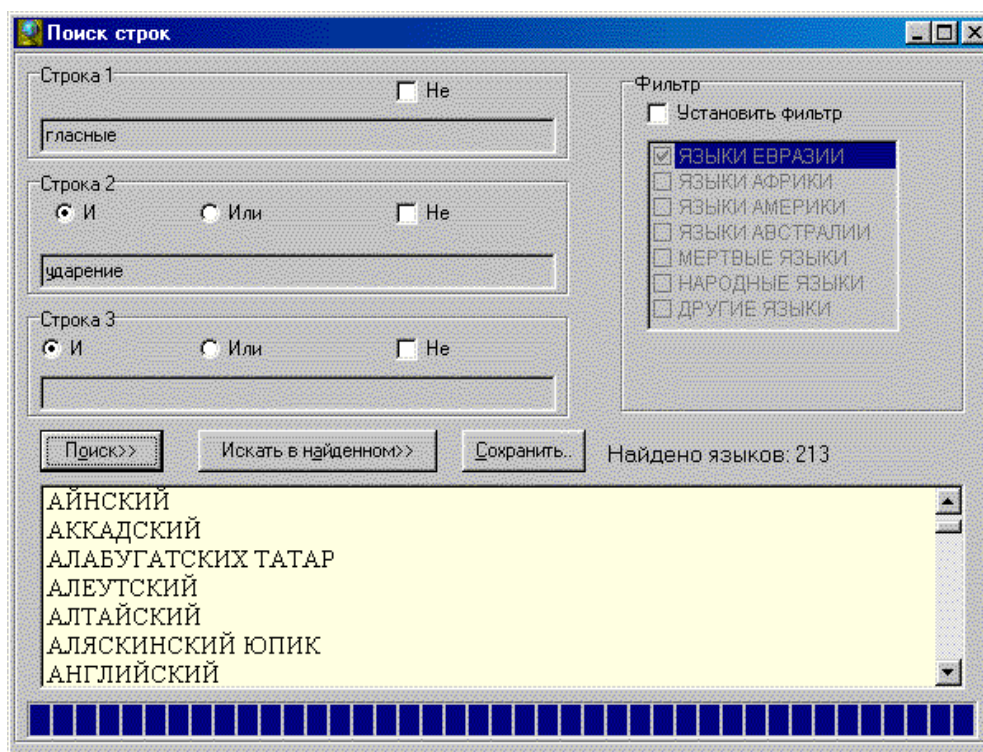


Рис. 3.3. Окно ввода параметров поиска

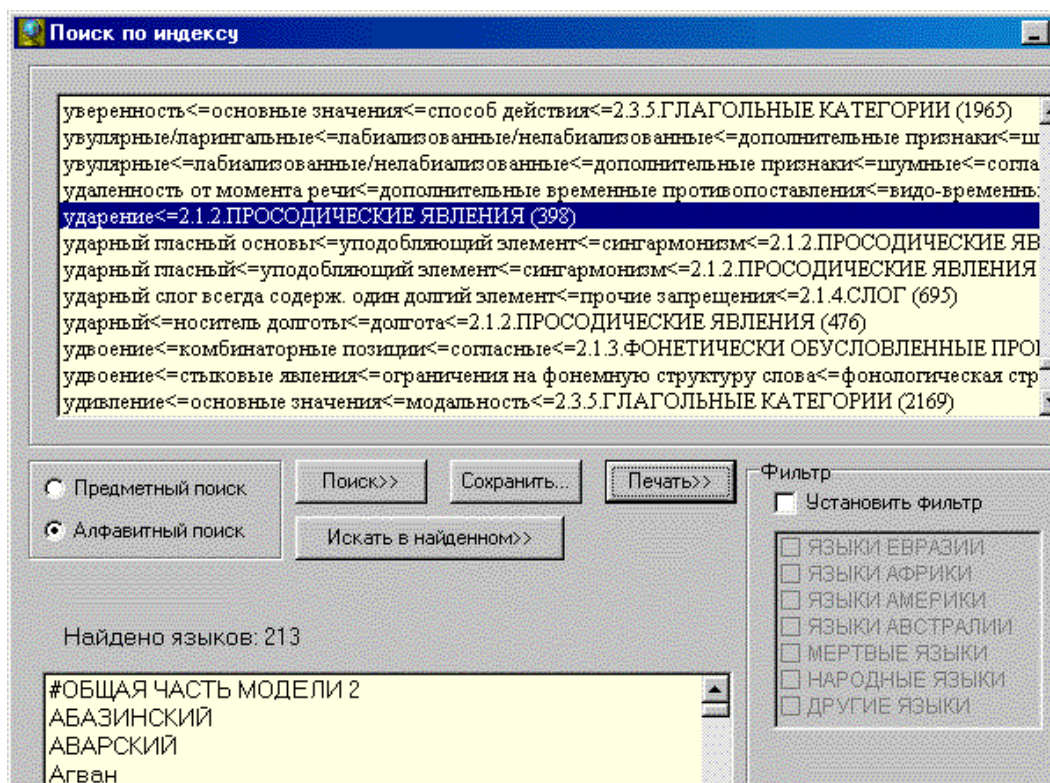


Рис. 3.4. Результаты поиска

Для удобства справа от индексов, представленных в алфавитном порядке, показаны классы и подклассы, к которым принадлежит данная строка с указанием ее порядкового номера. Для получения полной строки необходимо кликнуть мышкой по строке.

3.7. Работа с моделью и рефератом языка

Для работы с рефератом языка необходимо выбрать в главном меню **Язык → Выбрать группу** или нажать кнопку на панели. Далее выбирается группа языков. Можно просмотреть или добавить пояснение к группе, нажав кнопку **Пояснение**. Потом нажатием кнопки **Далее** в новом списке выбирается интересующий язык. Для просмотра и редактирования индивидуальной части реферата выбирается язык и нажатием правой кнопкой мышки – соответствующая команда во всплывающем меню. Также эта команда доступна в главном меню **Реферат → Индивидуальная часть**.

Для удаления языка правой кнопкой мыши делается клик на языке и выбирается функция **Удалить**. Далее необходимо подтвердить удаление. Эта команда доступна также на инструментальной панели (синий минус) или в главном меню **Язык → Удалить**. Для добавления языка правой кнопкой мыши на языке выбирается из меню функция **Вставить**. Эта команда доступна на инструментальной панели (синий плюс) и в главном меню **Язык → Вставить**.

Для редактирования названия языка необходимо кликнуть на поле языка и изменить значение. Далее необходимо нажать на инструментальной панели «галочку» для подтверждения изменений. Для отмены нажимается «крестик».

Для поиска языка в списке выбирается в меню пункт **Поиск → Поиск языка**. Для продолжения работы с рефератом необходимо нажать кнопку **Далее**. В открывшемся списке категорий модели можно выбрать те строки, которые входят в реферат. Для этого надо рядом с категорией поставить «галочку». Также можно открыть и сохранить реферат.

3.7.1. Команды главного меню

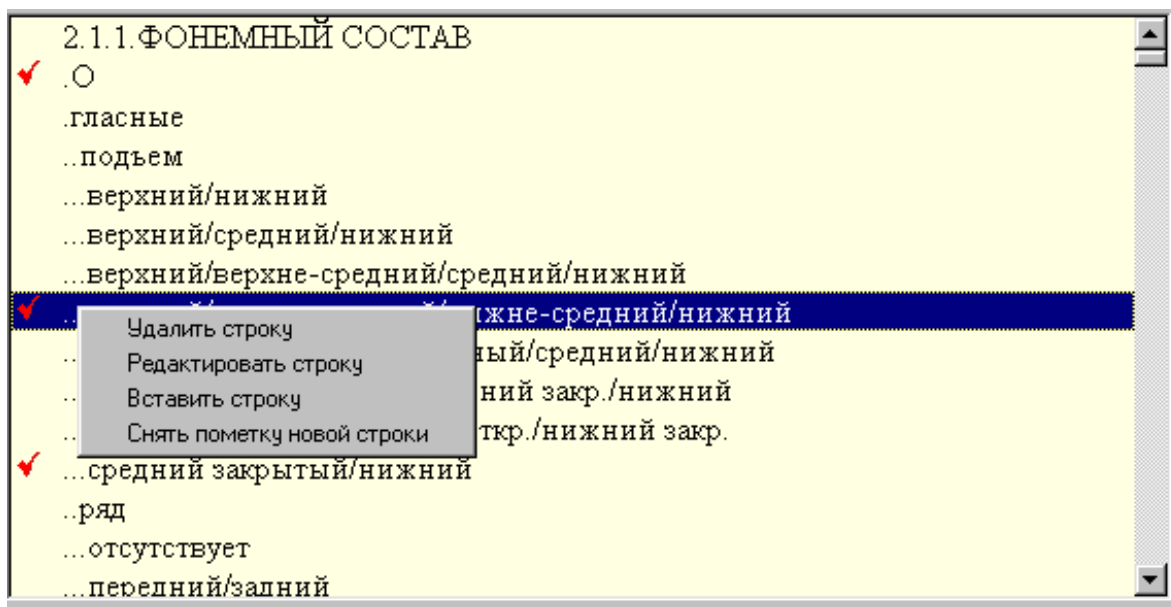


Рис. 3.5. Редактирование модели реферата

Команда **Реферат → Очистить** снимает отметку (красную «галочку») со всех строк модели.

Команда **Реферат → Модель → Редактировать** осуществляет переход к редактированию, добавлению и удалению категорий модели. Чтобы удалить категорию модели необходимо выбрать категорию, правой кнопкой мыши кликнуть на строке и выбрать команду **Удалить строку**. Также в этом меню доступны команды **Вставить строку**, **Редактировать строку** и **Снять пометку новой строки**. Новая строка добавляется после выделенной курсором строки. Для подтверждения редактирования строки в режиме **Редактировать строку** надо нажать клавишу **Enter**. Все новые строки помечаются красной «галочкой».

Команда **Реферат → Индивидуальная часть** обеспечивает работу с индивидуальной частью реферата. Для подтверждения редактирования необходимо нажать **Да**, в обратном случае – **Отмена**.

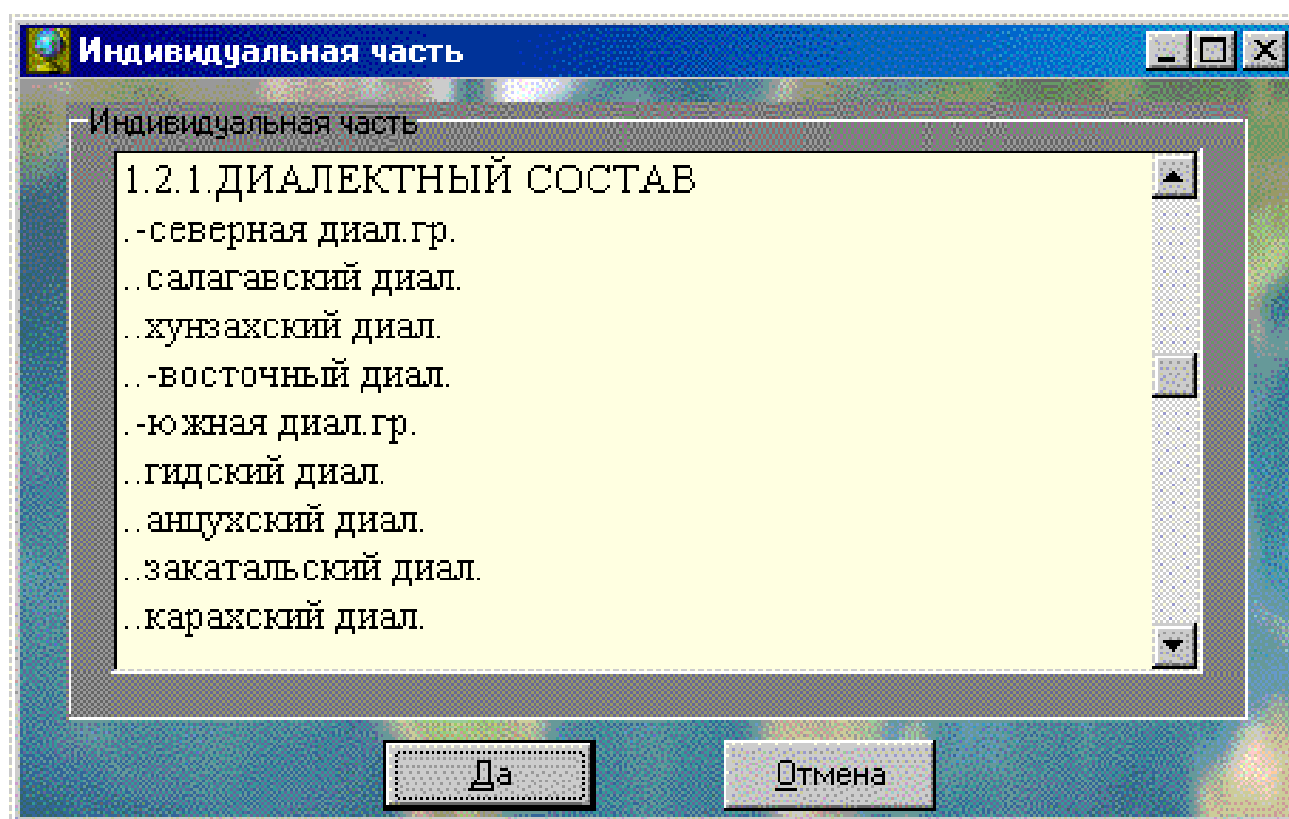


Рис. 3.6. Редактирование индивидуальной части реферата

Команда **Реферат → Инспекция** дает переход к инспекции реферата. Инспектируются все отмеченные строки модели, т.е. те строки, которые составляют реферат.

Команда **Реферат → Редактировать**: переход к редактированию реферата языка. В открывшемся списке категорий модели можно выбрать те строки, которые входят в реферат. Для этого надо рядом с категорией поставить галочку (рис. 3.7).

Реферат → Сохранить: сохранение изменений в реферате.

Реферат → Сохранить индивидуальную часть: сохранение изменений в индивидуальной части реферата (используется после открытия реферата из файла).

Реферат → Печать: печать реферата вместе с индивидуальной частью.

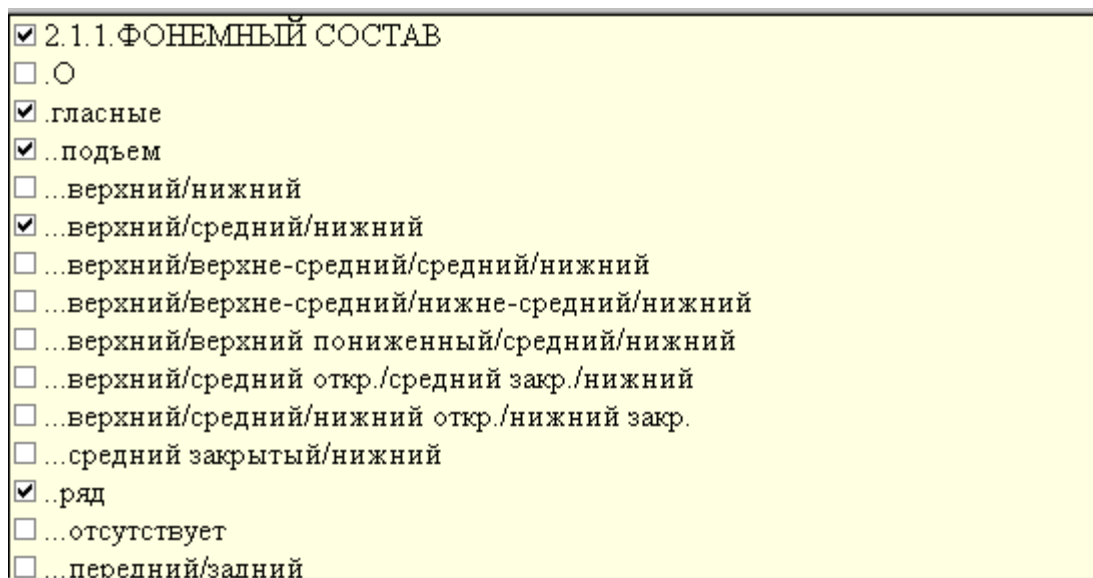


Рис. 3.7. Редактирование бинарной части реферата

3.7.2. Загрузка реферата в БД из файла

Файл → Открыть. Выбирается в диалоге нужный файл (*.txt). Далее выбирается кодировка, в которой выполнен файл (Win или Dos). Если какая-нибудь строка не найдена в модели, то выводится соответствующее предупреждение и заполнение реферата прекращается. Для правильного открытия после этого необходимо либо добавить в нужное место модели строку, либо удалить ее из файла. При открытии файла нужные строки модели отмечаются в реферате и временно заносятся в базу данных. Для окончательного занесения в базу данных необходимо выбрать **Реферат → Сохранить** и далее **Реферат → Сохранить индивидуальную часть**. Программа сама спросит пользователя о сохранении при вызове следующей функции из меню или при нажатии на кнопку **Закончить**.

3.7.3. Сохранение реферата в файл

Файл → Сохранить. Выбрать кодировку.

Файл → Сохранить новые строки модели (пункт доступен во время режима **Реферат → Модель → Редактировать**) – сохранение новых строк модели в файл.

После окончания работы с рефератом нажмите кнопку **Закончить**, и программа вернется в исходную позицию. Во время работы с рефератом недоступны функции поиска. Для поиска необходимо вернуться в начало программы.

3.8. Навигация в пределах модели и реферата

Для быстрой навигации в пределах модели реферата в системе предусмотрены две функции: **Найти признак** и **Перейти к разделу**. Переход к этим функциям осуществляется через контекстное меню (правая кнопка мыши) в режимах **Реферат → Редактировать** и **Реферат → Модель → Редактировать**.

3.9. Разработка Web-версии БД

Учитывая уникальность БД, ее потенциальные возможности для дальнейших научных исследований в области количественной типологии и компаративистики, с целью предоставления доступа к ней широкому кругу специалистов, аспирантов и сту-

дентов, в ИЯ РАН были выполнены работы по созданию Web-версии БД. Сейчас идет апробация пилотного варианта Web-версии, которая стала доступна в сети Интернет по адресу www.dblang.ru, начиная с 15 марта 2006 г.

Web-версия БД представляет собой серверное Интернет-приложение, созданное в среде ASP.NET на язык C# с использованием СУБД MS SQL. Содержимое Web-версии БД полностью идентично Windows-версии. В дальнейшем планируется их полная синхронизация в случае пополнения БД новыми языками.

3.9.1. Просмотр сайта

Запуск Web-версии БД и ее просмотр осуществляется в среде Интернет-браузера путем набора в адресной строке URL www.dblang.ru. Сайт был оптимизирован для браузера MS Internet Explorer (версия 6.0).

После загрузки главной страницы рабочее поле браузера выглядит следующим образом (рис. 3.8):

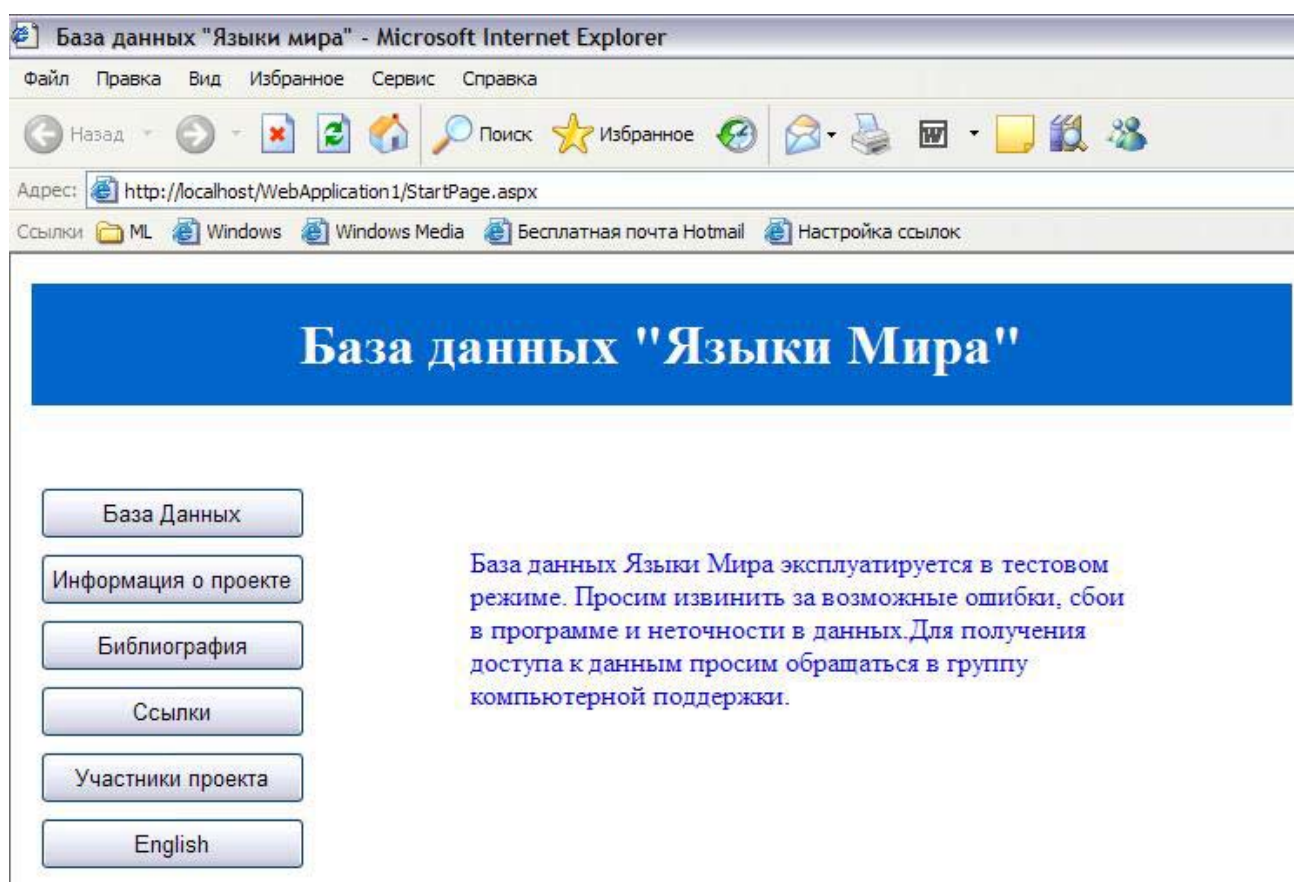


Рис. 3.8. Главная страница Web-версии БД

Переход к базе данных осуществляется нажатием на одноименную кнопку. После ввода имени и пароля пользователь получает доступ к списку языков.

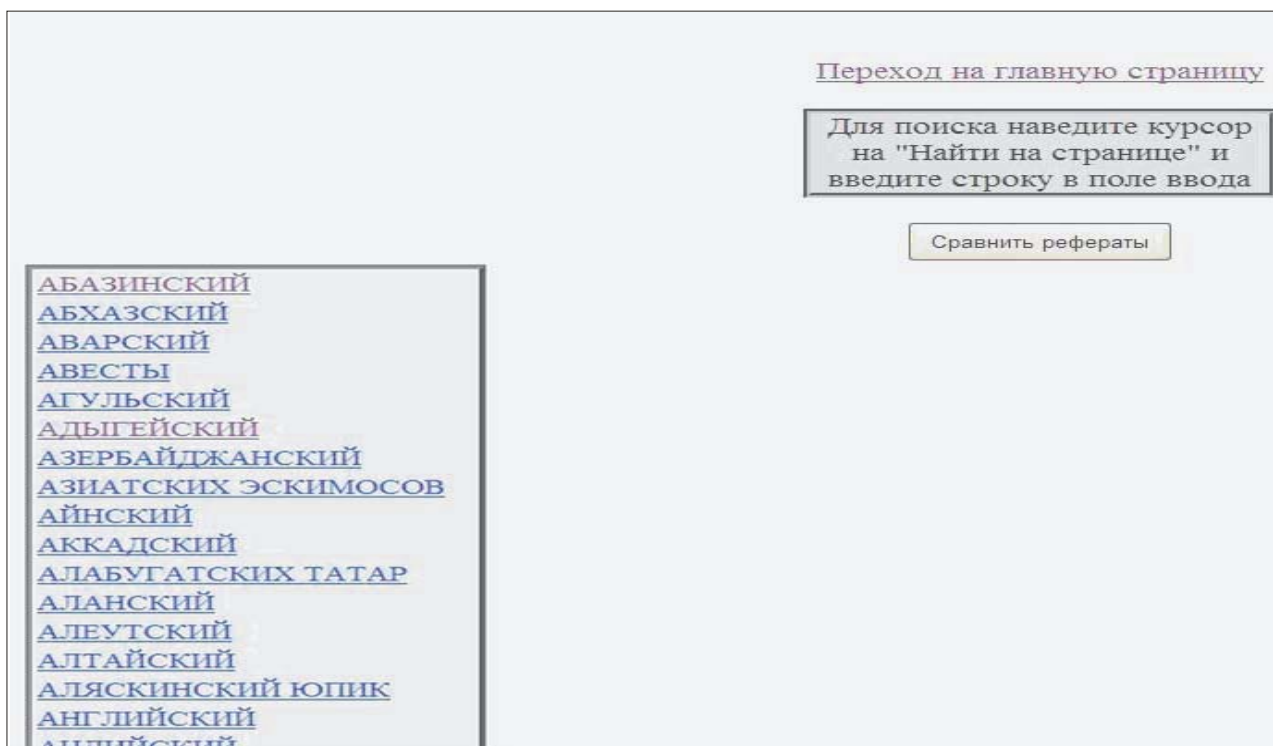


Рис. 3.9. Список языков Web-версии БД

3.9.2. Просмотр реферата языка

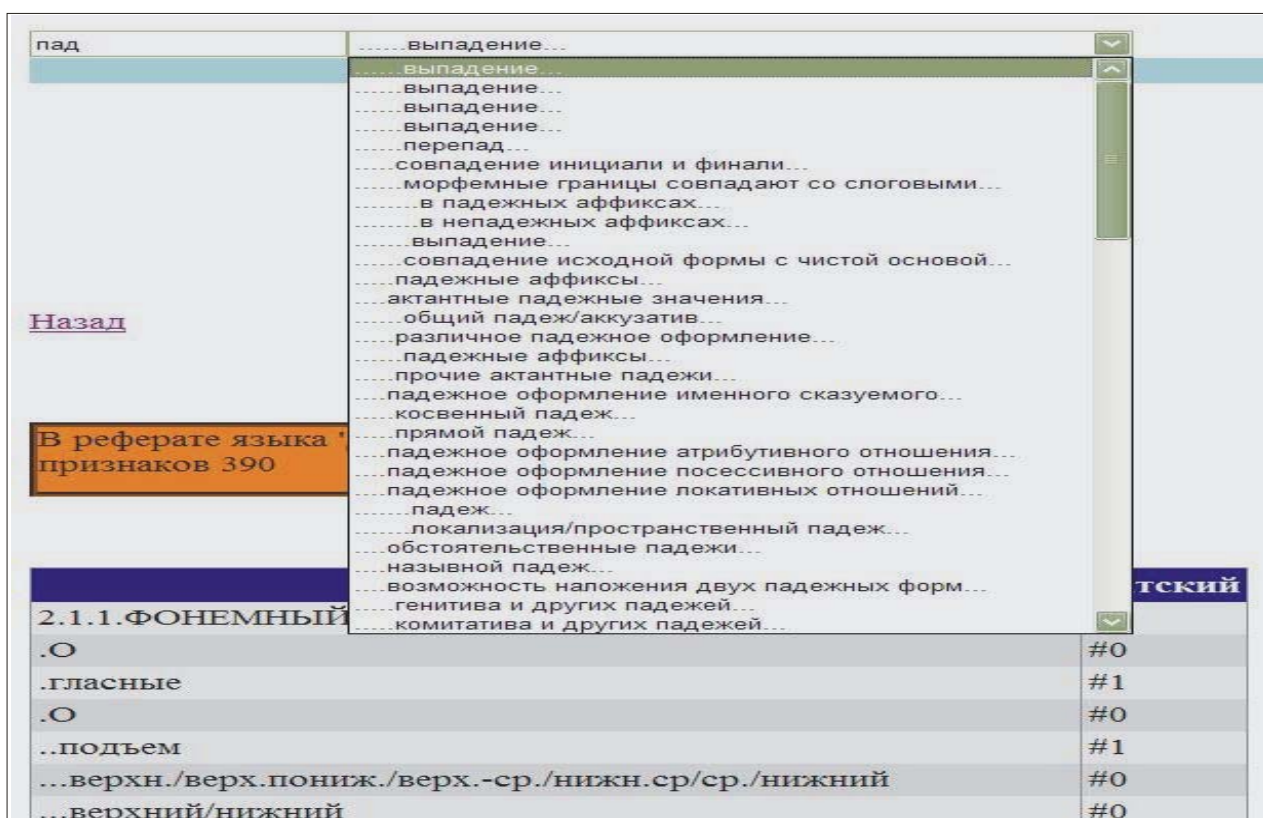


Рис. 3.10. Вид реферата в сети Интернет: режим поиска

Для просмотра реферата языка пользователь должен выбрать имя языка из списка и вызвать его. Для облегчения навигации по списку языков предусмотрен режим быстрого поиска. В результате на экране появляется реферат в виде списка признаков (рис. 3.10)

Для облегчения поиска нужного признака или раздела в БД предусмотрена навигация в виде окна быстрого поиска и последующего выбора строки реферата из списка (показано на рис. 3.10).

3.9.3. Сравнение рефератов

При необходимости можно сравнить два реферата. Для этого с помощью кнопки «Сравнить рефераты» эксперт переходит к диалогу выбора языков из списка.

3.10. Экспертиза БД

Выборочная экспертиза БД¹⁹ показала наличие ошибок и неточностей, которые в первую очередь связаны с большим объемом водимых данных (в БД описано свыше 1,2 млн. бинарных состояний). Количество ошибок по разным языкам колеблется в широких пределах. Рассмотрим для примера результаты экспертизы трех скандинавских языков германской группы (датский, норвежский, шведский; экспертизу проводила Д.Б. Никуличева). Для шведского количество признаков, присутствующих в языке, составляет 378. Обнаружено 50 ошибок, что от числа заполненных признаков составляет 13,2 %, от общего числа признаков (3800) – 1,3 %. При этом 18 ошибок (4,7%) обнаружено непосредственно в печатном издании энциклопедии. Для норвежского количество признаков, изначально заполненных в базе данных составляет 415. Обнаружено 36 ошибок (8,6 %), от общего числа – 0,9 %. 12 ошибок (2,9 %) обнаружено непосредственно в энциклопедии. Для датского количество признаков в базе данных составляет 391. Обнаружено 85 ошибок, что от числа заполненных признаков составляет 21,7 %, от общего числа признаков – 2,2 %. 18 ошибок (4,6 %) обнаружено в энциклопедии. То есть мы видим, что процент ошибок колеблется от 9 до 22 % (от 1 до 2 % в пересчете на весь объем реферата) и существенная доля этих ошибок зафиксирована непосредственно в самой энциклопедии. Эти данные являются весомым аргументом в пользу проведения дальнейшей экспертизы БД.

Вместе с тем, результаты предварительных расчетов, выполненных на существующих данных, убедительно показали возможность получения качественной меры близости и выполнения на ее основе дальнейших количественных расчетов.

3.11. Назначение БД

База данных Языки мира имеет три основных области применения:

- она может быть использована как основа для количественных исследований;
- она представляет интерес как справочный лингвистический ресурс;
- и может быть использована для учебных целей.

3.11.1. Количественные исследования

База данных «Языки мира» с самого начала ее создания была ориентирована преимущественно на количественные исследования. Эта целевая установка во многом определила существующий формат представления данных и набор инструментальных средств. В настоящее время количественные исследования сосредоточены по трем

¹⁹ Экспертиза проводилась в рамках Российского научно-образовательного центра по лингвистике в 2006 г.

направлениям (классам задач): расчет мер парной близости языков; кластеризация языков и интеллектуальный анализ данных; расчет количественных портретов языковых сообществ (макро-семей, семей, ветвей, групп и подгрупп). Разумеется, что решение этих количественных задач не является самоцелью, а подчинено достижению предметных целей, среди которых можно было бы выделить общие и частные.

Общей целью является разрешение хорошо известных проблем сравнительно-исторического языкознания:

- исследование общих закономерностей языковой эволюции и языкового разнообразия;

- выявление закономерностей наследования и приобретения структурных черт языков;

- выявление причин и факторов, определяющих типологические черты языков, типологическая классификация языков;

- формирование общей генеалогической языковой картины.

Среди частных целей, способствующих достижению общих, можно было бы назвать:

- повышение надежности методики выявления языкового родства;

- подтверждение или опровержение известных генетических гипотез, выдвижение новых;

- выявление, подтверждение или опровержение ареальных контактов;

- анализ общего структурного строя языков.

Для решения перечисленных задач, кроме ядра БД, в исследовательской группе выполнялись разработки различных инструментальных средств. К таким программным средствам относятся:

- программа Similarity – для расчета мер парной близости языков;

- программа LangFam – для расчета языковых портретов и выявления генетических маркеров;

- «Программный комплекс для проведения статистических исследований в области компаративистики» – преимущественно для выполнения автоматической кластеризации языков.

Для выполнения интеллектуального анализа данных использовался статистический пакет R.

Методы и результаты количественных исследований описаны в главах 4, 6-7. В главе 5 представлена модель структурной эволюции языков. Предсказанные моделью количественные закономерности эволюции языков находят хорошее подтверждение на материале БД.

3.11.2. Справочные цели

База данных представляет собой уникальный справочный ресурс. Она содержит описание грамматической структуры 315 языков Евразии. Кроме того, модель реферата представляет собой самостоятельную ценность, так как является концентрированное воплощение опыта и знаний нескольких лингвистических школ в области типологии и компаративистики. Размещение БД в сети Интернет сделало ее доступной для всего научного сообщества.

3.11.3. Учебные цели

По материалам Базы данных «Языки мира» и инструментальным средствам была создана учебная программа «Базы данных для типологических и компаративистских исследований»²⁰. По этой программе были прочитаны спецкурсы на филфаке МГУ (ОТиПЛ, Москва) и на факультете «Лингвистика» ЮУрГУ (г. Челябинск).

3.12. Резюме по главе 3

1. Разработан комплекс инструментальных средств доступа к данным, позволяющий выполнять ввод и редактирование описания языков в заданном формате, их поиск, а также парное сравнение языков.
2. БД доступна в различных вариантах: существуют DOS-, Windows-, Web-, Excel-версии.
3. БД может использоваться для исследовательских, учебных и справочных целей.

²⁰ Авторы: В.Н. Поляков, В.Д. Соловьев.

Глава 4. НОВЫЕ ВОЗМОЖНОСТИ КВАНТИТАТИВНЫХ ИССЛЕДОВАНИЙ

4.1. Расчет мер парной близости языков

Меры близости являются удобным способом оценки близости сложных объектов, свойства которых заданы набором дискретных параметров. Большое внимание мерам близости в квантитативных расчетах уделяется потому, что они являются надежной основой для автоматической кластеризации языков. Также меры близости используются в качестве исходных данных в алгоритмах интеллектуального анализа данных (Data Mining).

Выбор наилучшей меры является сложной задачей, что определяется целым рядом обстоятельств.

Во-первых, существует большое разнообразие подходов к расчету мер, что само по себе затрудняет выбор удачной формулы или алгоритма.

Во-вторых, язык, а точнее, структура грамматического строя языков представляет собой чрезвычайно сложный информационный объект. Это нашло свое отражение в структуре модели описания языка в базе данных.

В-третьих, большой объем данных в БД, накладывает существенные ограничения на прямые эксперименты по выбору мер близости.

В-четвертых, надо хорошо понимать, что численная характеристика близости, рассчитанная по выбранной формуле, отражает не только следы родства, но и следы заимствований в языке. Таким образом, проблема устранения типологического фона становится чуть ли не главной при выборе наилучшей меры близости.

Именно поэтому выбор, или, возможно, конструирование новой меры близости, которая удачно передаст генетическое родство языков, является непростой экспертной задачей, в которой проявляются не только знания, но и интуиция исследователя. В связи с этим, прежде чем сравнивать меры близости, было бы разумным сначала определить критерии их качества.

4.1.1. Подходы к расчету мер близости

Согласно (Раушенбах, 1985), меры близости объектов $X=\{x_1, x_2 \dots x_n\}$ и $Y=\{y_1, y_2 \dots y_n\}$, описываемых дихотомическими признаками x_i и y_i , принято выражать в виде функции четырех переменных:

$$S=F(a,b,c,d), \quad (4.1)$$

где a — число признаков, отсутствующих у X и Y одновременно, d — число совпадающих признаков, b (или c) — число признаков, присутствующих у X , но отсутствующих у Y (или наоборот). Сумма $(a + b + c + d)$ равна общему числу признаков n .

Среди наиболее часто используемых мер в (Раушенбах, 1985) выделяются следующие:

$$S_1 = \frac{a+d}{a+b+c+d} ; \quad (4.2)$$

$$S_2 = \frac{2(a+d)}{2(a+d)+b+c} ; \quad (4.3)$$

$$S_3 = \frac{a+d}{a+d+2(b+c)} ; \quad (4.4)$$

$$S_4 = \frac{d}{a+b+c+d} ; \quad (4.5)$$

$$S_5 = \frac{d}{b+c+d} ; \quad (4.6)$$

$$S_6 = \frac{a}{a+b+c} ; \quad (4.7)$$

Меры близости S_1 - S_6 могут быть представлены в виде функций от $v(X)$, $v(Y)$, $v(X \cap Y)$ или $v(X \cup Y)$. Это можно объяснить тем, что близость объектов X и Y определяется наличием у них общих признаков, информационная ценность которых ($v(X \cap Y)$) должна быть неким образом соотнесена с информацией об объектах, рассматриваемых порознь ($v(X)$, $v(Y)$) или в совокупности ($v(X \cup Y)$). Сходство объектов может быть измерено не только по наличию, но и по отсутствию у них одних и тех же признаков – «сходство по антипатиям». Нормировка мер близости связана с ограничением ее значения сверху.

Можно сконструировать общее выражение для мер, представленных в формулах (4.2- 4.7):

$$S_7 = \frac{R1*a + R4*d}{R1*a + R2*b + R3*c + R4*d} , \quad (4.8)$$

где $R1 \dots R4$ – коэффициенты, определяющие вклад в общую меру одновременного присутствия признаков (параметр a), одновременного отсутствия признаков (d) или противоречий (b и c) у сравниваемых объектов.

4.1.2. Учет структуры и объема признакового пространства

С одной стороны, бинарное кодирование признаков представляется удачным способом организации данных, так как позволяет поставить в соответствие значению «ИСТИНА» число 1, а значению «ЛЮЖЬ» – число 0. Это весьма распространенный подход при подсчете близости на основе бинарных дихотомических данных. С другой стороны, существует ряд особенностей в описании языков в БД (иерархия признаков, парадигмы, разделы реферата и другие), которые вносят факторы вариативности и расширяют признаковое пространство. Рассмотрим этот вопрос подробнее.

Формулы (4.2-4.8) описывают сравнительно простые меры близости, построенные по аддитивному принципу. Коэффициенты, используемые в формулах (4.3), (4.4) и (4.8), структурируют признаковое пространство модели по принципу наличия/отсутствия и совпадения/несовпадения признаков у объектов X и Y . Эти выражения можно усложнить, если каждому признаку поставить в соответствие вес w , кото-

рый в свою очередь может являться функцией от частоты f или уровня L в иерархии признаков.

$$a = \sum_{i=1}^n w_i * p_i, \quad (4.9)^{21}$$

где $w_i^f = W(f)$, если мы используем весовую частотную меру, или $w_i^L = W(L)$, если мера весовая иерархическая.

Для подсчета a :

$$p_i = \begin{cases} 0, & \text{если } (x_i \neq y_i) \wedge (x_i = y_i = 0); \\ 1, & \text{если } x_i = y_i = 1; \end{cases} \quad (4.10)$$

для подсчета d :

$$p_i = \begin{cases} 0, & \text{если } (x_i \neq y_i) \wedge (x_i = y_i = 1); \\ 1, & \text{если } x_i = y_i = 0. \end{cases}$$

Информация о частоте того ли иного признака или его уровне в иерархии в БД в явном виде не представлена, однако без труда может быть эксплицирована. Частота f для признака i задается как отношение

$$f = N^+ / N_{\text{я}}, \quad (4.11)$$

где N^+ – количество языков в БД, в которых встречается признак i , $N_{\text{я}}$ – общее количество языков в БД.

Использование такой континуальной переменной, как частота признака и веса, построенного на ее основе, расширяет число вариантов мер близости до бесконечности. Вариативность параметрического пространства определяется тремя возможными способами задания веса, а также их комбинациями.

Способ 1. Вес используется как пороговая функция для принятия решения о включении/невключении признака i в расчет меры. Например, с помощью выражения (4.12) будут «отсекаться» низкочастотные признаки.

$$w_i = \begin{cases} 0, & \text{если } f_i < f_0 \\ 1, & \text{если } f_i \geq f_0 \end{cases} \quad (4.12)$$

Способ 2. Вклад частоты можно учитывать не линейно, а с помощью специально подобранной функции. Например, можно задать

$$w_i = f^2. \quad (4.13)$$

²¹ Выражения для d не приводим, так как его можно получить из (4.9) по аналогии.

Способ 3 (как разновидность способа 2). Частотные веса могут рассчитываться «от обратного»:

$$w_i = (N_j - N^+)/N_j. \quad (4.14)$$

Аналогичные частотным могут быть сконструированы весовые меры, в которых вес будет являться функцией от уровня признака в иерархии. Максимальное количество уровней в глубину в БД равно семи²². Для весовых иерархических мер, также как и для частотных, возможно задание порога отсечения признаков по номеру уровня как снизу, так и сверху. Можно сконструировать нелинейные весовые функции от номера уровня и в том числе обратную функцию.

Рассматривая возможности структурирования признакового пространства в базе данных «Языки мира» на основе иерархической организации, необходимо отметить такую важную дихотомию признаков, как «Классифицирующие/Фактографические». Будем называть фактографическими все признаки, которые оказываются на самом нижнем уровне иерархии в представлении той или иной категории структурного описания языков.

Обратимся к фрагменту модели из раздела 2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ (табл. 4.1). Как можно видеть из таблицы для категории *местоимения и местоименные слова* признак *..разряды местоимений* являются классифицирующими, в то время как признаки (*...адвербиальные, ...анафорические, ...атрибутивно-указательные, ...безличные, ...взаимные, ...взаимно-возвратные, ...взаимно-личные, ...взаимно-притяжательные*) попадают в категорию фактографических. Далее, признак *...возвратные* является классифицирующим, а расшифровывающие его признаки *...инклюзивность/эксклюзивность* и *...субъектно/объектные* – вновь фактографическими. Хотелось обратить внимание на то, что признак *...возвратные* попал в разряд классифицирующих именно по той причине, что имеет подчиненные признаки более низкого уровня, в то время как его «коллеги» с того же третьего уровня трактуются в системе как фактографические.

Т а б л и ц а 4.1

Пример фактографических признаков

2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ	Номер уровня	Тип признака
.О		
.местоимения и местоименные слова	1	Классифицирующий
..разряды местоимений	2	Классифицирующий
...адвербиальные	3	Фактографический
...анафорические	3	Фактографический
...атрибутивно-указательные	3	Фактографический
...безличные	3	Фактографический
...взаимные	3	Фактографический
...взаимно-возвратные	3	Фактографический

²² Нулевым уровнем считается заголовок раздела в модели реферата.

...взаимно-личные	3	Фактографический
...взаимно-притяжательные	3	Фактографический
...возвратные	3	Классифицирующий
...инклюзивность/эксклюзивность	4	Фактографический
...субъектно/объектные	4	Фактографический

В базе данных информацию об этом типе признаков (Классифицирующие /Фактографические) можно эксплицировать путем анализа уровней подчиненности. Отключение нужной нам категории можно организовать путем наложения специального фильтра.

Также представляется интересным провести исследование вклада различных частей модели в расчет общей меры близости. Таким образом, отнесенность признака к разделу реферата является еще одним параметром при конструировании меры близости.

Сводная таблица мер близости, которые позволяет рассчитывать существующая структура БД, приведена в таблице 4.2.

Подведем предварительные итоги анализа возможностей структурирования признакового пространства БД. При разработке мер парной близости были выработаны следующие общие подходы:

- Мера близости рассчитывается на основе совпадения/несовпадения признаков имеющих бинарный характер.

- В простейшем случае при совпадении признака у двух языков мера увеличивается на единицу.

- Мера может нормироваться на величину максимального совпадения, тогда она принимает значение от 0 до 1.

- При расчете простых мер близости можно учитывать следующие параметры²³:

- Тип признака (классифицирующий/фактографический)
- Значение признака (одновременное наличие/одновременное отсутствие)
- Уровень признака в иерархии
- Частоту встречаемости признака в БД
- Можно ограничить расчет меры каким-либо разделом БД, значением частоты, уровнем в иерархии
- Можно ввести вес признака в зависимости от его частоты/уровня в иерархии
- Весовые коэффициенты могут быть как прямыми (возрастать), так и обратными (убывать)
- Можно строить комбинированные меры близости на сочетании разнообразных факторов.

²³ Подробную информацию о структуре данных в базе можно найти в публикациях на сайте www.dblang.ru.

Сводная таблица основных мер близости

Тип меры	Мера близости	Сводная формула
Простые аддитивные	Все признаки	$S_{\text{все}} = \frac{a+d}{a+b+c+d}$
	Только присутствующие в языках	$S_{\text{прис}} = \frac{a}{a+b+c+d}$
	Только отсутствующие ²⁴	$S_{\text{отс}} = \frac{d}{a+b+c+d}$
	Только классифицирующие	$S_{\text{клас}} = \frac{a_{\text{клас}} + d_{\text{клас}}}{a+b+c+d}$
	Только фактографические	$S_{\text{факт}} = \frac{a_{\text{факт}} + d_{\text{факт}}}{a+b+c+d}$
	Только присутствующие классифицирующие	$S_{\text{прис.клас}} = \frac{a_{\text{клас}}}{a+b+c+d}$
	Только присутствующие фактографические	$S_{\text{прис.факт}} = \frac{a_{\text{факт}}}{a+b+c+d}$
Весовые	$S_{\text{вес}} = \frac{a_{\text{вес}} + d_{\text{вес}}}{a+b+c+d}; \dots \text{ и т.д.}; a_{\text{вес}}, d_{\text{вес}} = \sum_{i=1}^n w_i * p_i;$	
	Весовая частотная, линейная прямая	$w_i = f_i$
	Весовая частотная, квадратичная прямая	$w_i = f_i^2$
	Весовая частотная, линейная обратная	$w_i = 1/f_i$
	Весовая иерархическая, линейная прямая	$w_i = k * L_i$
	Весовая иерархическая, линейная обратная	$w_i = 1/(k * L_i)$
Компоненты меры близости		
a – число совпадающих присутствующих признаков (x=1, y=1)		
b – число несовпадающих признаков (x=0, y=1)		
c – число несовпадающих признаков (x=1, y=0)		
d – число совпадающих отсутствующих признаков (x=0, y=0)		
N=a+b+c+d – общее число признаков		
p – фактор, задающий условие совпадения признаков. Формула (4.10)		
f =n/N, частота встречаемости признака в БД		
n – число языков в БД, в которых признак присутствует		
L – уровень признака в иерархии		
k – коэффициент, задающий вклад каждого уровня в вес		

²⁴ Как видно из таблицы 4.5, эта категория также представляет интерес, так как существует семья (монгольская), для которой значение показателя качества меры K_i достигает максимума только в этом варианте расчетов.

Фильтры (условия отбора признаков)²⁵	
По частоте	Значение частоты лежит в заданном диапазоне: $f_1 \leq f \leq f_2$
По уровню признака в иерархии	Значение уровня лежит в заданном диапазоне: $L_1 \leq L \leq L_2$
По разделу реферата	Признак j принадлежит разделу R : $j \in R$.
Нормировка	Для языка i : $S_i \text{ норм} = S_i/S \text{ макс}$

Принципиально имеется возможность конструирования более сложных мер близости, как на основе подбора весовых коэффициентов (см. например формулу 3.1 в главе 3), так и на основе различных методов свертки информации из БД.

Как можно видеть, возможности варьирования параметров при расчете меры парной близости языков весьма широки. Это, с одной стороны, дает основания надеяться на то, что путем экспериментальных расчетов можно подобрать формулу и алгоритм для расчета наилучшей меры близости. С другой стороны, такое разнообразие в сочетании с большими затратами времени на полный расчет мер близости по всему списку языков в БД определило необходимость разработки научно-обоснованной методики отбора мер близости и критериев этого отбора.

4.1.3. Методика проверки расчетов

Большое разнообразие способов конструирования мер парной близости языков не позволяет априорно утверждать о преимуществах той или иной меры близости. Это заставило авторов искать пути верификации расчетов. Предложенная оригинальная методика проверки расчетов основывается на эталонной выборке из 48 языков по 8 языковым семьям (см. табл. 4.3)²⁶. Мера близости по языкам последних трех строк считается, однако группы не используются для оценки качества меры.

Т а б л и ц а 4.3

Эталонная выборка языков для проверки качества меры близости

Группа	Языки	Язык-прототип	Ng	Ki
Уральские	Эстонский, финский, вепсский, венгерский, хантыйский	Финский	5	K ₁
Тюркские	Татарский, башкирский, турецкий, азербайджанский, туркменский	Турецкий	5	K ₂
Монгольские	Монгольский, калмыцкий, бурятский, монгорский	Калмыцкий	4	K ₃
Славянские	Русский, белорусский, польский, болгарский, македонский	Белорусский	5	K ₄
Иранские	Персидский, таджикский, дари, шугнанский	Персидский	4	K ₅

²⁵ В случае наложения одновременно нескольких фильтров они взаимодействуют по правилу конъюнкции логических условий (операция «логическое И»).

²⁶ Состав эталонной выборки практически совпадает с теми языками, по которым проводится выборочная экспертиза БД. Это позволит в дальнейшем оценить влияние ошибок и погрешностей в описании на результаты расчетов.

Германские	Немецкий, датский, норвежский, исландский, английский	Немецкий	5	K ₆
Романские	Итальянский, испанский, галисийский, португальский, румынский	Испанский	5	K ₇
Кавказские-1	Агульский, лезгинский, багвалинский,	Лезгинский	3	K ₈
Кавказские-2	Грузинский, абхазский	-	-	-
Палеоазиатские	Чукотский, корякский, ительменский, нивхский, бурушаски	-	-	-
Другие	Аккадский, армянский, ассамский, бенгальский, бирманский	-	-	-

Для оценки качества меры применяется критерий:

$$K = (K_1 + K_2 + K_3 + K_4 + K_5 + K_6 + K_7 + K_8) / 8, \quad (4.15)$$

где K_i – коэффициент совпадения по каждой из семей (уральские, тюркские, монгольские, славянские, иранские, германские, романские, кавказские)²⁷.

Значения K и K_i лежат в диапазоне от 0 до 1. Коэффициент K_i рассчитывается путем ранжирования языков по мере близости относительно языка-прототипа в семье. Он равен максимуму (т.е. единице), если все языки из одной семьи (ветви) оказываются упорядоченными без включений из других семей (ветвей), иначе происходит уменьшение пропорционально нарушению ранга.

$$K_i = N_p / N_g, \quad (4.16)$$

где N_p – количество языков из целевой группы, которые стоят за языком-прототипом подряд без включений «чужих» языков. N_g – число языков в группе (см. табл. 4.3).

²⁷ Коэффициенты качества K_i рассчитываются только по генетически однородным семьям и ветвям в эталонной выборке языков.

4.1.4. Результаты расчетов мер

Рис. 4.1. Экранная форма программы Similarity

Для выполнения расчетов парных мер близости на языке VBA была написана программа Similarity²⁸. В качестве среды хранения данных использовался лицензионный пакет MS Excel, куда данные были проконвертированы из БД. На рис. 4.1 показана экранная форма задания параметров расчетов.

Программа позволяет рассчитывать не только простые меры, представленные в таблице 4.2, но и различные их комбинации. Как можно видеть из экранной формы на рис. 4.1, программа позволяет задавать 10 параметров расчета, представленных в таблице 4.4.

Т а б л и ц а 4.4

Описание дискретных параметров расчета в программе Similarity

№ пп	Наименование параметра	Значение	Описание параметра
1	Вид меры близости	1) Простая аддитивная 2) Весовая иерархическая	Задаёт способ учета вклада каждого признака

²⁸ Автор программы В.Н.Поляков.

№ пп	Наименование параметра	Значение	Описание параметра
		3) Весовая частотная	
2	Тип признака 1	1) Все признакаи 2) Фактографические (терминальные) 3) Классифицирующие (нетерминальные)	Задаёт ограничение по роли совпадающих признаков в общей модели реферата
3	Тип признака 2 (3 значения)	1) Все общие признаки 2) TRUE-признаки 3) FALSE-признаки	Задаёт ограничение по виду совпадающих признаков с позиций их присутствия / отсутствия у сравниваемых языков
4	Фильтр по признакам «Число уровней» (2 значения)	1) Нет ограничений 2) Ограничить (прим.1)	Задаёт ограничения по уровню учитываемых признаков (сверху и снизу в иерархии модели). Вводится минимальное и максимальное значения номера уровня, которые принимаются при расчетах
5	Фильтр по признакам «Частота признака» (2 значения)	1) Нет ограничений 2) Ограничить (прим.1)	Задаёт ограничения по частоте учитываемых признаков (сверху и снизу в диапазоне частоты распределения по всей БД). Вводится минимальное и максимальное значения частоты. Максимум ограничен общим количеством языков в БД
6	Фильтр по признакам «Объем по модели» (2 значения)	1) Нет ограничений 2) Ограничить (прим.1)	Задаёт ограничения расчета мер по одному из разделов модели. Раздел реферата выбирается из списка
7	Фильтр по языкам «Объем по языкам» (2 значения)	1) Нет ограничений 2) Ограничить (прим.2)	Задаёт ограничения по расчету меры для языков, если количество присутствующих признаков в языке выходит за указанные пределы. Вводится минимальное и максимальное число признаков в языке. Максимум ограничен общим количеством признаков в БД
8	Весовая мера (2 значения)	1) Прямая 2) Обратная	Определяет способ учета влияния частоты. Если весовая мера прямая, то вес с ростом частоты (или уровня в иерархии) растет, иначе – падает
9	Вклад частоты (2 значения)	1) Линейный 2) Квадратичный	Определят вид функции, по которой рассчитывается весовой коэффициент частотной меры
10	Оценка меры	1) Да 2) Нет	Определяет, нужно ли рассчитывать значение качества меры близости автоматически

Примечания:

1. Параметр дает возможность дополнительного варьирования путем ввода численных значений.
2. Параметр не влияет на меру близости, так как всего лишь определяет, нужно ли считать меру для данного языка или нет.

Общее количество комбинированных мер, которые можно образовать варьированием восьми²⁹ дискретных параметров в программе Similarity составляет $M = 3*3*3*2*2*2*2*2 = 864$. Кроме того, возможность ввода дополнительных численных параметров, при специфицировании значений фильтров 4 и 5 (табл. 4.4) и выбор номера раздела делают задачу сплошного перебора всех возможных вариантов расчета меры близости чрезвычайно трудно выполнимой.

Результаты расчетов представляются в виде таблицы Excel, в которой в заголовках строк и столбцов расположены названия языков, а в ячейках на пересечении строк и столбцов значение парной меры близости. На рис. 4.2 показан фрагмент таблицы простой аддитивной меры, полученной по всему пространству признаков. Мера представлена без нормировки.

В ячейку таблицы A1 заносятся все исходные данные, определяющие вид меры близости и, если необходимо, комментарий.

Если был указан параметр «Оценка меры» = «Да», программа последовательно сортирует данные по столбцам относительно каждого языка-прототипа в группе и подсчитывает коэффициенты K_i по формуле (2) и итоговый коэффициент K по формуле (1). На рис. 4.3 показан пример данных о близости, полученных по всему пространству признаков, отсортированных по отношению к финскому языку. Как мы видим языки финно-угорской группы в этом варианте расчетов оказались ближе всех к языку-прототипу, что дает значение коэффициента $K_1=1$.

A1		Вид меры близости: Простая аддитивная. Тип признаков -1: Все признаки. Тип признаков-2 задано. Ограничение частоты сверху: Не задано. Раздел реферата: Не задан. Мин.число										
	А	В	ВЕНГЕРС	ФИНСКИЙ	АССАМСК	ДАРИ	ИТЕЛЬМЕ	ПОРТУГА	ГРУЗИНС	БУРУШАС	АККАДСК	НОРВЕЖС
1	Вид меры											
2	ВЕНГЕРС	3821	3597	3593	3585	3582	3538	3570	3537	3565	3522	
3	ФИНСКИЙ	3597	3821	3585	3571	3540	3552	3536	3525	3541	3548	
4	АССАМСК	3593	3585	3821	3613	3574	3588	3568	3567	3587	3540	
5	ДАРИ	3585	3571	3613	3821	3574	3598	3594	3583	3589	3562	
6	ИТЕЛЬМЕ	3582	3540	3574	3574	3821	3539	3595	3544	3566	3487	
7	ПОРТУГА	3538	3552	3588	3598	3539	3821	3549	3552	3550	3615	
8	ГРУЗИНС	3570	3536	3568	3594	3595	3549	3821	3572	3570	3511	
9	БУРУШАС	3537	3525	3567	3583	3544	3552	3572	3821	3557	3524	
10	АККАДСК	3565	3541	3587	3589	3566	3550	3570	3557	3821	3528	
11	НОРВЕЖС	3522	3548	3540	3562	3487	3615	3511	3524	3528	3821	

Рис. 4.2. Фрагмент таблицы простой аддитивной меры, полученной по всему пространству признаков

²⁹ Параметры 7 и 10 в таблице 4.4 не влияют на характеристики меры близости.

Расчет одного варианта меры по эталонной выборке из 48 языков на компьютере с процессором Intel Pentium 1,6 ГГц занимает около 20 мин. Расчет по одному разделу реферата около 5 мин. Полный расчет по всей базе (315 языков) выполняется свыше 10 часов.

За время тестирования БД был проведен большой объем работ по выбору наилучшего варианта меры близости и исследованию влияния различных факторов (типы признаков, их частота в БД, иерархия в структуре реферата, вклад различных разделов реферата) на качество меры.

		ФИНСКИЙ			ФИНСКИЙ
1	ФИНСКИЙ	3821	25	НОРВЕЖСКИЙ	3548
2	ВЕПССКИЙ	3637	26	ГАЛИСИЙСКИЙ	3548
3	ЭСТОНСКИЙ	3630	27	АНГЛИЙСКИЙ	3543
4	ВЕНГЕРСКИЙ	3597	28	АККАДСКИЙ	3541
5	ХАНТЫЙСКИЙ	3593	29	ТАДЖИКСКИЙ	3541
6	КАЛМЫЦКИЙ	3590	30	ИТЕЛЬМЕНСКИЙ	3540
7	АЗЕРБАЙДЖАНСКИЙ	3589	31	НЕМЕЦКИЙ	3539
8	АССАМСКИЙ	3585	32	ГРУЗИНСКИЙ	3536
9	БАШКИРСКИЙ	3585	33	ИСПАНСКИЙ	3531
10	ИСЛАНДСКИЙ	3578	34	ИТАЛЬЯНСКИЙ	3529
11	ТУРЕЦКИЙ	3577	35	БУРУШАСКИ	3525
12	БУРЯТСКИЙ	3576	36	АГУЛЬСКИЙ	3525
13	МОГОЛЬСКИЙ	3575	37	ПЕРСИДСКИЙ	3524
14	ДАРИ	3571	38	БОЛГАРСКИЙ	3514
15	ТАТАРСКИЙ	3571	39	ПОЛЬСКИЙ	3509
16	МОНГОРСКИЙ	3565	40	ШУГНАНСКИЙ	3502
17	АРМЯНСКИЙ	3563	41	НИВХСКИЙ	3501
18	БЕНГАЛЬСКИЙ	3563	42	БЕЛОРУССКИЙ	3500
19	ТУРКМЕНСКИЙ	3562	43	ЧУКОТСКИЙ	3492
20	БАГВАЛИНСКИЙ	3560	44	КОРЯКСКИЙ	3490
21	РУМЫНСКИЙ	3556	45	АБХАЗСКИЙ	3488
22	МАКЕДОНСКИЙ	3553	46	ДАТСКИЙ	3488
23	ПОРТУГАЛЬСКИЙ	3552	47	БИРМАНСКИЙ	3459
24	ЛЕЗГИНСКИЙ	3549	48	РУССКИЙ	3455

Рис. 4.3. Мера близости, ранжированная относительно финского языка

Фактически было установлено³⁰, что наилучших значений мера достигает при простом аддитивном сложении всех совпадающих признаков без ограничений на их частоту, иерархию или принадлежность к разделу (см. табл. 4.5). В этом случае по двум группам (уральские, тюркские) достигается полное совпадение с традиционным генетическим представлением и коэффициент качества К равен 0,67. Все остальные комбинации признаков давали худший результат.

³⁰ Данные приводятся по состоянию расчетов на 7.06.2006.

Качество мер близости – простые аддитивные меры

	Все признаки	Только присутствующие в языках	Только отсутствующие в языках	Только классифицирующие	Только фактографические	Только присутствующие классифицирующие	Только присутствующие фактографические
К ₁ -Уральские (5 яз.)	1	0,5	0,75	0,5	1	0,25	0,5
К ₂ -Тюркские (5 яз.)	1	0,75	0,5	0,75	0,75	0	0,75
К ₃ -Монгольские (4 яз.)	0,67	0,33	1	0,33	0,67	0,33	0,33
К ₄ -Славянские (5 яз.)	0,5	0,5	0	0,25	0,5	0,25	0,75
К ₅ -Иранские (4 яз.)	0,67	0,33	0	0,33	0,67	0	0,33
К ₆ -Германские (5 яз.)	0,5	0,75	0,25	0,75	0,5	0,75	0,5
К ₇ -Романские (5 яз.)	0,5	0,5	0,5	0,5	0,5	0,25	0,75
К ₈ -Кавказские (3 яз.)	0,5	0	0,5	0,5	0,5	0	0
К-Итого	0,67	0,46	0,4375	0,49	0,64	0,23	0,49

В таблице 4.6 представлены результаты расчетов мер по разделам реферата. Здесь мы наблюдаем такой же эффект. Сужение объема расчетов не приводит к улучшению качества меры. Варьирование ограничений по частоте и иерархии, выбор различных весовых коэффициентов и конструирование различных комбинаций из перечисленных факторов также не позволили улучшить результат, полученный первоначально путем использования простой аддитивной меры. Это свидетельствует о хорошей системной проработке формата представления данных, заложенного на этапе проектирования БД.

Как видно из таблицы 4.6, наилучшим с точки зрения маркирования генетической близости языков оказывается раздел реферата 2.3.7. Части речи ($K=0,56$ для меры, учитывающей только присутствующие признаки). Однако по сумме всех значений показателя качества K по всем разделам мера «Все признаки» остается лидирующей (см. нижнюю строку таблицы 4.6). В таблице 4.6 курсивом выделены значения, максимальные для столбцов, подчеркиванием – максимальные для строк. По количеству максимумов наилучшим среди разделов остается раздел «2.3.7. Части речи» – 3 максимума. Лидером среди мер близости по количеству максимумов является мера, учитывающая только присутствующие признаки.

Результаты расчета К по разделам реферата

Раздел реферата	Все признаки	Только присутствующие в языках	Только классифицируемые	Только фактографические	Только присутствующие классифицирующие	Только присутствующие фактографические
2.1.1. ФОНЕМНЫЙ СОСТАВ	0,30	0,35	0,32	0,30	0,23	0,36
2.1.2. ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ	0,29	0,34	0,24	0,26	0,19	0,33
2.1.3. ФОНЕТИЧЕСКИ ОБУСЛ. ПРОЦЕССЫ	0,09	0,10	0,17	0,06	0,06	0,14
2.1.4. СЛОГ	0,20	0,16	0,50	0,09	0,29	0,14
2.2.1. ФОНОЛОГИЧЕСКАЯ СТРУКТУРА	0,13	0,22	0,13	0,16	0,19	0,16
2.2.2. ФОНОЛОГИЧЕСКИЕ ПРОТИВОП. МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ	0,16	0,19	0,19	0,13	0,13	0,16
2.2.3. МОРФОЛОГИЧЕСКИ ОБУСЛОВЛЕННЫЕ ЧЕРЕДОВАНИЯ	0,33	0,17	0,18	0,17	0,09	0,10
2.3.0. МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА	0,34	0,46	0,33	0,26	0,21	0,23
2.3.1. КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ	0,07	0,07	0,07	0,07	0,07	0,07
2.3.2. ИМЕННЫЕ КЛАССИФИКАЦИИ	0,23	0,09	0,15	0,23	0,03	0,20
2.3.3. ЧИСЛО	0,16	0,13	0,21	0,20	0,17	0,20
2.3.4. ПАДЕЖНЫЕ ЗНАЧЕНИЯ	0,46	0,44	0,30	0,53	0,26	0,51
2.3.5. ГЛАГОЛЬНЫЕ КАТЕГОРИИ	0,32	0,29	0,20	0,29	0,14	0,20
2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ	0,47	0,29	0,41	0,39	0,36	0,29
2.3.7. ЧАСТИ РЕЧИ	0,44	0,56	0,20	0,53	0,07	0,53
2.4.0. ПАРАДИГМЫ	0,42	0,40	0,30	0,32	0,22	0,27
2.5.1. СТРУКТУРА СЛОВОФОРМЫ	0,17	0,30	0,18	0,13	0,17	0,27
2.5.2. СЛОВООБРАЗОВАНИЕ	0,18	0,20	0,17	0,18	0,10	0,17
2.5.3. ПРОСТОЕ ПРЕДЛОЖЕНИЕ	0,41	0,34	0,36	0,33	0,06	0,34
2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ	0,10	0,17	0,18	0,10	0,09	0,20
Сумма по столбцу:	5,26	5,18	4,77	4,73	3,14	4,86

Интересно отметить, в каких разделах достигается максимум показателя качества меры близости по той или иной языковой семье (ветви) (см. табл. 4.7).

Таблица 4.7

Максимумы показателя качества по различным языковым семьям

Языки	Мера	Раздел	Ki
K ₁ -Уральские (5 яз.)	Только присутствующие фактографические	2.3.4.	1
K ₂ -Тюркские (5 яз.)	Только присутствующие	2.3.7.	1
K ₃ -Монгольские (4 яз.)	Все признаки	2.5.3.	1
K ₄ -Славянские (5 яз.)	Только фактографические	2.5.3.	0,75

Языки	Мера	Раздел	Ki
	Только присутствующие фактографические	2.1.1., 2.5.3.	0,75
	Только присутствующие	2.5.3.	0,75
K ₅ -Иранские (4 яз.)	Только фактографические	2.3.6., 2.3.7.	0,67
	Только присутствующие фактографические	2.3.7.	0,67
	Только присутствующие	2.1.2., 2.3.7.	0,67
	Только классифицирующие	2.5.4.	0,67
K ₆ -Германские (5 яз.)	Только присутствующие	2.3.6.	0,75
	Только классифицирующие	2.1.1.	0,75
	Только присутствующие классифицирующие	2.2.1.	0,75
K ₇ -Романские (5 яз.)	Только присутствующие фактографические	2.3.7.	0,75
	Только присутствующие	2.3.7.	0,75
	Только классифицирующие	2.1.4., 2.3.7.	0,75
	Только присутствующие классифицирующие	2.3.6.	0,75
K ₈ -Кавказские (3 яз.)	Все признаки	2.2.3, 2.3.4.	1
	Только фактографические	2.3.4.	1
	Только присутствующие	2.3.0.	1
	Только классифицирующие	2.1.4.	1

4.1.5. Предварительные обсуждения результатов

Результаты, представленные в таблице 4.5, можно рассматривать как вполне обнадеживающие. Однако, анализ расчетов позволил выявить проблемы, на решение которых предполагается направить дальнейшие усилия исследовательской группы. Наиболее серьезными оказались две проблемы:

- недостаточно высокий уровень качества меры близости;
- слишком близкое расстояние между неродственными языками в ранге.

Чрезвычайно важным представляется вопрос: «Существуют ли пути улучшения мер близости?» Ответ на этот вопрос предстоит найти в ближайшие годы. На наш взгляд, существенного улучшения качества мер можно добиться скоординированными действиями сразу по нескольким направлениям.

Во-первых, это выполнение полной экспертизы БД и устранение ошибок и неточностей, внесенных как в результате кодирования информации, так и обнаруженных в первоисточнике – энциклопедии «Языки мира».

Во-вторых, существует потенциал для улучшения качества мер путем использования различных эвристик. К таким перспективным идеям можно отнести: использование признаков – генетических маркеров, подбор весов при различных частях БД, формирование сложных мер близости, основанных на различных способах свертки информации.

4.2. Кластеризация³¹

В этом разделе главы мы уделим внимание описанию следующих аспектов квантитативных исследований на материале БД «Языки мира»:

- задача кластеризации языков компьютерными методами;
- задача автоматического выявления ошибок в БД;
- расчет частоты встречаемости признаков в выбранной группе языков;
- задачи интеллектуального анализа данных с использованием специальных пакетов прикладных программ.

Все эти задачи находятся еще в стадии первоначальной проработки. Методы кластеризации и интеллектуального анализа данных очень сильно зависят от качества предварительно рассчитанных мер близости. Именно поэтому основные результаты исследований в этой сфере еще предстоит получить и осмыслить после окончательной экспертизы БД. Поэтому в этой части нашей книги основное внимание будет уделяться общим описаниям методов и постановке задач.

4.2.1. Общие сведения о задаче кластеризации

Кластерный анализ – множество вычислительных процедур, которые формируют либо выявляют иерархии (разбиения), лежащие в основе тех или иных совокупностей данных. Алгоритм кластеризации действует в два этапа: на первом этапе вычисляется метрика (показатели сходства объектов классифицируемой совокупности), на втором этапе происходит пошаговое построение классов. Выбор способов расчета метрик обусловлен способом описания входных данных.

Обозначим множество языков в базе данных как I . Согласно (Жамбю 1988), кластеризация³² на множестве I – разбиение его на определенное количество непустых попарно-непересекающихся подмножеств. Автоматическая кластеризация³³ – применение вычислительных методов для построения систем классов (кластеров), которые могут быть либо иерархическими, либо непересекающимися. На первом этапе в нашей исследовательской группе для кластеризации применялся наиболее простой метод кластеризации ISODATA (Поляков, Соловьев 2006-1), который относится к классу неиерархических. Можно сказать, что этот этап исследований показал принципиальную возможность применения методов автоматической кластеризации в данной сфере квантитативных исследований. Дальнейшие исследования показали, что иерархические алгоритмы дают лучшие результаты на заданной совокупности языков, чем неиерархические. Это объясняется, с одной стороны, более «тонким» механизмом работы иерархических методов по сравнению с неиерархическими, и, с другой, самой природой исследуемого объекта, в которой генеалогическое дерево является базовой иерархической структурой. В данном разделе описаны два алгоритма иерархической классификации: метод «ближайшего соседа» (одной связи) и метод «дальнего соседа» (полной связи).

³¹ В разделе использованы материалы диплома студентки КГУ М.В. Двоеносовой (руководитель В.Д. Соловьев).

³² Иногда в этом случае используется термин *классификация*, однако мы будем считать классификацией задачу отнесения элемента к известному множеству, а кластеризацией – задачу разбиения множества I на неизвестные априори подмножества (кластеры).

³³ Методам автоматической кластеризации могут быть противопоставлены человеко-машинные процедуры по-уавтоматической кластеризации, однако применимость такого подхода для решения поставленной задачи еще требует изучения.

Введем необходимые обозначения. Пусть: I – множество языков, J – множество признаков, K_{IJ} – таблица бинарных данных, определяющих свойства языков. На пересечении строки i (=языка i) и столбца j (=признака j) этой таблицы находится значение признака $k(i, j)$. Таблица логического описания (K_{IJ}) – таблица, элементы которой равны 0 или 1 в булевском смысле. То есть $k(i, j) = 1$ означает, что элемент i обладает свойством j . $k(i, j) = 0$ означает, что свойство j для объекта i не выполнено.

$$K_{IJ} = \begin{pmatrix} k_{11} & \dots & k_{1n} \\ \dots & \dots & \dots \\ k_{m1} & \dots & k_{mn} \end{pmatrix}, \quad (4.17)$$

$i = 1 \dots m$, m – число языков в БД; $j = 1 \dots n$, n – число признаков в БД; $k_{ij} \in \{0, 1\}$.

Матрица расстояний D представляет собой таблицу, в которой и столбцам, и строкам поставлены в соответствие элементы множества языков I . На пересечении строки и столбца находится значение меры близости $d(i_k, i_e)$ между языком i_k и языком i_e , рассчитанной в соответствии с выбранной исследователем метрикой, часть которых была описана в раздел 3.1.

$$D = \begin{pmatrix} d(i_1, i_1) & \dots & d(i_1, i_m) \\ \dots & \dots & \dots \\ d(i_m, i_1) & \dots & d(i_m, i_m) \end{pmatrix}, \quad (4.18)$$

D – симметричная матрица, для элементов которой справедливы соотношения $d(k, e) \in [0, 1]$, $d(k, k) = 1$, $d(k, e) = d(e, k)$. m – число языков в БД.

Входными данными для алгоритмов иерархической классификации может быть матрица расстояний между элементами одного из множеств I или J . В нашем случае используется матрица расстояний (мер близости), построенная на множестве языков I . Матрица расстояний D получается в результате расчета мер близости языков с помощью программы Similarity, описанной в разделе 3.1.

4.2.2. Кластеризация языков

1. Общий алгоритм

На вход алгоритма, реализованного в данной работе, подается предварительно вычисленная по формулам, описанным в предыдущем разделе, матрица расстояний между классифицируемыми объектами. Матрица расстояний D имеет вид:

$$\begin{array}{cccccc} & i_1 & i_2 & \dots & \dots & \dots & i_m \\ \hline i_1 & 0 & d(i_1, i_2) & \dots & \dots & \dots & d(i_1, i_m) \\ i_2 & d(i_1, i_2) & 0 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ i_m & d(i_1, i_m) & \dots & \dots & \dots & \dots & 0 \end{array} \quad (4.19)$$

Ее элементами являются расстояния (значения мер близости) между объектами классифицируемой совокупности.

Общий алгоритм кластеризации выглядит следующим образом:

входная информация – таблица расстояний между элементами множества I (объектов) или J (дескрипторов) по желанию исследователя.

1. Шаг 1: Пусть есть таблица расстояний на множестве I (причем элементы I и соответствующие ему терминальные классы пронумерованы от 1 до $|I|$). Выполняется операция объединения двух ближайших элементов по метрике $d_{II}(i, i')$, что дает:

- a. Уровень $m' = |I| + 1$ – первый сформированный узел.
- b. $A(m') = i$ – старший преемник для узла $m' = |I| + 1$
- c. $B(m') = i'$ – младший преемник для узла $m' = |I| + 1$
- d. $P(m') = 2$ – число элементов для узла m' .
- e. $v(m') = \min(d_{II}(i, i'))$ – индекс узла m' .

Объединив элементы i и i' , мы можем получить вместо I новое разбиение P . Для этого исключим классы $\{i\}$ и $\{i'\}$ из I и добавим класс $\{i, i'\}$.

Для повторного применения этого процесса нужно вычислить метрику $\{d_{PP}(q, q') : q, q' \in P\}$ на классах разбиения P , по возможности используя только данные предыдущего шага $d_{II}(i, i')$. При этом вычисляются расстояния между вновь созданным классом $n' = \{i, i'\}$ и другими одноэлементными классами.

2. Общий шаг: Отыскиваем два класса q и q' , расстояние между которыми $\{d_{PP}(q, q')\}$ минимально: q и q' ближайшие друг к другу классы разбиения P . Таким образом, создаем новое разбиение, которое характеризуется узлом уровня m'' , на 1 больше предыдущего $m'' = m' + 1$. При этом определяем:

- a. $A(m'') = q$ – старший преемник узла m'' .
- b. $B(m'') = q'$ – старший преемник узла m'' .
- c. $E(m'') = E(q) + E(q')$ – число элементов узла m'' .
- d. $v(m'') = \min(d_{PP}(i, i'))$

3. Конец процесса: получаем разбиение из двух классов, дающих в объединении I .

Различие алгоритмов иерархической классификации заключается в способе вычисления расстояния между классами и записи рекуррентных формул для вычисления этого расстояния. Рассмотрим два способа расчета расстояния между классами: метод «ближайшего соседа» и метод «дальнего соседа».

1. Расстояние между классами и рекуррентные формулы для алгоритма одной связи («ближайшего соседа»).

Пусть $d(i, i')$ – исходное расстояние между элементами множества I (классифицируемое множество объектов). A и B – два подмножества множества I . Обозначим $\delta_{\min}(A, B) = \min\{d(i, i') : i \in A, i' \in B\}$ – минимальное из исходных расстояний между отдельными элементами разных классов (рис. 4.4).

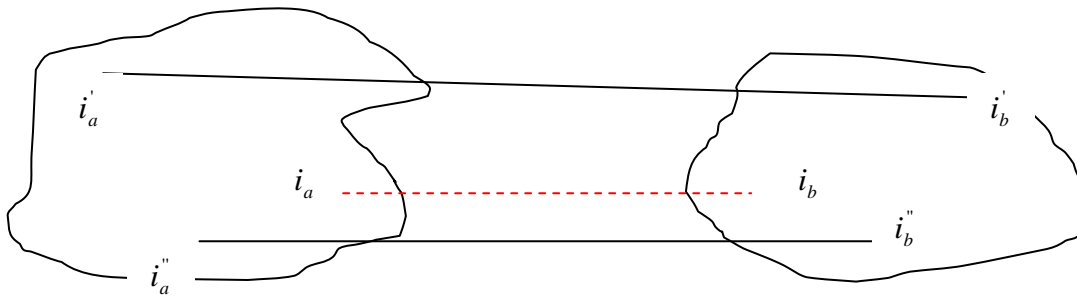


Рис. 4.4. Схематичное представление минимального исходного расстояния для метода «ближайшего соседа»

Рекуррентная формула примет вид:

$$\delta_{uim}^h(t, S_h \cup S_{h'}) = \min[\delta_{uim}^{h-1}(t, S_h), \delta_{uim}^{h-1}(t, S_{h'})], \quad (4.20)$$

где t – новый класс, S_h и $S_{h'}$ – классы-преемники.

2. Расстояние между классами и рекуррентные формулы для алгоритма полной связи («дальнего соседа»).

Пусть $d(i, i')$ – исходное расстояние между элементами множества I (классифицируемое множество объектов). A и B – два подмножества множества I . Обозначим $\delta_{uim}(A, B) = \max\{d(i, i') : i \in A, i' \in B\}$ – максимальное из исходных расстояний между отдельными элементами разных классов (рис. 4.5). Рекуррентная формула примет вид:

$$\delta_{uim}^h(t, S_h \cup S_{h'}) = \max[\delta_{uim}^{h-1}(t, S_h), \delta_{uim}^{h-1}(t, S_{h'})], \quad (4.21)$$

где t – новый класс, S_h и $S_{h'}$ – классы-преемники.

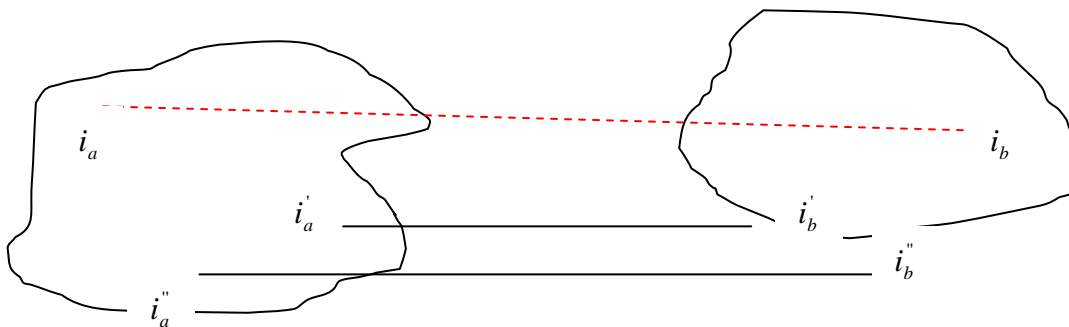


Рис. 4.5. Схематичное представление максимального исходного расстояния (метод «дальнего соседа»)

4.2.3. Кластеризация признаков

Вычисление мер близости и кластеризация признаков, производятся по тем же алгоритмам, что и для языков. Для применения этих алгоритмов необходимо транспонировать таблицу исходных данных (таблицу логического описания).

4.2.4. Проверка результатов методом Data Splitting

Выбор метода проверки зависит от алгоритма кластеризации. Реализуемый в программе алгоритм требует на вход матрицу расстояний между классифицируемыми элементами. Точно также должен действовать и соответствующий алгоритм проверки. Метод Data Splitting был выбран как наиболее подходящий.

Метод проверки Data Splitting (разделение данных) заключается в следующем: если в конечном решении найдутся кластеры, содержащие отдельные элементы выборочной совокупности, то весь набор данных может быть случайным образом разбит на две части, каждая из которых должна быть проанализирована отдельно. Получаемый таким образом результат должен по своей структуре напоминать решение, полученное на предыдущем этапе.

4.2.5. Архитектура программного комплекса

Результатом проделанной работы стал разработанный «Программный комплекс для проведения статистических исследований в области компаративистики», обладающий необходимыми возможностями для проведения кластеризации и статистических расчетов, связанных с ней (автор программы – М.В.Двоеносова).

В настоящее время комплекс проходит тестовую эксплуатацию в Российском научно-образовательном центре по лингвистике им. И.А. Бодуэна де Куртенэ. Эксплуатация программного комплекса в процессе разработки позволила получить предварительные сведения о закономерностях объектов базы данных «Языки мира», провести статистические расчеты, получить некоторые результаты.

После запуска программы происходит аутентификация пользователя и загрузка данных. Затем исследователь может приступить к работе. При нажатии кнопки «Расчеты», пользователь переходит к одноименной форме, предназначенной для проведения предварительных перед процедурой кластеризации статистических расчетов (рис. 4.6).

В окне «Языки» представлены языки базы данных, распределенные по языковым семьям. В окне «Признаки» представлены признаки в соответствии с иерархией (рис. 4.7).

Нажатием кнопок «Все» пользователь может выбрать для кластеризации все языки или признаки, а кнопку «Очистить» – снять метки соответственно.

После выбора объектов исследования, пользователь обращается к меню. На вкладке меню «Меры близости» содержатся кнопки вызова процедур расчета. Выбор процедуры расчета производится нажатием соответствующей кнопки, после чего результаты расчета меры близости отображаются в окне «Исходные данные» формы «Классификатор». Меню языковые семьи содержит подменю «Портрет языковой семьи и мера близости» и «Мера близости языковых семей». В первом случае рассчитывается портрет языковой семьи и расстояние между семейным портретом и языками, представляющими семью. Во втором – расстояние между языковыми семьями. Также предусмотрены функции: «Расчет расстояния между признаками», «Автоматическое выявление возможных ошибок», «Сохранение и загрузка меры близости в специальном формате CLW», «Корректировка исходных данных».

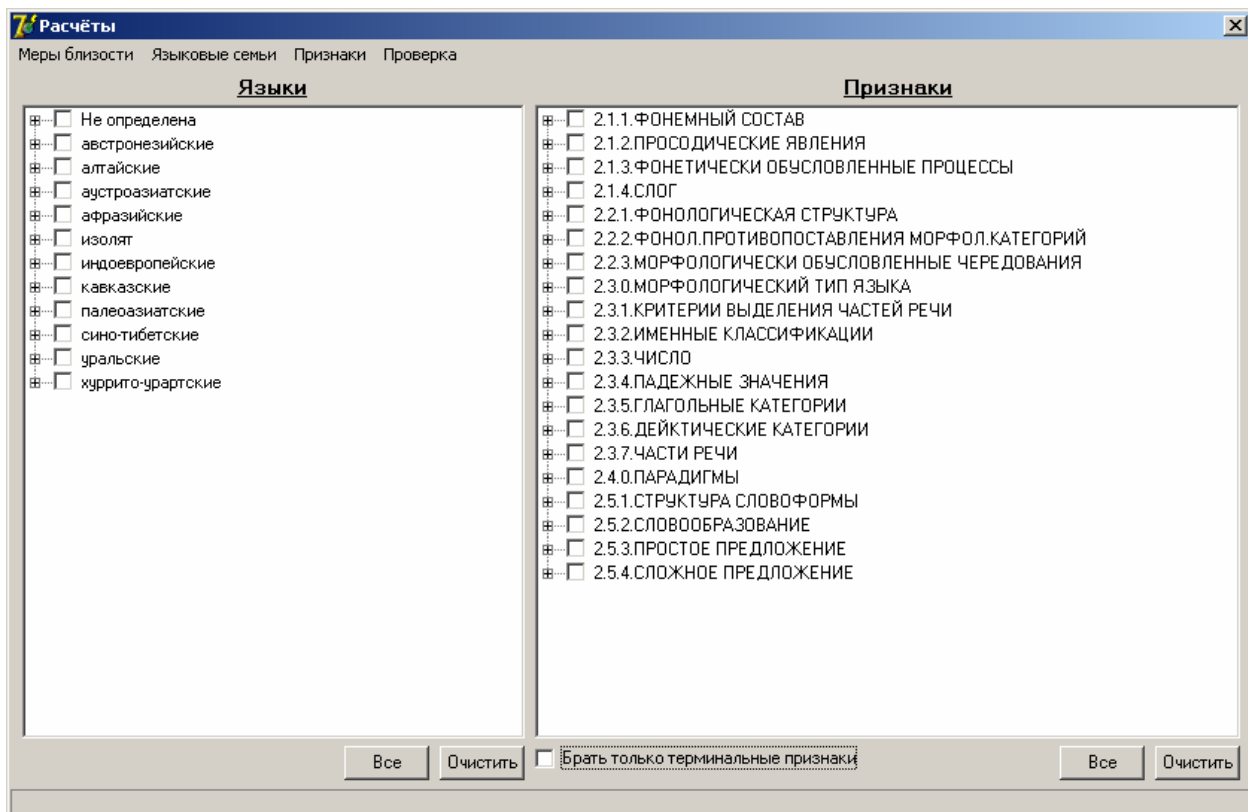


Рис. 4.6. Форма «Расчеты» программного комплекса

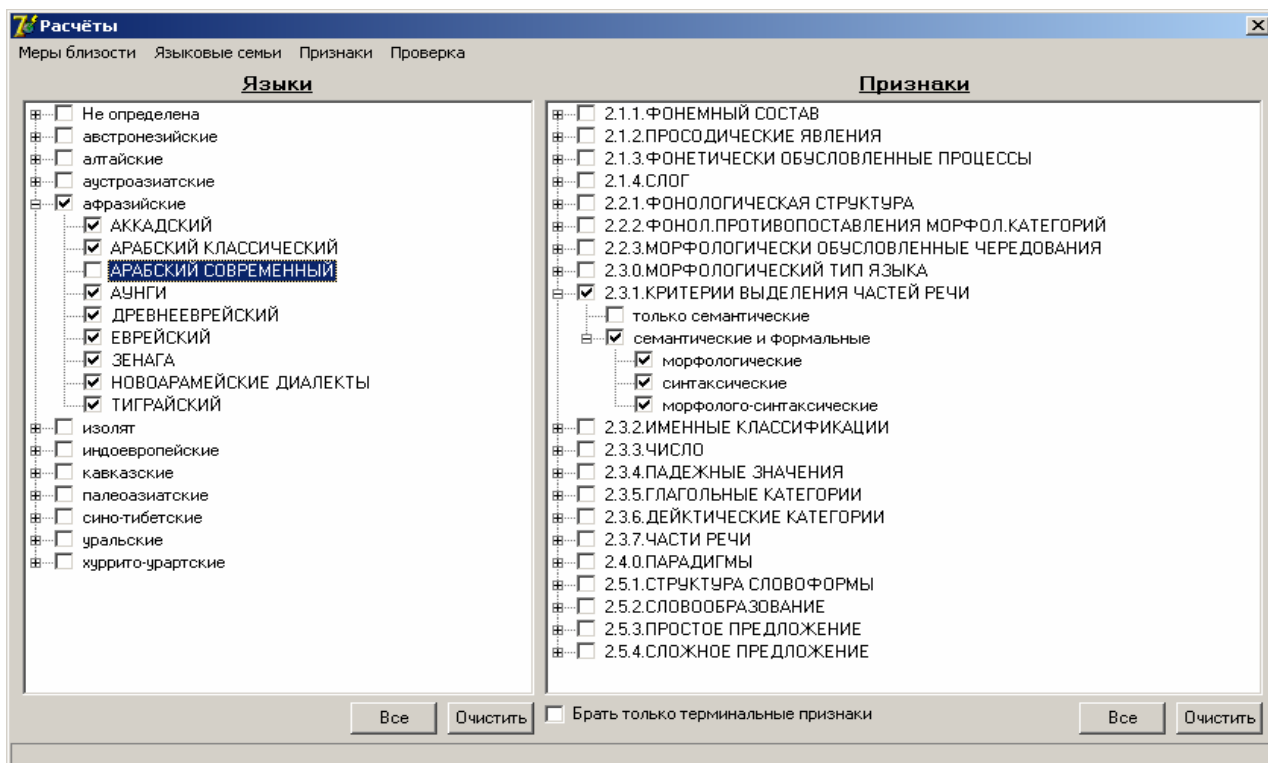


Рис. 4.7. Графическое отображение объектов базы

После вычисления мер близости (или после загрузки ее из внешнего файла), пользователь переходит на форму «Классификатор». Кнопка «Мах» предназначена

для вызова процедуры кластеризации по методу «Дальнего соседа». Кнопка «Min» предназначена для кластеризации по методу «Ближайшего соседа». Для того, чтобы проклассифицировать языки только по терминальным признакам базы данных, необходимо на форме «Расчеты» поставить галочку «Брать только терминальные признаки». Если на форме «Классификатор» поставить галочку «Проверка», то вместе с результатом классификации будет выведен результат проверки классификации методом «Data Splitting» (рис. 4.8).

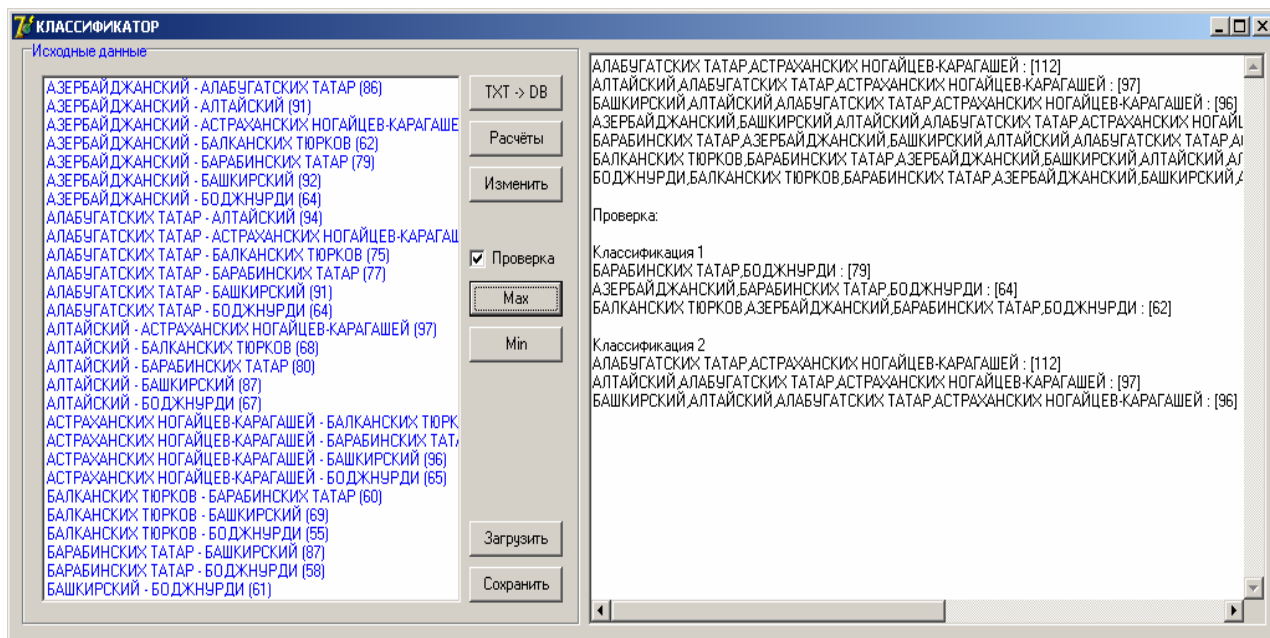


Рис. 4.8. Работа с программным комплексом в режиме «Классификатор»

Результаты всех расчетов выводятся в текстовые файлы с разделителями табуляции (L_Family (портрет), L_Dist (расстояние между языками), L_PropDist (расстояние между признаками), L_Errors (возможные ошибки в данных)). Указанные файлы создаются по умолчанию в одной папке с исполняемым файлом программы.

4.2.6. Выявление возможных ошибок в данных

В системе исследователю предоставлены средства для выявления возможных ошибок в данных. В данной программе используется следующий критерий: если по какому-либо признаку в группе из n языков $n-1$ язык принимает одно значение, а оставшийся язык – другое, то это может быть ошибка. Таким образом, исследователь может заранее исключить из расчетов возможные ошибочные сведения. Как показали результаты экспериментов, применение такого метода оправдано на этапе, когда данные в БД еще не прошли полную экспертизу. В этом случае метод позволяет выявлять до 70 % ошибок. Однако, эти ошибки распределяются примерно поровну: первая половина приходится на язык-носитель уникалий, вторая половина – на группу родственных языков. То есть, для получения достоверных данных, результаты, выявленные программой, необходимо уже проверять вручную. По мере выполнения полной экспертизы необходимость в данной опции отпадет.

4.2.7. Расчет частот встречаемости признаков в выбранной группе языков

Частоты встречаемости признаков рассчитываются следующим образом:

$$\vartheta(m) = \frac{\sum_i k(i, j)}{|I|}, \quad i = \overline{1, m}, \quad (4.22)$$

где i – конкретный язык, j – конкретный признак, $k(i, j)$ – значение из исходной таблицы логического описания для данных языка и признака, $|I| = m$ – количество языков в выбранной для исследования группе.

Анализ полученных результатов дает исследователю возможность заранее отобрать для проведения расчетов признаки, информация по которым наиболее полно представлена в базе данных.

4.3. Интеллектуальный анализ данных

Интересно, что использование двоичных признаков для описания языков с точки зрения возникающих при этом задач имеет сходство с исследованием видового состава растительных сообществ с использованием флористических списков. В этом случае аналогом площадки описания является язык, а аналогом списка присутствующих видов – список имеющихся у него признаков. Для каждого языка доля присутствующих признаков составляет лишь небольшую часть от всей совокупности признаков, также как и доля видов растений, встретившихся на данной площадке, составляет лишь небольшую часть от всех возможных видов. Подобные методы активно применялись при анализе растительных сообществ (Савельев 2004).

Существует аналогия между языками и описаниями растительных сообществ, а так же общие проблему с вычислением мер сходства: относительно небольшое количество признаков (со значением ‘true’) у каждого языка, разная степень информативности признаков, которая может быть обусловлена как их слишком большой распространенностью, так и уникальностью, разная степень подробности описания для различных языков и т. д., которые создают дополнительные трудности при их сопоставлении. Кроме того, априорно неизвестно какую информацию несет каждый конкретный совпадающий признак в паре языков: информацию о непосредственном генетическом родстве этой пары, информацию о дальнем родстве (имеют общего родственника) или информацию о контактах между языками и заимствовании. Для преодоления этих трудностей были предложены различные меры сходства, в том числе ориентированные на использование именно двоичных признаков. Кроме того, для выявления и представления структуры объектов в пространстве признаков здесь разработаны и используются различные методы ординации, т.е. представления объектов в пространстве низкой размерности, обычно на плоскости.

Наиболее известным способом ординации, основанным на использовании матрицы расстояний, является неметрическое многомерное шкалирование, отображающее объекты в пространство низкой размерности таким образом, что в нем Евклидовы расстояния между образами объектов пропорциональны расстояниям в исходной матрице. Это позволяет представить на плоскость «взаимоположение» объектов. Такие методы ординации основаны на вычислительных процедурах линейной алгебры и известны достаточно давно (Kruskal 1964). Кроме использования расстояний, существуют и другие методы ординации, объединяемые общим названием – анализ соответ-

ствия – и позволяющие анализировать таблицы признаков с большим количеством нулей. Кроме представления структуры самих объектов, такой анализ позволяет выявлять структуру признаков и связь признаков с объектами (Legendre 1998).

Расчеты проводились на эталонной группе из 48 языков, представленной ранее. По техническим причинам нумерация, используемая на последующих рисунках, начата с 2 (см. табл. 4.8).

Т а б л и ц а 4.8

Нумерация языков в эталонной группе

№	Язык	№	Язык
2	ВЕНГЕРСКИЙ	26	МОНГОРСКИЙ
3	ФИНСКИЙ	27	ЭСТОНСКИЙ
4	АССАМСКИЙ	28	МАКЕДОНСКИЙ
5	ДАРИ	29	НЕМЕЦКИЙ
6	ИТЕЛЬМЕНСКИЙ	30	БЕНГАЛЬСКИЙ
7	ПОРТУГАЛЬСКИЙ	31	РУМЫНСКИЙ
8	ГРУЗИНСКИЙ	32	ЛЕЗГИНСКИЙ
9	БУРУШАСКИ	33	РУССКИЙ
10	АККАДСКИЙ	34	КОРЯКСКИЙ
11	НОРВЕЖСКИЙ	35	ПЕРСИДСКИЙ
12	АНГЛИЙСКИЙ	36	ТАДЖИКСКИЙ
13	ИСЛАНДСКИЙ	37	ЧУКОТСКИЙ
14	БУРЯТСКИЙ	38	ТУРКМЕНСКИЙ
15	АЗЕРБАЙДЖАНСКИЙ	39	ТАТАРСКИЙ
16	ВЕПССКИЙ	40	ИСПАНСКИЙ
17	ХАНТЫЙСКИЙ	41	ИТАЛЬЯНСКИЙ
18	ТУРЕЦКИЙ	42	ГАЛИСИЙСКИЙ
19	БИРМАНСКИЙ	43	АБХАЗСКИЙ
20	АРМЯНСКИЙ	44	БЕЛОРУССКИЙ
21	БАГВАЛИНСКИЙ	45	БОЛГАРСКИЙ
22	АГУЛЬСКИЙ	46	ДАТСКИЙ
23	МОГОЛЬСКИЙ	47	НИВХСКИЙ
24	КАЛМЫЦКИЙ	48	ШУГНАНСКИЙ
25	БАШКИРСКИЙ	49	ПОЛЬСКИЙ

Для них были рассчитаны расстояния, выполнена ординация и рассчитаны ошибки ординации. Все вычисления осуществлялись с помощью профессионального статистического пакета R (R 2006), который является, de facto, международным стандартом для проведения статистических исследований³⁴.

Для оценки качества полученной ординации, т.е. соответствия расстояний в матрице и Евклидовых расстояний на ординационной плоскости, используются как методы визуализации, так и численные характеристики.

³⁴ Расчеты выполнялись А.А. Савельевым.

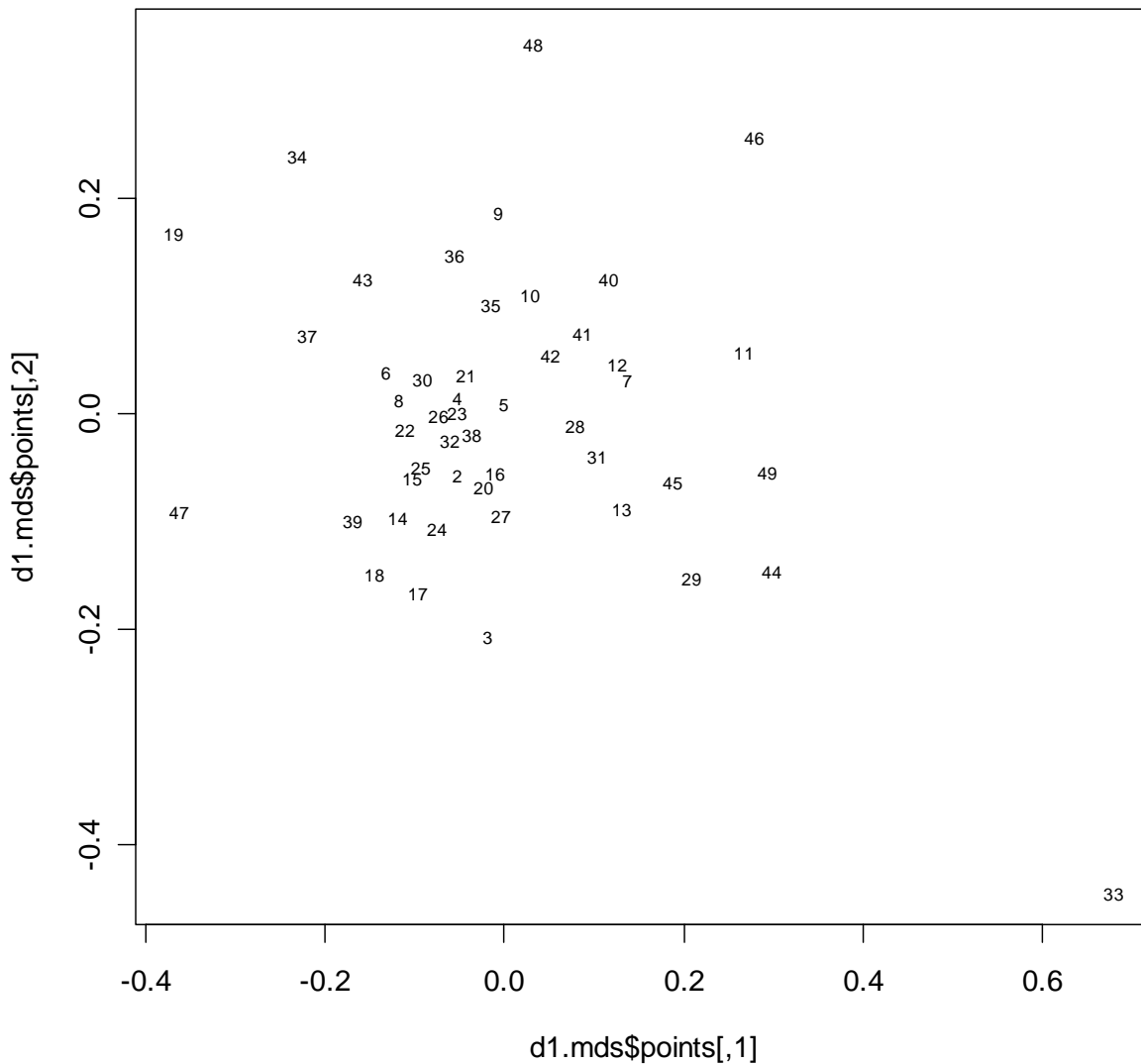


Рис. 4.9. Пример ординации

Наиболее распространенной численной характеристикой является STRESS, вычисляемая по формуле:

$$\sum_{ij} (d_{ij} - \hat{d}_{ij})^2 / \sum_{ij} \hat{d}_{ij}^2 \quad (4.23)$$

где суммирование ведется по всем парам ij объектов, d_{ij} есть исходная мера различия (расстояние) между i -м и j -м объектами (языками), а \hat{d}_{ij} – Евклидово расстояние между i -м и j -м объектами на ординационной плоскости. Результаты вычисления STRESS для различных методов вычисления расстояния показали, что ни один из методов вычисления расстояний не дает существенно лучших результатов. В нижеследующей таблице для каждого метода слева указывается его STRESS. Рассматрива-

лись различные метрики, включенные в систему R (начиная с manhattan) и предложенные авторами статьи.

Т а б л и ц а 4.9

Числовая оценка качества ординации

18.0 Additiv-all-all-all	14.3 bray	14.3 bray
20.3 Additiv-all-True-all	18.2 kulczynski	18.2 kulczynski
19.4 Additiv-Fact-all-all	19.8 gower	19.8 gower
22.6 Additiv-Fact-True-all	14.3 horn	36.2 morisita
19.7 Additiv-Klass-all-All	20.0 mountford	14.3 horn
24.2 Additiv-Klass-False-All	14.3 jaccard	20.0 mountford
18.5 Additiv-Klass-True-All	20.1 binomial	14.3 jaccard
20.3 manhattan	20.3 manhattan	36.2 raup
19.8 euclidean	19.8 euclidean	20.1 binomial
14.3 canberra	14.3 canberra	

Для визуальной оценки качества ординации используется график Шепарда, в котором по горизонтали откладывается расстояние в исходной матрице, а по вертикали – монотонная регрессия на него Евклидовых расстояний на ординационной плоскости. Если все точки лежат на ломаной, представляющей монотонную (неубывающую) регрессию, то ординация сохраняет отношение «не больше» для расстояний: если $d_{ij} \leq d_{kl}$, то и $\hat{d}_{ij} \leq \hat{d}_{kl}$.

Для содержательной оценки полученных ординаций предложен следующий подход. Языки, относящиеся к одной близкородственной группе языков (в данной выборке это – славянские, германские, романские, иранские, тюркские, монгольские, уральские, северокавказские, палеоазиатские), должны располагаться на ординационной плоскости компактно. По этому критерию наилучшие результаты дает мера близости Add-Fact-all-all (только фактографические). На рисунке 4.9 как раз и представлена ординация по этой мере.

Можно обратить внимание на следующие явно выделяющиеся данные на этом рисунке. На периферии оказались следующие языки: №19 – бирманский язык – принадлежит далеко отстоящему семейству языков (сино-тибетских), №47 – нивхский язык – изолят, его происхождение не известно, №46 – датский язык, как оказалось, его описание содержит ошибки (это, в частности, указывает на одно из возможных применений данного метода – поиск ошибок), №33 – русский язык оказался описан слишком подробно – «переописан».

Другим методом, совмещающим кластеризацию и ординацию, является использование непараметрических методов, относящихся к так называемым топографическим отображениям, и основанным на самоорганизации. Сюда относятся нейронные сети Кохонена (саморганизующиеся карты свойств), совмещающие классификацию методом к-средних с ординацией на плоскости (Kohonen 1997), и генеративные топографические отображения, совмещающие нечеткую классификацию на основе Гауссовых смесей распределений с отображением классов в подпространство низкой размерности (например, на плоскость) (Bishop 1998).

В результате применения метода нейронных сетей Кохонена с использованием всех признаков и Евклидова расстояния, был получен результат самоорганизации, представленный на рисунке 4.11.

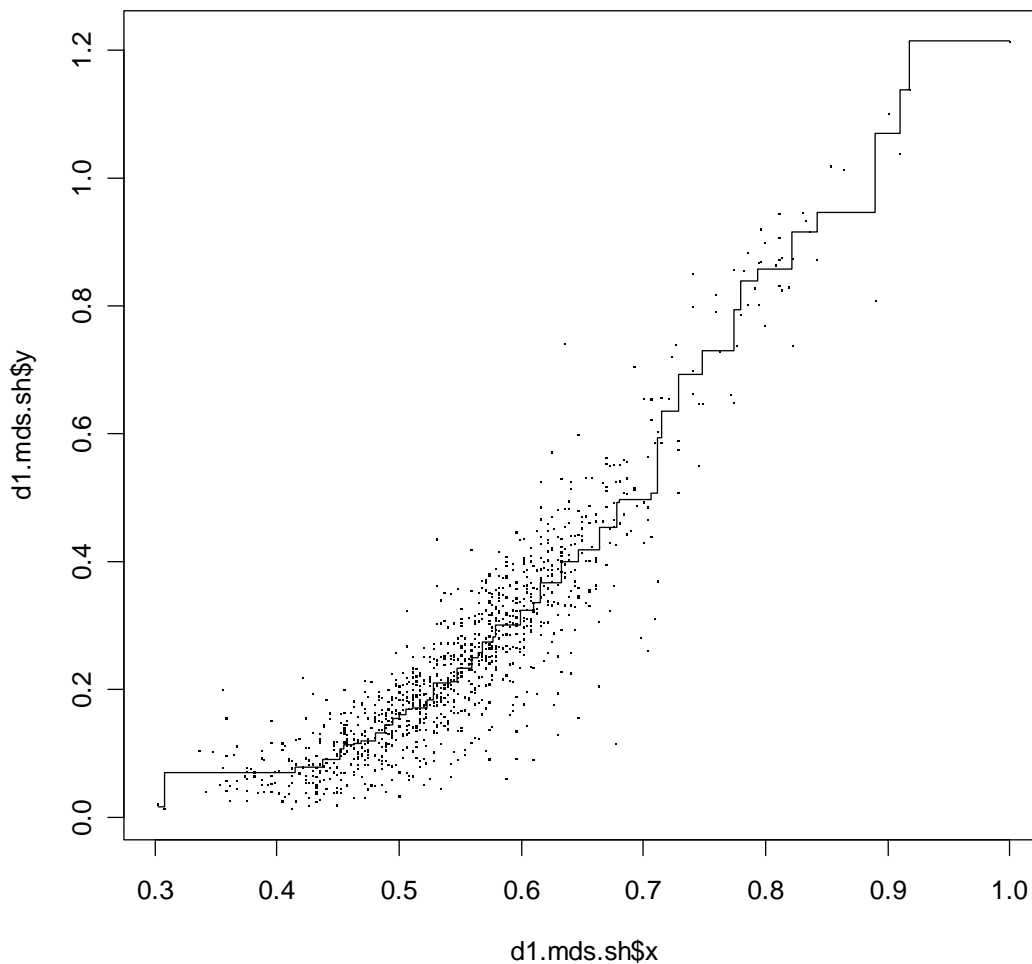


Рис. 4.10. График Шепарда для ординации рисунка 4.9.

Для обучения использовалась нейронная сеть Кохонена SOM, топологически организованная в виде прямоугольной решетки из $7 \times 7 = 49$ классов. Для более полного представления о структуре классов при выводе использовалось отображение Сэммона центров классов, кроме того, вместо топологического соседства на рисунке показано минимальное остовое дерево классов. Отсутствие «перекрещиваний» в отображении Сэммона для минимального остового дерева интерпретируется как результат хорошей ординации, т.е. положения классов на ординационной плоскости соответствуют расстояниям в пространстве признаков.

Размеры классов показывают результаты калибровки полученной ординации с использованием 48 выделенных языков (калибровка осуществляется классификацией языков методом «ближайшего соседа»). Размер символа класса пропорционален числу отнесенных к нему языков; классы, к которым не отнесен ни один язык, показаны черными точками.

Таблица отнесения языков к классам дана ниже (табл. 4.10), жирным шрифтом выделены большие группы языков, отнесенных к одному классу SOM.

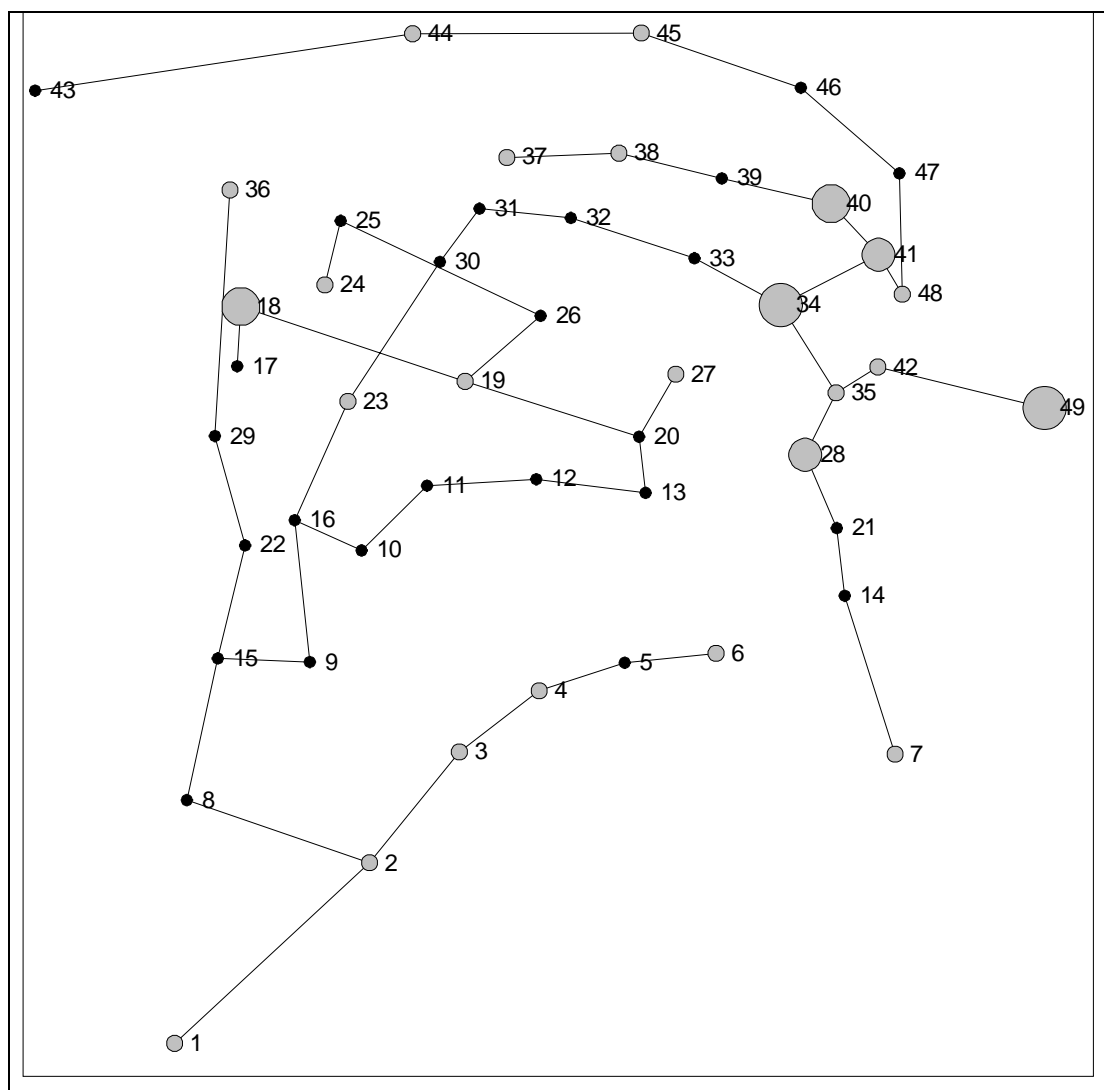


Рис. 4.11. Кластеризация языков нейронной сетью Кохонена

Легко видеть, что нейронная сеть правильно объединила уральские языки (34 класс), алтайские (близкорасположенные классы 40 и 41), имеется хорошая корреляция германских языков с классом 49 и некоторые другие. В то же время ряд решений, например, помещение лезгинского языка (32) в одну группу с германскими явно ошибочно. Требуется дальнейшая работа по настройке сети.

Т а б л и ц а 4.10

Таблица отнесения языков к классам

Язык	SOM
34	1
39	2

Язык	SOM
29	28
30	28

Язык	SOM
24	40
26	40

35	3
36	4
6	6
44	7
40	18
41	18
42	18
43	18
45	18
7	19
37	23
49	24
5	27
10	28

2	34
3	34
16	34
17	34
27	34
33	34
4	35
19	36
48	36
8	37
22	38
15	40
18	40
23	40

14	41
25	41
38	41
20	42
31	42
28	44
21	45
9	48
11	49
12	49
13	49
32	49
46	49
47	49

4.4. Обсуждение предварительных результатов интеллектуального анализа данных и кластеризации

Вполне очевидно, что залогом надежности результатов, полученных методами кластеризации и Data Mining, являются валидные исходные данные. Предварительные расчеты по мерам близости и эксперименты с указанными количественными методами показали, что точности этих данных пока недостаточно для получения однозначных выводов. Вместе с тем, эти результаты можно рассматривать как обнадеживающие. Они дают все основания для продолжения начатых работ по экспертизе контента БД и совершенствованию количественных методов.

Продолжением этих работ явилась формализация методов количественных исследований, не имеющих такой сильной зависимости от качества данных, как кластеризация и Data Mining. Для этого была разработана методика исследований, основанная на расчетах количественных портретов языковых семей, речь о которых пойдет в следующем разделе.

4.5. Резюме по главе 4

1. Мера близости языков является базовым параметром, который может быть полезен для решения многих классификационных задач. Однако выбор из континуального множества различных мер конкретной меры близости, которая была бы адекватна изучаемому объекту – сложная самостоятельная задача.
2. Выработаны общие принципы построения требуемых мер близости и с помощью численных экспериментов проанализирована группа типовых мер.
3. Создан программный комплекс, позволяющий строить классификации языков с помощью строгих методов кластерного анализа.
4. Продемонстрированы возможности применения других методов математического анализа, из которых наиболее перспективным представляется ординация с помощью пакета статистических исследований R.

Глава 5. МОДЕЛЬ СТРУКТУРНОЙ ЭВОЛЮЦИИ ЯЗЫКОВ

5.1. Общие соображения

Как отмечалось во введении, существует аналогия между языками и биологическими видами, распространяющаяся, в частности, и на их эволюцию. Родословное дерево языков – прекрасный пример заимствованной в биологии модели. Современная компаративистика рассматривает фонетические изменения как случайные (Бурлак, Старостин 2001, с.10) аналогично случайности мутаций в биологической эволюции. Но, разумеется, есть и принципиальные отличия. Как отмечается в (Nichols 1992, с.214), в языках нет аналога генома, и, соответственно, нет возможности строить теорию эволюции языков в духе законов Менделя и других законов генетики, очень точно описывающих эволюцию биологических видов.

С другой стороны, в эволюции языков важную роль играют заимствования – эволюционный механизм, у которого нет прямого аналога в живой природе. Биологические виды не обмениваются между собой генами. Исключение составляют простейшие живые организмы – вирусы, для которых прямой обмен генами, видимо, возможен (Глас 2006, с.43).

Следует обратить внимание на то, что в компаративистике, фактически, не находится места заимствованиям. Изменения ядра лексики происходят по внутренним законам развития языка, и заимствования не оказывают на них существенного влияния. На это обращено внимание в основополагающей работе М. Сводеша (русский перевод: Сводеш 1999). Он заметил, что, например, в английском языке практически половина лексики может считаться заимствованной, но из ядра лексики заимствованных оказывается лишь 6% слов (за 2 тыс. лет). По данным М. Сводеша в других языках процент заимствований в лексическом ядре еще меньший. Более того, заимствования не влияют (не увеличивают) на темпы изменений в лексическом ядре. М. Сводеш пишет (Сводеш 1999, с. 26): «Таким образом, язык, который заимствует некоторые слова из чужого языка, не изменяет свой основной словарный запас быстрее, чем другой язык, где единственным видом языкового изменения является замена одной исконной формы другой». Хотя в дальнейшем было показано, что, например, в норвежском языке целых 16% заимствований (Бурлак, Старостин 2001, с.84), и для уточнения результатов было предложено разделять в ядре исконную и заимствованную лексику (Бурлак, Старостин 2001, с.84), все же заимствования для компаративистики являются скорее досадной помехой, с которой надо как-то справляться, а не фундаментальным механизмом развития языков, требующем самостоятельного изучения.

Заимствования происходят на всех языковых уровнях, в том числе и на уровне структурных свойств, отраженных в базе данных «Языки мира». В случае, когда два языка или языковые семьи в течение продолжительного времени контактируют друг с другом, они могут приобрести много общих черт. Иногда из-за этого возникают значительные трудности в построении родословного дерева – не ясно, близки ли языки ввиду общности их происхождения или ввиду массивованных заимствований. Во введении уже упоминался сложный случай с балто-славянской группой.

Большие трудности в классификации языков возникают в так называемых диалектных континуумах. Хороший пример такового представляют тюркские языки. Крымско-татарский язык имеет три диалекта: северный, средний и южный. При этом по формальным классификационным параметрам они относятся к разным генетическим группам: северный диалект – к кыпчакско-половецкой подгруппе кыпчакской

группы тюркских языков, средний диалект – к кыпчакско-ногайской подгруппе кыпчакской группы, а южный – к огузской группе тюркских языков (Языки мира: Тюркские 1997, с.299)! И это не единичный случай: та же ситуация в узбекском, азербайджанском и в ряде других языков (Языки мира: Тюркские 1997, с.436, 171).

Широкие заимствования структурных свойств – одно из характерных отличительных свойств residual зон по (Nichols 1992, с.21). Структурные заимствования обычно изучаются на материале так называемых языковых союзов, классическим примером которого является Балканский языковой союз (Атлас 1998, с.34). Однако четких закономерностей заимствований пока не выявлено.

Еще один интересный и принципиально важный механизм языковой эволюции – появление у языков новых свойств. В базе данных «Языки мира» языки описываются с помощью почти 4000 параметров, в среднем 300-400 на один язык. Согласно наиболее популярной гипотезе моногенеза, протоязык всего человечества возник около 100 тыс. лет назад в одном месте (Африке). Естественно полагать, что вначале это был простой язык, который можно было бы описать небольшим числом параметров. Пока мы не знаем, было ли возникновение языка скачкообразным (в духе теорий Хомского) или это все же плавный процесс. Но даже если протоязык сразу возник как язык высокой степени сложности, то и в этом случае он описывался бы не более чем 400 признаками, т. е. за этот период должно было возникать, в среднем, по 4 новых признака каждые 100 лет.

Очевидно, когнитивный аппарат человека (или языковой модуль), обрабатывая имеющуюся в его распоряжении языковую информацию, стремится, с одной стороны, как-то ее упорядочить, а с другой стороны, полнее учесть действие внешних (социальных, культурных, природных) факторов. Ясно, что этот механизм генерации новых признаков существовал не только в начале языковой эволюции, но действует и по сей день, обеспечивая, в частности, дивергенцию разошедшихся языковых групп (диалектов).

По аналогии с вопросом – возник ли язык в одном месте или в нескольких местах, независимо друг от друга, – можно поставить и вопрос, возникали ли новые признаки один раз с последующим распространением путем заимствований и миграции, или имела место параллельная эволюция, когда признак возникал независимо в нескольких местах планеты.

Приведем два интересных примера возникновения нового свойства, речь идет об иерархическом строе.

М. Митун (Mithun 2004, с.20) исследовала вопрос о возникновении иерархического строя в ряде языков Северной Калифорнии (кстати, Калифорния в (Nichols 1992, с.21). отнесена к residual зонам). Эти языки (Chimariko, Yana, Yurok, Karuk) не относятся к одному семейству, поэтому их строй нельзя отнести к общему предку. М. Митун полагает, что независимое возникновение иерархического строя маловероятно, т.к. это очень редкий строй. По данным (Nichols 1992, с.90) иерархический строй встречается в 3% языков. В итоге она приходит к выводу, что строй этих языков заимствован из одного источника, хотя определить его и не удастся.

Все нахско-дагестанские языки имеют эргативный строй, который, естественным образом, восстанавливается и для их протоязыка (Языки мира: Кавказские 2001, с.163). Однако у некоторых из них (лакский, табасаранский, даргинский, бацбийский) наблюдаются элементы иерархического строя, что признается явным новообразованием, возникшим не более 500 лет назад. Соответствующие языковые примеры мож-

но найти в (Языки мира: Кавказские 2001, Табасаранские 1982). Конечно, некоторые из этих языков могли заимствовать элементы иерархического строя у какого-то из перечисленных языков. Но, по крайней мере, в одном из них эти свойства должны были возникнуть не в результате заимствований. Иерархического строя нет ни у одного из языков на расстоянии тысяч километров от северокавказского региона! Контакты с языками иерархического строя в течение последних 500 лет исключаются. Поэтому, по крайней мере, в двух регионах мира иерархический строй возник независимо друг от друга.

К сожалению, таких ясных примеров генерации новых признаков с полностью исключенными заимствованиями мало, и они плохо изучены.

База данных «Языки мира» предоставляет новые исследовательские возможности для изучения генерации новых свойств и механизмов заимствований, давая исчерпывающий материал по статистике распределения свойств по языкам мира.

5.2. Заимствования vs. параллельная эволюция

В данном параграфе будет рассмотрен вопрос: каков основной способ появления в языке новых признаков – заимствования из других языков или независимая генерация. Сравнительно-исторический метод не может помочь при ответе на этот вопрос в виду того, что он лишь в небольшой степени применим к анализу структурных (типологических) свойств языков и, кроме того, он ориентирован на «вертикальные» связи языков (предок-потомок), а не на «горизонтальные» заимствования.

Учитывая огромное число признаков и языков, нет реальной надежды в обозримой перспективе получить детальную картину индивидуального поведения признаков – где каждый из них возник и каким языкам передавался в процессе эволюции. Так же как и в статистической физике, здесь придется довольствоваться статистическими закономерностями. Для этих целей удобно воспользоваться массивом информации, хранящейся в базе данных «Языки мира». Приведем далее два примера анализа распределения признаков. Используемые здесь методы исследований больше напоминают методы естественных наук (в частности, эксперимент), чем собственно лингвистические.

Эксперимент 1. Рассмотрим сначала простейшую ситуацию – распределение признаков, существующих ровно в двух языках. В нашей базе данных таких насчитывается 455. Выделим из них те, в которых языки относятся к разным языковым семьям, причем будут учтены только основные семьи и ареальные группы, имеющие достаточно большое количество языков. Это индоевропейская, алтайская, уральская, афразийская семьи, палеоазиатские и кавказские языки.

Методика подсчетов. Ясно, что если одна группа содержит больше языков, чем другая, то у нее больше шансов будет представленной в искомым признаках. Чтобы нивелировать влияние числа языков в группе, применяем следующий метод. Для каждой пары групп подсчитаем число признаков указанного вида (т. е. признаков, присутствующих ровно в двух языках – один из одной группы, другой – из другой), затем поделим его на произведение числа языков в данных группах. Вычисления велись с точностью до 0,001.

Априорно можно ожидать следующие результаты. Если основным механизмом являются заимствования, то следует ожидать, что чем ближе географически друг к другу две группы, тем большее число таких признаков они будут содержать. Если основным механизмом является независимая генерация новых свойств, то рассматриваемые признаки должны быть распределены по парам групп достаточно равномерно.

Наконец, можно предположить, что признаки, присутствующие всего в двух языках – это остатки признаков, ранее существовавших в протоязыках и затем по какой-то причине исчезающих. В этом случае следует ожидать обнаружения генетической близости соответствующих групп. Например, если исчезают какие-то признаки, присутствовавшие в протоностратическом языке, то их следы естественно обнаружить преимущественно в индоевропейской, алтайской, уральской и афразийской семьях, а не среди палеоазиатских и кавказских языков. Упорядоченные по величине результаты приведены в таблице 5.1.

Т а б л и ц а 5.1

Частоты признаков, встречающихся ровно в двух языках по парам групп

Пара языковых сообществ	Частота
Алтайские – афразийские	0,009
Уральские – палеоазиатские	0,0081
Уральские – алтайские	0,007
Алтайские – палеоазиатские	0,007
Индоевропейские – афразийские	0,007
Афразийские – кавказские	0,006
Индоевропейские – кавказские	0,006
Кавказские – уральские	0,006
Кавказские – палеоазиатские	0,006
Индоевропейские – уральские	0,005
Афразийские – уральские	0,005
Индоевропейские – алтайские	0,004
Алтайские – кавказские	0,003
Индоевропейские – палеоазиатские	0,003
Афразийские – палеоазиатские	0

Проанализируем полученные данные. Высокая частота в паре алтайские – афразийские является исключением (возможно, случайностью или следствием ошибки в данных), требующим отдельного исследования. Но в последующих нескольких строках присутствуют только пары географически близких семей. Причем строки 2-4 занимают семьи из северо-восточной части Евразии, а строки 5-7 – семьи из юго-западной части. Характерно, что последнюю строку занимает пара наиболее удаленных друг от друга групп: палеоазиатские и афразийские языки, среди которых не обнаружилось ни одного признака рассматриваемого вида.

Представляется, что эти данные однозначно свидетельствуют о приоритетности механизма заимствований. Хотя, как говорилось выше, вполне возможна независимая генерация в разных языках одинаковых свойств, однако совершенно неясно, как на основе такого механизма параллельной эволюции можно объяснить полученные данные. Нет данных и о приоритете генетической близости. Среди самых высокочастотных оказались пары уральские – палеоазиатские и алтайские – палеоазиатские, в которых языки относятся к разным макросемьям.

Эти результаты также не могут быть и следствием ошибок в базе данных. Случайные ошибки образуют шум, который может затемнить существующие закономерности, но не может создать закономерность там, где ее нет.

Отметим, что отбор признаков, присутствующих именно в двух языках, не является принципиальным. Аналогичные результаты получаются и при рассмотрении признаков, встречающихся в трех, четырех и т. д. языках. Два языка выбрано просто для демонстрации.

Таким образом, анализ распределения признаков может дать информацию о заимствованиях. В эксперименте 1 речь шла о распределении признаков в глобальном масштабе – семей и еще больших групп языков. В следующем примере мы применим аналогичную методику для несколько более тонкого анализа – на уровне ветвей. Будут рассмотрены некоторые структурные заимствования тюркских языков. Естественно ожидать, что заимствования – как лексические, так и структурные – будут происходить из одних и тех же языковых групп. Поскольку лексические заимствования для тюркских языков достаточно хорошо изучены, можно сравнить полученные результаты и тем самым протестировать предложенный метод.

Эксперимент 2. Рассмотрим все признаки, присутствующие ровно в двух языках, в точности один из которых – тюркский. Таких оказалось 60. Вторые языки, в которых встречаются эти признаки, распределены по группам следующим образом.

Иранские – 10, славянские – 8, уральские – 8, тунгусо-маньчжурские – 6, монгольские – 5, палеоазиатские – 4, семитские – 3, абхазо-адыгейские – 3, дардские – 3, романские – 2, нахско-дагестанские – 2, индоарийские – 2, по одному – тибето-бирманские, шумерский, германские, нуристанские.

Полученное распределение хорошо коррелирует и с данными о лексических заимствованиях (Языки мира: Тюркские 1997). Хорошо известно, что значительное влияние в плане лексики на тюркские языки оказали иранские, славянские, монгольские, семитские языки. Известны и заимствования из уральских и тунгусо-маньчжурских языков. Хотя в наших данных доля уральских и тунгусо-маньчжурских языков несколько неожиданно велика, однако этот результат можно объяснить следующим образом. Полученные таким методом данные свидетельствует лишь о контактах тюркских языков с другими группами, а не о направлении заимствований. Так что большая доля уральских и тунгусо-маньчжурских языков может объясняться заимствованиями, наоборот, из тюркских языков в эти группы. По крайней мере, в отдельных случаях это верно (Языки мира: Монгольские 1997, с.302).

Как и следовало ожидать, в этом эксперименте не обнаружилось заимствований из кельтских, австронезийских и других языков, с которыми у тюрков вряд ли были серьезные контакты. Важно отметить высокую разрешающую способность данного метода: обнаруживаются даже слабые влияния. Например, заимствования из тибетских языков, что подтверждается на лексическом уровне (Языки мира: Тюркские 1997). Как и в предыдущем эксперименте, выбор числа 2 не является принципиальным.

Важно заметить, что результаты о заимствованиях и контактах получены применением чисто статистической процедуры (элементарных подсчетов) к заложенным в базу данных описаниям структурных свойств языков. Здесь ни в какой форме не применяется сравнительно-исторический метод с установлением регулярных соответствий и реконструкцией протосостояний.

5.3. Диаграммы распределения частотности признаков

Выше мы видели, что интерес могут представлять признаки, присутствующие ровно в двух языках. Перейдем к общей картине, рассматривая вместо двойки произвольное число. База данных «Языки мира» позволяет получить следующую статистику: для каждого числа n , сколько признаков существует ровно в n языках. На диаграмме (рис. 5.1) приведены соответствующие данные в форме графика. В дальнейшем графики этого типа будем называть ЯП-диаграммами (язык-признак диаграмма). Признак, существующий ровно в n языках, будем называть n -языковым признаком.

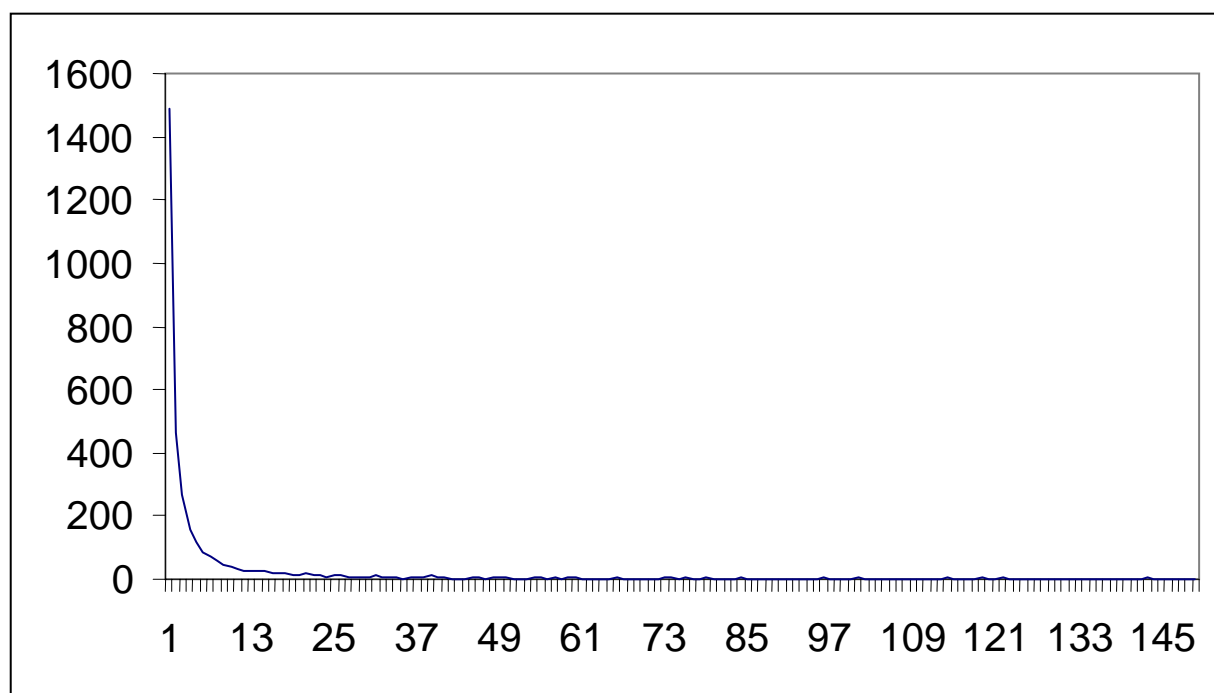


Рис. 5.1. ЯП-диаграмма всех языков базы данных (левая половина графика)

Правая половина графика имеет такой же вид практически прямой линии, как и часть от 50 до 150, представленная здесь. Внешний вид графика (крайне близкий к гиперболе) на первый взгляд вызывает удивление. Без обширной базы данных получить информацию такого рода не представляется возможным, поэтому ранее ничего похожего не публиковалось. Объединение в одном месте огромного количества данных практически неизбежно создает новые объекты для исследований и обнаруживает новые явления. Неясно, как можно теоретически предсказать этот вид ЯП-диаграммы на основе, скажем, теории Хомского или любой другой продвинутой лингвистической теории.

Объяснения требуют несколько моментов.

- а) Почему большая часть графика является почти константной?
- б) Почему такое большое число 1-языковых признаков?
- в) Почему кривая так быстро падает и имеет вид гиперболы?
- г) Почему имеет место небольшой подъем в правой части графика?

Прежде чем отвечать на эти вопросы, построим еще несколько ЯП-диаграмм. График на рис. 5.1 отражает свойства языков области LENCА. Стоит сначала выяснить, как будет выглядеть график, построенный для иных групп языков. К сожалению, в нашей базе данных пока нет языков из других регионов мира, но можно построить ЯП-диаграммы для более узких

групп языков. На диаграмме (рис. 5.2) приведены графики, построенные для тюркских и кавказских языков. В дальнейшем будут приведены ЯП-графики и еще для ряда языковых групп.

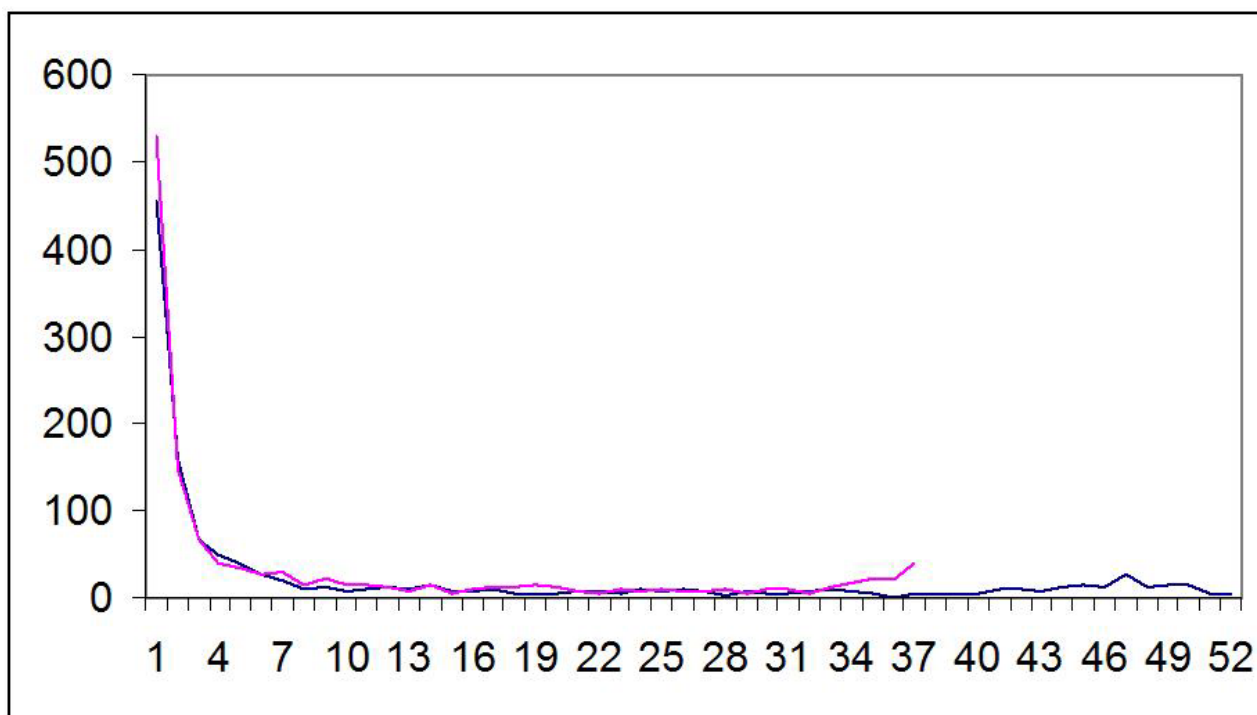


Рис. 5.2. Кавказские (более короткая линия) и тюркские языки

Как мы видим, основные особенности графиков сохраняются. Кроме того, бросается в глаза поразительная схожесть ЯП-диаграмм этих двух групп языков. И это несмотря на то, что взяты существенно отличающиеся друг от друга группы. Тюркские языки образуют одну семью, причем, весьма однородную. Кавказские же включают три семьи, относящиеся к двум разным макросемьям: картвельская семья относится к ностратической макросемье, а абхазо-адыгская и нахско-дагестанская семьи – к сино-тибетской макросемье (Бурлак, Старостин 2001).

Схожесть ЯП-диаграмм, построенных для совершенно различных групп языков, указывает на то, что имеют место некие весьма жесткие законы, определяющие распределение языковых признаков. Конечно, возникает вопрос, не является ли эта форма графиков неким артефактом?

Ясно, что столь четкую закономерность нельзя объяснить никакими случайными ошибками базы данных. Единственное возможное альтернативное объяснение состоит в наличии некой системной ошибки во всей современной методологии описания языков. Если ситуация такова, то это потребует пересмотра едва ли не большей части существующей лингвистики. В дальнейших исследованиях мы исходим из предположения, кажущегося более реалистичным: современная лингвистика дает в целом удовлетворительные описания языков, и существуют некие законы, управляющие распределением признаков и обуславливающие такой вид ЯП-диаграмм в целом. Отвечая на вопросы А-Г, мы как раз и планируем выйти на эти законы.

Таким образом, с помощью простых числовых подсчетов на материале базы данных «Языки мира» обнаружено новое явление, состоящее в гиперболической форме ЯП-диаграммы и постоянстве этой формы для различных языковых групп.

Будут рассматриваться три типа языковых изменений: «внутренняя» эволюция языка, заимствования и генерация новых признаков.

Посмотрим, как отражаются на графике внутренние изменения в языках, происходящие в результате имманентного языкам свойства изменяться. Согласно установленному в типологии взгляду, эволюция языков не является направленной. Это и понятно, если бы эволюция имела направленный характер, то за 100 тыс. лет все языки уже пришли бы к своему финальному состоянию.

Различные состояния сменяют друг друга по циклу. Напомним классическую последовательность фаз А. Шлейхера: изолированные → агглютинативные → флективные, согласно современным воззрениям (Croft 1990, с.229) дополняемую завершающим цикл переходом: флективные → изолированные.

В каждый момент времени часть языков находится в одном из этих состояний, часть в другом, часть в третьем. В итоге устанавливается некий баланс между частотами состояний.

Хотя это относится к совокупности всех языков мира, но то же самое можно сказать и о множестве языков, представленных в нашей базе данных. Дело в том, что, во-первых, представленные в базе данных языки покрывают четко отграниченную область – LENSА, характеризующуюся географической, культурной и языковой спецификой. Языковые контакты LENSА с языками соседних областей – Африки к югу от Сахары, Южной и Юго-восточной Азии очень малы. Во-вторых, эта область была заселена при расселении человечества из Африки первой, не менее 60 тыс. лет назад (Nichols 1992, с.275), т. е. имелось вполне достаточно времени для установления на этой территории стационарного состояния.

Средней части ЯП-диаграмм (причем большая их часть, приблизительно от $x=20$ до $x=300$) отражает частоту признаков, получивших достаточное распространение (присутствующих в среднем в 160 языках), т. е. не новых, а уже достаточно укоренившихся в системах языков и, вероятно, по большей части вошедших в вышеописанные циклы. Цикличность и ответственна за константный вид средней части ЯП-диаграмм.

Допустим, что признак X заменяется в некотором языке на признак Y . Если до этой замены X встречался в n языках, а Y – в k языках, то после замены X будет встречаться в $n-1$ языке, а Y – в $k+1$ языках. На ЯП-диаграмме это отразится тем, что значение в точке n уменьшится на 1, в точке $n-1$ – увеличится на 1, в точке k – уменьшится на 1, в точке $k+1$ – увеличится на 1. Короче это можно сформулировать так: признак X сместится на графике на одну позицию влево, а признак Y – на одну позицию вправо.

Такие изменения происходят параллельно в различных языках. В силу цикличности развития, в каком-то другом языке осуществляется переход от признака Y к признаку X и на графике эти признаки сдвинутся в противоположную сторону. В развиваемом здесь подходе не прослеживается судьба конкретных признаков, а изучаются лишь общие статистические закономерности, т. е., говоря математическим языком, строится вероятностная модель.

В рамках вероятностного подхода принимается, что структурные изменения независимы друг от друга. Это верно в двух смыслах. Во-первых, независимыми являются изменения одного и того же признака в разных языках. Во-вторых, независимыми являются изменения различных признаков в одном и том же языке. Отметим, что постулат о независимости структурных изменений в разных языках точно соответствует постулату о независимости изменений в базисной лексике (Арапов, Херц 1974, с.21, 25), на котором строится глоттохронология.

Считая, что рассматривается достаточно большой интервал времени, будем моделировать динамику признаков случайными блужданиями точек (признаков), причем вероятность сдвига признака вправо равна вероятности сдвига признака влево (это следует из вышеописанной цикличности процессов внутренних изменений). Для признака X обозначим эту вероятность $\alpha_1(X)$.

Ясно, что если бы для какого-то признака равенство вероятностей сдвигов вправо и влево нарушалось на протяжении длительного периода времени, то признак скатился бы либо к концу ЯП-диаграммы, приобретя стопроцентную встречаемость и превратившись в языковую универсалию, либо к ее началу, и вообще выпал бы из употребления.

Не имея априорной информации о зависимости $\alpha_1(X)$ от X , полагаем, что для каждого признака X из рассматриваемой части графика $\alpha_1(X)$ выбирается случайным образом из некоторого интервала, скажем, от 0 до 0,05. При таких допущениях мы получим почти прямую линию с небольшими случайными флюктуациями. Именно такая картина и наблюдается на ЯП-диаграммах, что является подтверждением адекватности выбранного подхода к моделированию и правильности сделанных допущений. Фрагмент графика (от $x = 76$ до $x = 100$) из рис. 5.1 приведен на рис. 5.3 в увеличенном виде.

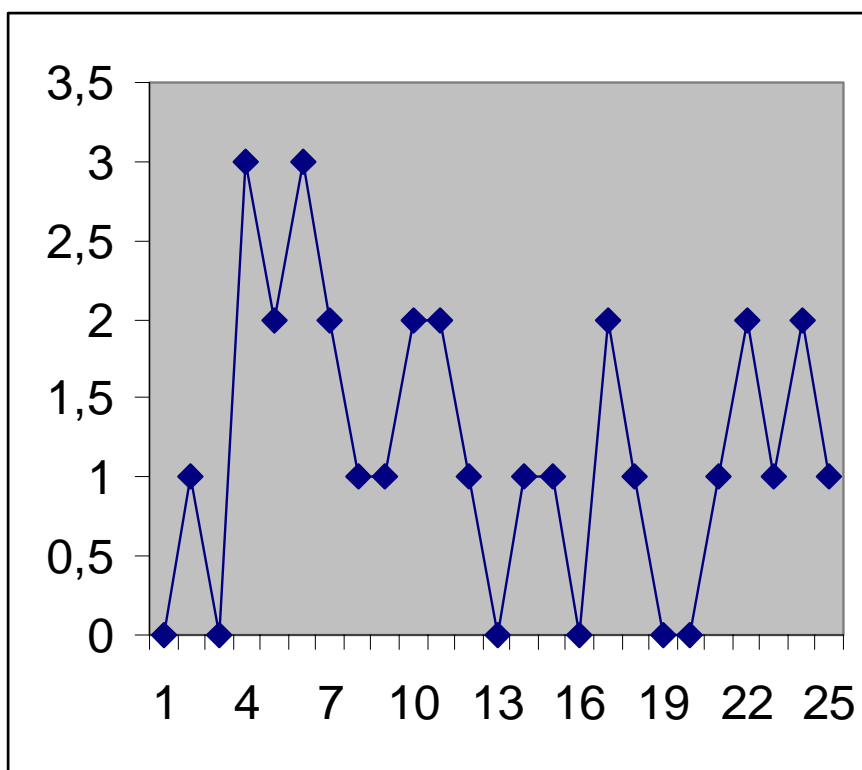


Рис. 5.3. Фрагмент ЯП-диаграммы БД для $x = 76 \dots 100$

Перейдем к вопросу Б. Для объяснения столь большого числа 1-языковых признаков и вообще признаков, присутствующих в небольшом числе языков, может быть предложена модель в духе когнитивной лингвистики. Одна из основных и наиболее удачных идей когнитивной лингвистики состоит в том, что многие языковые структуры имеют четко выраженный стабильный центр и нечеткую, подвижную перифе-

рию. Это подтверждается и психологическими экспериментами (Rosch 1975). Стабильность центра обеспечивает преемственность языка, его передачу из поколения в поколение. Гибкость же периферии обеспечивает возможность изменений языка.

Рассматривая совокупность языковых признаков, естественно также выделить центр, состоящий из давно возникших и имеющих достаточно широкое распространение признаков, изучением которых собственно и занимаются типологи. На ЯП-диаграмме они занимают среднюю и правую часть графика. Периферию же составляют молодые признаки, еще не успевшие широко распространиться и занимающие левую часть графика. Является невероятным, чтобы все вновь возникающие признаки сохранялись и распространялись в языках мира. Тогда бы признаков в итоге оказалось слишком много. Как и остальные признаки, 1-языковые признаки могут подвергнуться вытеснению другим признаком, но в этом случае они исключаются из пула признаков окончательно. Для n -языковых признаков при небольших n также относительно велика вероятность исчезновения в результате последовательного вытеснения из всех n языков.

Таким образом, возникает следующая картина. В языках постоянно генерируются новые признаки, которые как бы ведут борьбу за существование. Многие из них вскорости исчезают, некоторым удается распространиться на несколько языков и закрепиться в пуле признаков на более или менее длительный срок. Эти признаки занимают левую часть ЯП-диаграммы и как раз и образуют динамичную периферию, необходимую для эволюции языка. Естественно, что периферия должна быть достаточно большой по размерам. Если каждый язык для обеспечения возможностей развития генерирует хотя бы по несколько новых признаков, то все 315 языков базы данных в совокупности будут иметь более тысячи 1-языковых признаков, что и наблюдается в действительности.

Разумеется, 1-языковые признаки могут и не быть абсолютно новыми, а могут быть заимствованы из каких-либо языков, не включенных в базу данных, в том числе из языков, располагающихся вне области LENCА. Но это не меняет картину принципиально, так как даже если бы мы располагали исчерпывающей базой данных, содержащей все языки мира, все равно она должна включать определенное количество новых признаков, появление которых обеспечивает языковое разнообразие.

Можно провести аналогию между языковыми признаками и коммерческими структурами. Лишь небольшая часть из них достигла финансовых высот и обеспечивает стабильность мировой экономической системы. В условиях свободного рынка ежедневно создаются тысячи мелких фирм, большинство из которых быстро разоряется.

Описанный при обсуждении вопроса А механизм циклического развития касается давно укоренившихся в языках мира признаков. Вновь возникающие признаки еще не включены в систему, поэтому требуется отдельное рассмотрение механизма их распространения. Новый признак, существующий поначалу в одном-единственном языке, может распространиться на другие языки путем заимствований.

Если язык А заимствует признак Х в языке В, то Х должен быть хорошо интегрирован в систему языка А. Вряд ли он заимствуется на «пустое место», более вероятно, что в системе языка А уже существовало более или менее подходящее для Х место, занимавшееся другим признаком Y, т. е. происходит вытеснение признака Y признаком Х.

В качестве примера упомянем интересную ситуацию в языковом союзе Amdo Sprachbund, включающем тюркские, монгольские, синитские и тибетские языки. Как показал Ю. Янхунен (Janhunen 2004, с.70), все языки этого региона заимствовали

концепцию перспективы из тибетских языков. При этом тюркские языки потеряли личную маркировку предикатов. Ю. Янхунен это связывает с тем, что перспектива и личная маркировка предикатов очевидным образом перекрываются, а присутствие двух параллельных явлений в одном языке маловероятно. Таким образом, здесь постулируется вытеснение признака заимствованным. Причем следует заметить, что перспектива в языках мира встречается значительно реже, чем личная маркировка предикатов.

Итак, мы вновь пришли к той же схеме, что и выше, когда один из признаков сдвигается на 1 влево, а другой на 1 вправо, значит общее число сдвигов вправо равно общему числу сдвигов влево. Однако для численных расчетов здесь потребуются несколько более тонкие соображения, чем выше.

Под термином признак понимается абстрактный признак, применимый ко всем языкам, например, «существование артикля». Под конкретным признаком будем понимать значение абстрактного признака в конкретном языке. Например, «существование артикля в английском языке».

Вообще говоря, в данном случае, нельзя утверждать, что для каждого отдельного конкретного признака вероятности заимствования другим языком и вытеснения этого признака из языка равны. Однако поскольку в каждом акте заимствования у одного признака увеличивается его присутствие в пуле, а у другого уменьшается, то, усредняя по множеству всех конкретных признаков, участвующих в процессах заимствований, вводим единый параметр вероятности заимствования или вытеснения конкретного признака – α_2 . Такой подход представляется правомерным, поскольку нет априорной информации о том, будет ли признак заимствоваться или вытесняться, а также в виду большого, статистически значимого количества конкретных признаков (более 100 тыс.).

На основе сделанных допущений построим математическую модель динамики молодых признаков и сопоставим полученные с помощью этой модели результаты с фактическими данными. Переходя к ответу на вопрос В, сконцентрируем внимание на левой части ЯП-диаграммы. По внешнему виду зависимость является гиперболической.

Следующее допущение, которое мы делаем, – считаем современное состояние языков более или менее стационарным. К сожалению, лингвистика не располагает точными и подробными описаниями протоязыков, так что нет надежных данных для сопоставительных исследований. Некоторые результаты в этом направлении могут быть получены при изучении мертвых языков из базы данных. Эти результаты будут приведены в последующих параграфах. А пока, поскольку мы опять-таки не располагаем данными, позволяющими предположить существенное увеличение или уменьшение числа признаков в левой части ЯП-диаграммы за исторически обозримое время, полагаем, что оно не меняется со временем.

Пусть S – число новых признаков, добавляющихся в единицу времени, во всех языках базы данных вместе взятых. α_2 – вероятность исчезновения или заимствования конкретного признака. $Y(x)$ – число x -языковых признаков.

Запишем уравнения стабильности для первых нескольких точек графика. Для этого подсчитаем (приближенно) число добавляющихся и удаляющихся признаков.

В точке 1 имеется $Y(1)$ признаков. Поскольку процесс вероятностный, то невозможно точно сказать, сколько из них исчезнет и сколько будет заимствовано. Однако из теории вероятностей известно (Феллер 1984), что наиболее вероятным является число $\alpha_2 Y(1)$ для тех и других. Суммарно значение $Y(1)$ уменьшится на $2\alpha_2 Y(1)$.

В точке 2 имеется $Y(2)$ признаков, каждый из которых встречается в двух языках. Каждый из этих $Y(2)$ признаков имеет по две конкретные реализации в двух язы-

как. Вероятность того, что ровно один из этих конкретных признаков будет вытеснен, равна $2\alpha_2(1 - \alpha_2) = 2\alpha_2 - 2\alpha_2^2$. Полагая, что вероятность заимствований невелика, отбросим слагаемое $2\alpha_2^2$, как величину второго порядка малости. Например, при $\alpha_2 = 0,01$, $\alpha_2^2 = 0,0001$, что изменяет всю сумму лишь на 1%. Неточность исходных данных делает бессмысленным проведение вычислений со столь высокой точностью. $2\alpha_2 Y(2)$ – наиболее вероятное приращение величины $Y(1)$ за 2-языковых признаков. В принципе, вклад дают и 3-языковые признаки. Однако для этого признак должен быть вытеснен из двух языков одновременно, вероятность чего пропорциональна α_2^2 , и этим вкладом также можно пренебречь. В итоге для точки 1 уравнение стабильности выглядит следующим образом:

$$C + 2\alpha_2 Y(2) - 2\alpha_2 Y(1) = 0. \quad (5.1)$$

Преобразуем его в форму

$$Y(2) = Y(1) - S/2, \text{ где } S = C/\alpha_2. \quad (5.2)$$

Аналогично, уравнение стабильности для точки 2 (как и выше, учтен вклад только ближайших точек):

$$\alpha_2 Y(1) + 3 \alpha_2 Y(3) - 4 \alpha_2 Y(2) = 0 \quad (5.3)$$

или

$$Y(3) = Y(1) - 2S/3. \quad (5.4)$$

В общем случае получим

$$Y(x) = Y(1) - (x-1)S/x = (Y(1) - S) + S/x. \quad (5.5)$$

Это как раз уравнение гиперболы. Хотя построенная математическая модель является приближенной, все же, видно, что из теоретических соображений получена зависимость, на качественном уровне коррелирующая с формой левой части ЯП-диаграммы.

Полученный результат, с одной стороны, объясняет причины столь быстрого падения графиков на диаграммах (рис. 5.1-5.2), а с другой стороны, хорошее согласование теоретически предсказанных результатов с экспериментальными данными является сильным аргументом в поддержку сделанных теоретических допущений.

Обсуждение вопроса Г отложим до следующей главы. Пока перейдем к общему анализу модели и некоторых следствий из нее.

5.4. Анализ математической модели динамики признаков

Выбор статистического подхода мотивируется следующими обстоятельствами. Прежде всего, эволюционные изменения сами по себе имеют случайный характер (Бурлак, Старостин 2001, с.10): «Очень важно подчеркнуть, что изменения в принципе являются случайными: практически невозможно предсказать, произойдет или не

произойдет некоторое конкретное изменение. В этом смысле языковые изменений сходны с биологическими мутациями».

Во-вторых, даже если бы у нас была информация об индивидуальных процессах, проследить в точности динамику всей языковой системы крайне трудно в связи с ее большим размером. Используем терминологию математического аппарата статистической физики (Пенроуз 2004).

Если для описания состояния системы объектов требуется n параметров, то ее поведение описывается движением точки в n -мерном фазовом пространстве. В нашем случае состояние одного языка описывается в базе данных 3823 признаками, а поскольку всего языков 315, то общее число требуемых для описания состояния системы параметров равно $3823 \times 315 = 1204245$. Т. е. для моделирования потребуется более чем миллион-мерное пространство. Поведение одной частицы в классической механике описывается 6 параметрами (по три параметра на положение и импульс). Так что аналогичная размерность возникает при рассмотрении эволюции системы из 200000 частиц. При таких размерностях в физике применяется не точное, а вероятностное моделирование.

Итак, в основу развиваемой модели эволюции языков положены следующие постулаты:

- существование двух механизмов эволюции: внутренние изменения и заимствования;
- постоянное возникновение и исчезновение признаков из пула всех признаков языкового сообщества с преобладание возникновения новых признаков;
- стабильность языковой системы в целом;
- случайность изменений.

В математической формализации используется также постулат о независимости структурных изменений. В качестве обоснования этих постулатов и их математической формализации, приведенной в предыдущем параграфе, можно привести следующие доводы.

1. Модель построена на основе идей, развитых в компаративистике и типологии.
2. Хорошее согласование с другими классическими теориями, как в лингвистике, так и вне ее. Из лингвистических теорий здесь имеется в виду когнитивная лингвистика, из других наук – эволюционная биология. Это уже обсуждалось выше.

Хотя близость к теориям из других наук не является прямым аргументом, все же, как отмечает, например, Л. Лодэн (Laudan 1977), частично правдоподобность новой научной парадигмы связана с ее теоретической согласованностью с более установившимися научными парадигмами в близких областях. По мнению У. Крофта (Croft 1990), наиболее важной параллелью между эволюционной биологией и функционально-типологическим подходом в лингвистике является изменчивость систем и то, что структура систем частично объясняется их историей.

3. Как известно, лучшим критерием истинности теории является ее способность объяснить наблюдаемые факты. Предложенный математический формализм позволяет объяснить все основные особенности формы ЯП-диаграмм, построенных на основе базы данных.

4. Обычно новые теории хотя бы в некоторых аспектах демонстрируют лучшие результаты, чем предшествующие. В данном случае теория развита для объяснения совершенно новых экспериментальных данных, полученных с помощью уникальной базы данных «Языки мира». Естественно, что раньше не существовало теорий для

объяснения наблюдаемых эффектов. Поэтому имеет смысл попытаться предложить иные посылки и построить альтернативное объяснение.

Попробуем построить альтернативное объяснение высоким значениям функции $Y(X)$ в левой части графика. В предложенной модели это связывалось с преобладающим добавлением новых признаков. Предположим теперь обратное. Допустим, что сколько-то тысячелетий назад языки отличались большей сложностью (что само по себе несколько сомнительно) и описывались большим числом признаков, а затем пошел процесс упрощения языков с массивованным вымыванием признаков. В этом случае вероятность утраты признака будет больше вероятности заимствования, и будет происходить накопление 1-языковых признаков. Напишем аналогичные уравнения с α – вероятностью утраты признака, β – вероятностью заимствования и $\alpha > \beta$. Для точки 1 имеем:

$$2Y(2)\alpha - Y(1)\alpha - Y(1)\beta = 0 \text{ или } Y(2) = Y(1)/2 + Y(1)\beta/2\alpha. \quad (5.6)$$

Так как все входящие в уравнения величины положительны, то $Y(2) \geq Y(1)/2$. Это неравенство находится в противоречии с экспериментальными данными. На диаграмме 1 $Y(1) = 1488$ и $Y(2) = 466$ и требуемое неравенство даже приблизительно не выполняется. Аналогичная ситуация и на других диаграммах, построенных для более узких групп языков. Резкое несоответствие альтернативной модели экспериментальным данным является дополнительным аргументом в пользу основной модели.

5. Обычно формальные теории оказываются продуктивными и за пределами того круга вопросов, для решения которых они создавались. После того, как ответ на сформулированные выше вопросы А-В найден, имеет смысл посмотреть какие еще результаты могут быть получены на этом пути.

Значения параметров модели – констант S и α_2 только из полученных уравнений определить не удастся, т. к. система уравнений оказывается недоопределенной. Но можно определить их отношение $S = C/\alpha_2$. Из уравнения (*), подставив известные из диаграммы 1 значения $Y(2) = 466$ и $Y(1) = 1488$, получим $S = 2044$. S является абстрактным параметром, который можно интерпретировать как мощность процессов новообразования структурных свойств в группе языков. Посмотрим, какие значения принимает S для других групп языков и какую можно извлечь отсюда информацию о группе языков в привычных терминах.

Хотя абсолютные значения параметров S и α_2 , влияющих на величину S , неизвестны, но можно выделить факторы, которые предположительно влияют на их величину.

Рассмотрим влияние двух факторов: компактность языковой группы и число языков в ней. Для этого сначала сравним ЯП-диаграммы двух групп языков из разных семей, но близких по числу языков в группе и вполне сопоставимых по степени и генетической, и ареальной компактности. Это славянские языки и огузская группа тюркских языков. Для них $S = 468$ и 454 соответственно. Так как изучаемые параметры S и α_2 влияют только на левую часть графика, то в этом разделе будут даваться в увеличенном виде именно левые фрагменты графиков.

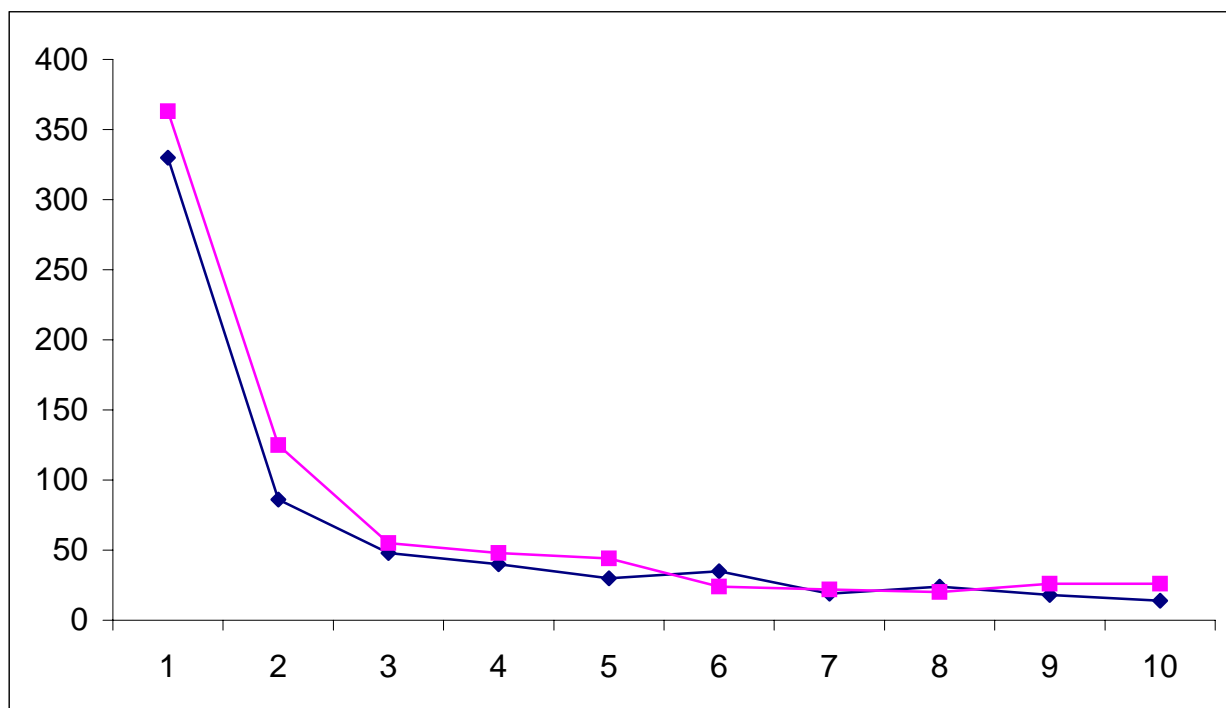


Рис. 5.4. Славянские (светлая линия) и огузские языки

Как видим, графики как по форме, так и по абсолютным значениям очень близки.

Сформируем тестовую группу из 52 языков (она понадобится в дальнейшем и для других целей), входящих в разных семьи и географически рассредоточенных. Она включает: афганский, тиграйский, бартаганский, ванеци, башкирский, астраханских ногайцев-карагашей, бретонский, немецкий, еврейский (совр. иврит), бирманский, карачаево-балкарский, дари, уйгурский, нганасанский, кара-калпакский, валлийский, ирландский, кетский, арагонский, мэнский, ногайский, нанайский, ненецкий, монгольский, саларский, сарыг-югурский, арумынский, сонкорско-тюркский, шотландский, польский, татарский, ваханский, ишкашимский, тофаларский, фарса диалекты, тувинский, монгорский, белорусский, астурийский, бирманский, турецкий, галисийский, хинди, гасконский, хазара, чулымско-тюркский, хуфский, урду, чувашский, печенежский, бурушаски, юкагирский.

Сравним ее с тюркскими языками. Для тюркских языков $S = 590$, для тестовой группы $S = 838$. Разрозненность языков в группе повышает значение параметра S примерно на 40%. Возрастание параметра S на качественном уровне можно объяснить следующим образом. При переходе к географически и генетически удаленным языкам вероятность заимствований (α_2) падает, и поскольку α_2 стоит в числителе, то отношение $S = C/\alpha_2$ возрастает.

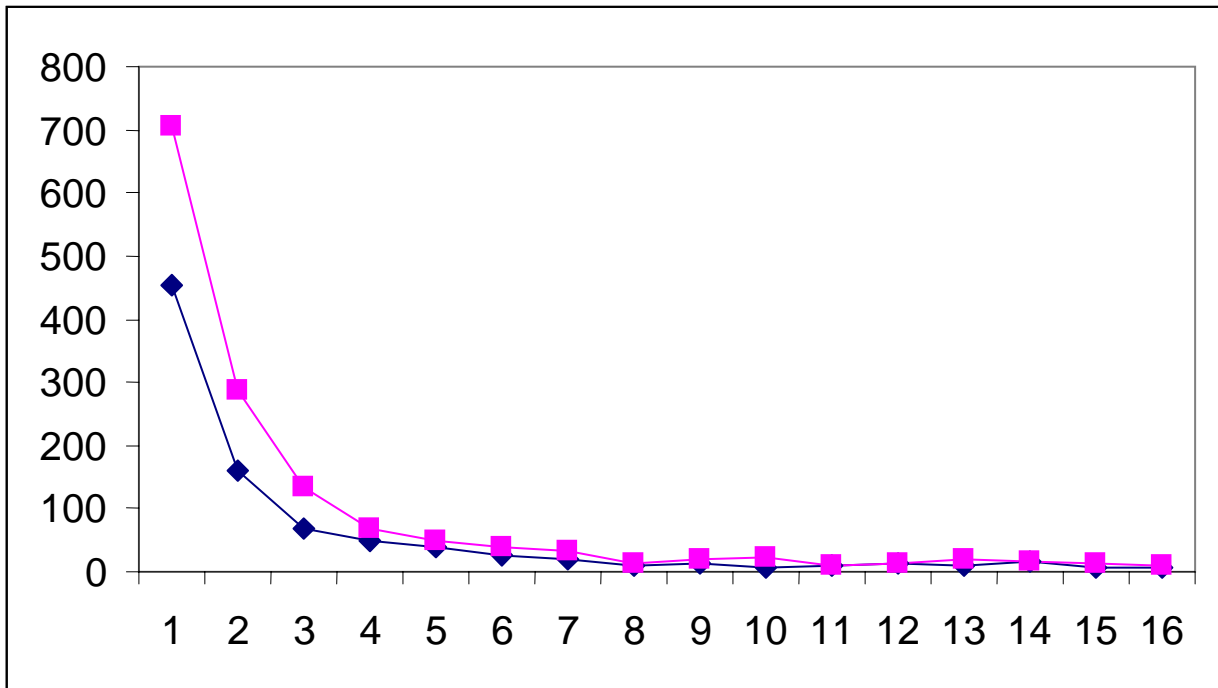


Рис. 5.5. Тюркские языки (темная линия) и тестовая группа

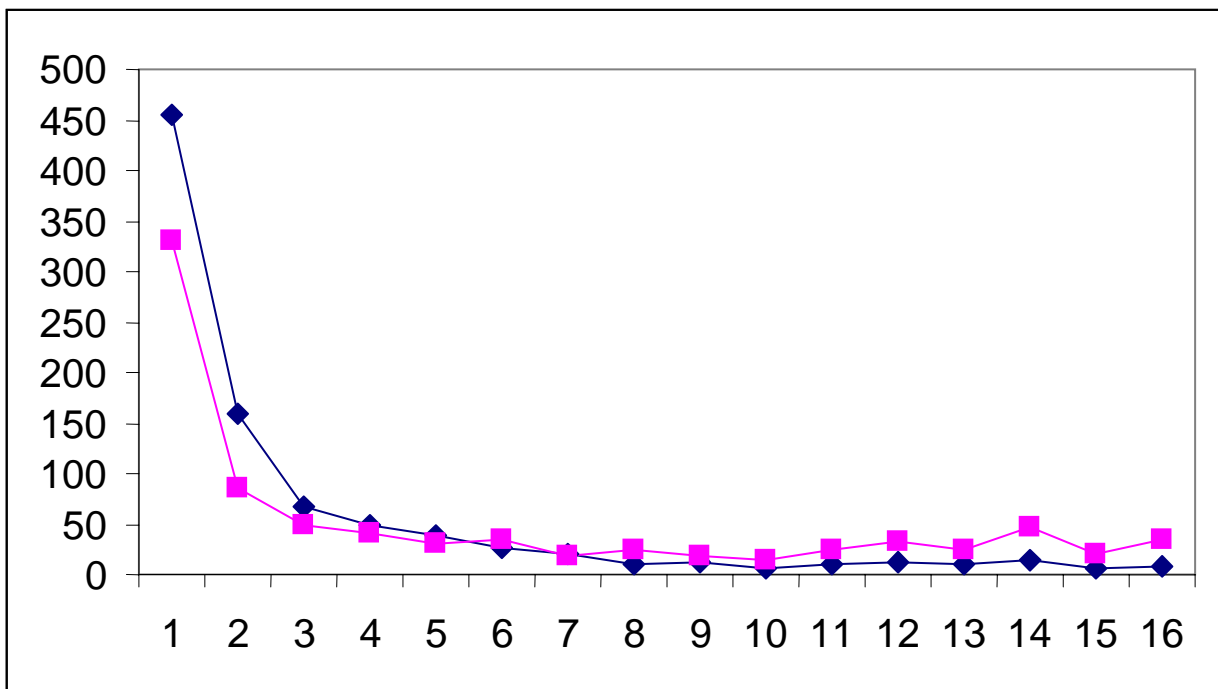


Рис. 5.6. Тюркские (темная линия) и огузские языки

Теперь сопоставим группы языков генетически и ареально однородных, но отличающихся численностью: огузские (16 языков) и тюркские (52) языки. Для первых $S = 488$, для вторых – $S = 590$, значение S возросло на 20%. Возрастание S здесь можно объяснить таким образом. С увеличением числа языков возрастает величина S , т. к. это общее число новых признаков, генерируемых всеми языками. В принципе, воз-

растет и α_2 , так как с увеличением числа языков увеличиваются и возможности для заимствований. Однако α_2 растет медленнее, т. к. количество заимствований из какого-то языка зависит от числа языков, с которыми он контактировал, которое, естественно, меньше общего числа языков в группе.

Наконец, сравним генетически и географически разнородные группы, отличающиеся числом языков. Возьмем снова тестовую группу (52 языка, $S = 838$) и все языки базы данных (315 языков, $S = 2044$). Увеличение числа языков в 6 раз приводит к увеличению S в 2,4 раза. В целом, увеличение S объясняется теми же факторами, что и в предыдущем случае. Однако в связи с большой сложностью изучаемых явлений и большим числом факторов, влияющих на параметры C и α_2 , точное теоретическое предсказание величины S для различных групп языков не представляется возможным.

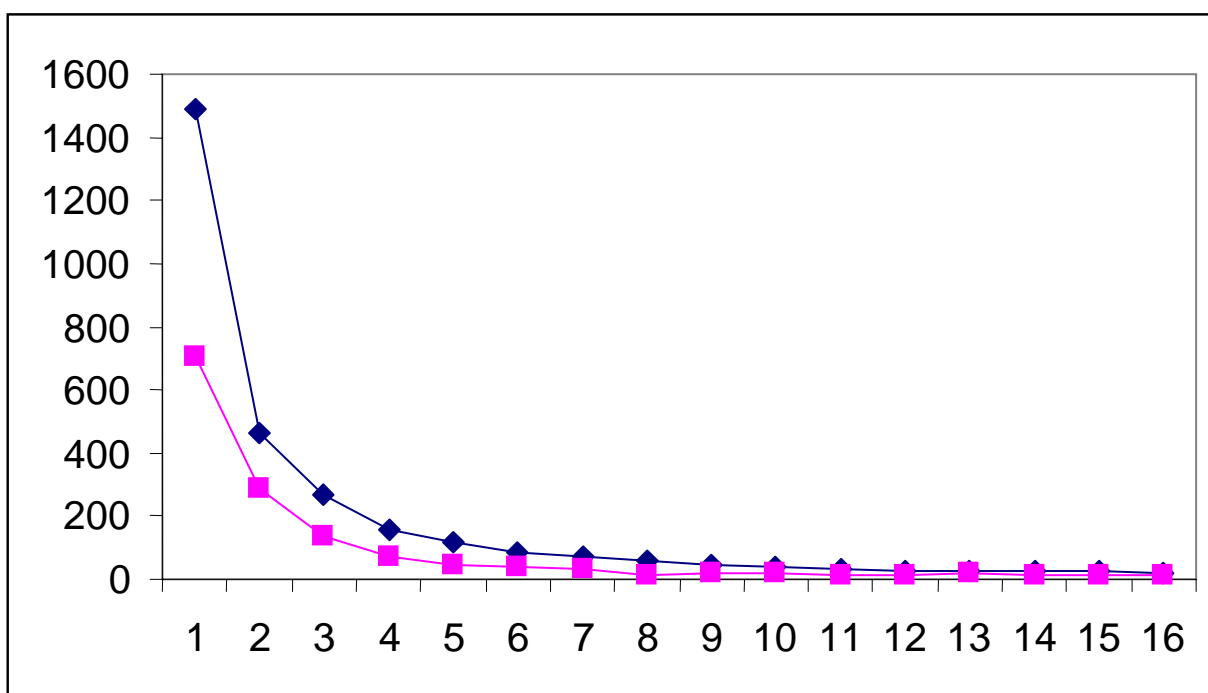


Рис. 5.7. Все языки (темная линия) и тестовая группа

Приведенные объяснения величины S в ряде случаев достаточно прозрачны. Рассмотрим теперь более тонкий случай с нетривиальными предсказаниями относительно иных свойств языков. Сопоставим огузские и кыпчакские языки (15 языков, тюркская семья).

Огузские и кыпчакские языки очень близки друг к другу. Тем не менее, их ЯП-диаграммы имеют заметные отличия. Огузские языки имеют значительно большее количество 1-языковых признаков, кыпчакские имеют больше высокочастотных признаков. Отличаются они и параметром S : $S = 488$ для огузских языков и только 262 для кыпчакских. Столь большая разница в значениях S , вероятно, должна означать, что C больше у огузских языков, а α_2 – у кыпчакских.

Действительно, большее число 1-языковых признаков является свидетельством большей активности языков в плане генерации новых признаков либо заимствования их из языков вне рассматриваемой группы (внешних заимствований). Это означает большие значения параметра C , а, значит, и S . Итак, имеющиеся данные позволяют

предположить, что для огузских языков, по сравнению с кыпчакскими, более характерны внешние заимствования, а для кыпчакских – внутренние (в пределах своей языковой группы). Последнее заставляет предположить, что кыпчакские языки занимают меньший ареал.

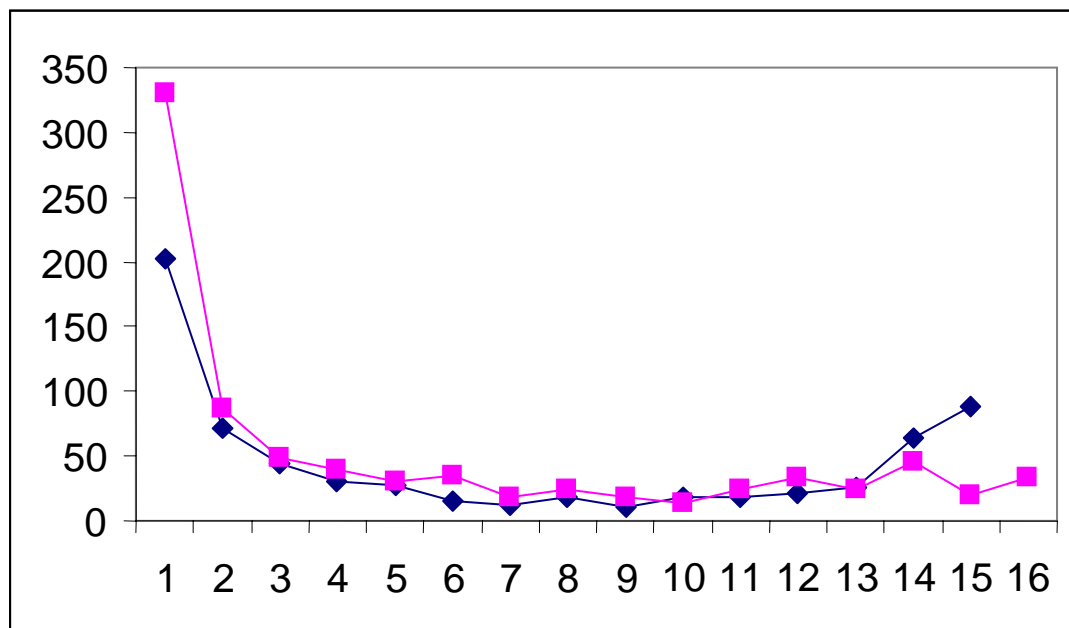


Рис. 5.8. Огузские (светлая линия) и кыпчакские языки

Данные выводы подтверждаются результатами других исследований. Так, в (Тюркские 1997, с.22) утверждается: «Факты показывают, что степень смешения огузизмов и кыпчакизов в отстаивающейся кыпчакской общности проявляется меньше (в смысле контрастности черт), чем в огузской. Последнее объясняется отчасти географическими условиями: отстаивание кыпчакских языков происходило на относительно узкой территории».

5.5. Резюме по главе 5

1. Собранные в БД значения языковых признаков могут анализироваться многими различными способами.
2. Анализ распределения признаков, существующих ровно в двух языках, позволил выявить определяющий характер заимствований для эволюции признаков.
3. Предложенный метод позволяет изучать контакты между языковыми группами, имевшие место в ходе эволюции, с помощью строгих математических методов.
4. Введены ЯП-диаграммы, характеризующие зависимость число признаков от степени их распространенности. Это совершенно новый интересный лингвистический объект исследований, появившийся только после создания БД.
5. Построена математическая модель распространения признаков, объясняющая внешний вид ЯП-диаграмм. Предложенная модель подтверждает недавно высказанную Далем (Dahl Ö. 2004) гипотезу о конкуренции молодых признаков.
6. Модель позволяет анализировать динамику признаков и в пределах отдельных языковых групп. Полученные результаты хорошо коррелируют с ранее полученными классическими методами.

Глава 6. РАЗРАБОТКА НОВЫХ МЕТОДОВ ВЕРИФИКАЦИИ ГЕНЕТИЧЕСКИХ ГИПОТЕЗ

6.1. Расчет квантитативных портретов по языковым семьям и ветвям

Анализ результатов расчетов мер парной близости позволяет делать вывод о том, что мера, рассчитанная с учетом прямого вклада всех признаков, содержит в себе генетически значимую информацию, о чем свидетельствует хорошее (хотя и не полное) совпадение результатов расчетов с традиционными представлениями о родстве языков в эталонной группе. В то же время существенные расхождения с традиционной картиной, особенно если использовать расчеты только по признакам, присутствующим в языке, с такой же убедительностью свидетельствуют о наличии типологического фона, природа которого вызывает большой интерес. Это явление подтверждается и квантитативными исследованиями, проведенными Г.Г. Сильницким (Сильницкий, 2004). Таким образом, все меры близости, рассчитанные по БД «Языки мира», несут в себе информацию двоякого рода: информацию о генетическом родстве двух языков и информацию об их типологической близости³⁵. Это заставило сосредоточить дальнейшие усилия исследовательской группы на детальном анализе признакового пространства в описании языков. Можно сказать, что в разгадке природы появления и динамики грамматических признаков в языках лежит ключ к созданию генетически представительных мер близости. Для исследования частотных закономерностей в распределении и динамике грамматических признаков были рассчитаны квантитативные портреты языковых семей по всему генетическому указателю БД «Языки мира». Методика расчета квантитативных портретов была уже описана ранее в разделе 3.2.

6.2. Классификация признаков

Для расчета и исследования квантитативных портретов была написана программа LangFamilies на языке VBA, которая рассчитывает частоту признаков по всем элементам генетического указателя БД «Языки мира» (макро-семьи, семьи, ветви, группы) и по всей БД в целом³⁶. После этого появляется возможность проведения сопоставительного анализа распределения признаков по разделам реферата и по семьям (ветвям).

В работе (Ярославцева 2005) введены понятия для следующих категорий признаков: *универсалии*, *раритарии*, *фреквенталии*, *уникалии*. Согласно этой работе, универсалиями считаются признаки, у которых общая частота по БД $F \Rightarrow 0,75$. Раритариями в (Ярославцева 2005) считаются признаки, у которых общая частота по БД $F \leq 0,25$. Уникалии, это признаки, которые встречаются только в одном языке. Фреквенталии – все остальные.

С другой стороны, строго говоря, универсалиями принято считать грамматические категории, присущие всем без исключения языкам. Кроме того, признак нельзя считать универсальным, если он распространен только в одной языковой семье (на-

³⁵ Существует точка зрения о том, что сильные расхождения в расчетах с экспертной информацией по эталонной группе объясняются наличием ошибок и систематических просчетов в описании языков, присущих текущей версии БД. Отметим, что этот вопрос заслуживает самого серьезного изучения. В настоящее время в ИЯ РАН проводится экспертиза БД, по результатам которой планируется внести необходимые коррективы и провести сравнительный вычислительный эксперимент. Вместе с тем, если на ошибки можно списать сравнительно низкий уровень показателя качества меры, то вторая проблема – близкое расстояние между неродственными языками в ранге, носит более фундаментальный характер, что будет продемонстрировано в главе 7.

³⁶ Автор программы В.Н. Поляков, генетический указатель составлен Е.И. Ярославцевой и актуализирован Ю.Б. Коряковым и В.Д. Соловьевым.

пример в ИЕ-семье), пусть и в достаточно большом количестве. Поэтому, будем использовать термин: *статистическая универсалия*. Более точным определением для статистической универсалии будет такое: признак должен иметь частоту $F \geq 0,5$ не менее чем в двух языковых семьях. Кроме того, для целей квантитативных исследований оказалось удобным разбить все признаки на 6 групп (см. табл. 6.1). Таблица составлена на примере индоевропейских языков.

Т а б л и ц а 6.1

Классификация признаков

Группа признаков		Общее число языков с указанным признаком по БД	Частота в ИЕ-семье	Частота в ветви (группе, подгруппе)	Частота в других семьях (ДС)
I группа	Абсолютные универсалии	$N_{бд} = 100\%$ от числа языков	-	-	-
II группа ³⁷	Статистические универсалии	-	$F_{ие} \geq 0,5$	-	$F_{дс} \geq 0,5$
III группа	Генетически устойчивые признаки для семьи	$N_{бд} \leq 2 * N_{ие}$	$F_{ие} \geq 0,5$	-	$F_{дс} \ll 0,5$
IV группа	Генетически устойчивые признаки для ветви	$N_{бд} \leq 2 * N_{в}$	$F_{ие} \leq 0,5$	$F_{в} \geq 0,5$	$F_{дс} \ll 0,5$
V группа	Генетически устойчивые признаки для части языков в ветви (группе, подгруппе)	$N_{бд} \leq 2 * N_{в}$	$F_{ие} \ll 0,5$	$F_{в} \ll 0,5$	$F_{дс} \ll F_{в}$
VI группа	Генетически неустойчивые признаки	-	$F_{ие} \ll 0,5$	$F_{в} \ll 0,5$	$F_{дс} \sim F_{в}$
VII группа	Уникалии	$N_{бд} = 1$	-	-	-

Как можно видеть, эта классификация основана не на конкретных значениях частоты распространения признака в БД, а на соотношении частоты его распространения в языковой семье (ветви) и во всем евразийском пространстве. Признаки в группах III и IV по традиционной классификации являются фреквенталиями. Группу V и VI чаще всего составляют раритарии. Очевидно, что признаки, входящие в группы III, IV, V и VII являются хорошими маркерами генетического родства.

³⁷ Коэффициент 2 выбран из соображений, что генетически устойчивые признаки могут заимствоваться другими языками, однако в целом не представляют в них типологической доминанты.

6.3. Пример расчета квантитативного портрета по ИЕ-языкам по разделу «Сложное предложение»

В таблице Е.1 (Приложение Е) приведен пример квантитативной карты ИЕ-семьи для раздела «2.5.4. Сложное предложение». В таблице 6.2 приведен обзор раритарных признаков (группы V и VI).

БД «Языки мира» позволила сделать еще одно важное наблюдение относительно квази-универсальных признаков. Было бы интересно провести наложение квантитативной карты на один или несколько древних языков ИЕ-семьи. В качестве такого языка был выбран язык Авесты, древнего религиозного письменного памятника предположительно 6 века до н.э. Результат сравнения представлен в таблице 6.2 и выделен курсивом. Мы видим, что большинство квази-универсалий в разделе «Сложное предложение» было представлено в Авесте. Таким образом, можно предположить, что эти общие черты были заимствованы большинством языков еще из древних письменных источников (или тогда еще живых языков). Подтверждением этому также служит хорошее совпадение признаков хеттского (столбец *анатолийские*, несовпадение выделено подчеркиванием), который имеет еще более древние письменные памятники, и языка Авесты.

Т а б л и ц а 6.2

Признаки групп IV и V

Номер в БД	Наименование	Языки	Группа
3776	главное предшествует придаточному	ГАГАУЗСКИЙ, ИТАЛЬЯНСКИЙ, ЛЮКСЕМБУРГСКИЙ, ПЕРСИДСКИЙ, СТАРОФРАНЦУЗСКИЙ, ТАДЖИКСКИЙ, УЗБЕКСКИЙ (*)	V
3782	особое оформление им. групп	ЗЕМИАКИ, КАРТАКАЛАИ, КЕТСКИЙ (*), САЛАРСКИЙ (*), СТАРОЯПОНСКИЙ (*)	VI
3783	особое оформление подлежащего	АРАБСКИЙ СОВРЕМЕННЫЙ (*), ИСЛАНДСКИЙ, КАЛМЫЦКИЙ (*), ЛАТИНСКИЙ, МАНЬЧЖУРСКИЙ (*), СТАРОПИСЬМЕННЫЙ МОНГОЛЬСКИЙ (*), ЮРТОВСКИХ ТАТАР (*)	VI
3784	особый порядок слов	АФРИКААНС, ДАТСКИЙ, НЕМЕЦКИЙ, НОРВЕЖСКИЙ, ФРИЗСКИЙ, ШВЕДСКИЙ, ЭСТОНСКИЙ (*)	V
3787	квотатив	АЙНСКИЙ (*), АНДИЙСКИЙ (*), КАРТАКАЛАИ, ОРХОНО-ЕНИСЕЙСКИХ НАДПИСЕЙ (*), ХВАРШИНСКИЙ (*), ЭЛАМСКИЙ (*)	VI
3792	таксисные конструкции	АЛАБУГАТСКИХ ТАТАР (*), АНГЛИЙСКИЙ, АСТРАХАНСКИХ НОГАЙЦЕВ-КАРАГАШЕЙ (*), БЕЛУДЖСКИЙ, ИСПАНСКИЙ, КАЗАХСКИЙ (*), КАРАЧАЕВО-БАЛКАРСКИЙ (*), ЛАТИНСКИЙ, НЕМЕЦКИЙ, НИДЕРЛАНДСКИЙ, НОРВЕЖСКИЙ, ПОРТУГАЛЬСКИЙ, ПРОВАНСАЛЬСКИЙ, САРДИНСКИЙ, ФИНСКИЙ (*), ФРАНЦУЗСКИЙ, ФРИЗСКИЙ, ЮРТОВСКИХ ТАТАР (*)	VI
3794	абсолютные обороты	ГОТСКИЙ, ПОРТУГАЛЬСКИЙ	V
3800	главное пред-	АВЕСТЫ, ИТАЛЬЯНСКИЙ, МАКЕДОН-	V

Номер в БД	Наименование	Языки	Группа
	шествует при- даточному	СКИЙ, НЕМЕЦКИЙ, ХУРРИТСКИЙ (*)	
3803	придаточное предшествует главному	ГУДЖАРАТИ, ХИНДИ	V
3805	союзное слово в придаточном	ГУДЖАРАТИ, МАКЕДОНСКИЙ, НЕМЕЦКИЙ, НОВОАРАМЕЙСКИЕ ДИАЛЕКТЫ, ОСК- СКИЙ, СИНГАЛЬСКИЙ, УРАТСКИЙ (*), ФИНСКИЙ(*), ХИНДИ, ХУРРИТСКИЙ (*)	VI
3810	только подчине- ние	ГЛАНГАЛИ, ЗЕМИАКИ, ИТЕЛЬМЕНСКИЙ (*), ОСКСКИЙ, СИНГАЛЬСКИЙ, СИРЕНИК- СКИХ ЭСКИМОСОВ (*), СТАРОЯПОНСКИЙ (*)	VI
3811	только сочине- ние	ГАЛЛЬСКИЙ, МОНГОРСКИЙ (*), ФАЛИСК- СКИЙ	VI

Примечание: знаком (*) отмечены языки, не относящиеся к ИЕ-семье.

Среди множества характеристик сложного предложения удалось выявить два генетически устойчивых признака, которые характерны для двух разных ветвей ИЕ-семьи (см. табл. E.1 в Приложении E). Это признак *..преобладает союзная* (3815), который является характерным генетическим указателем для кельтской ветви языков (частота 0,67) и признак *...знаменательные слова* (3817), характерный для большинства языков индоарийской ветви (частота 0,56). Семейных раритарных признаков для ИЕ-семьи в указанном разделе выявлено не было.

6.4. Автоматическое выявление признаков

Расчет количественных портретов можно проводить не только для семей, но и для языковых сообществ всех уровней: макро-семей, семей, ветвей, групп и подгрупп языков. Принимая во внимание, существенные объемы БД и разветвленность генетического указателя, выявление таких маркеров вручную представляется весьма трудоемкой задачей. Для облегчения ее выполнения в программе LangFam была реализована функция автоматического выявления генетически значимых признаков. Список этих признаков после тщательного анализа планируется опубликовать на сайте.

Расчет количественных портретов языковых сообществ представляет собой оригинальную методику, которая дает в руки исследователя новый инструментарий для проверки генетических гипотез. Основным достоинством этой методики является меньшая зависимость от случайных ошибок в БД. Кроме того, количественные портреты позволяют получить новые данные о динамике и заимствовании структурных грамматических признаков, что в свою очередь проливает свет на природу раритарий и квази-универсалий и дает новые факты для построения теории структурной эволюции языков.

6.5. Методика проверки генетических гипотез

Выявленные закономерности эволюции языкового континуума на территории Евразии и количественные портреты по семьям в БД позволили предложить новую методику верификации генетических гипотез. Методика основывается на использовании раритарных признаков. Идея использования раритариев для проверки спорных генетических гипотез возникла из предположения о том, что раритарные признаки могут нести более достоверную информацию о генетической близости, чем универсальные. Эта точка зрения убедительно подтверждается результатами расчетов количественных карт (раздел 6.3) и полученными закономерностями в динамике признаков (раздел 7.2).

Метод основывается на подборе таких фильтров на количественных портретах языковых семей, чтобы можно было выявить совпадение раритарных признаков в анализируемых языковых семьях, ветвях, языках-изолятах. Хорошее совпадение свидетельствует о генетической близости.

Приведем примеры.

В германской и романской ветвях ИЕ-семьи совпадают шесть признаков, из которых 4 являются доминирующими для каждой из ветвей (таблица 6.3). Все признаки являются фактографическими. Это хорошо подтверждается альтернативными данными о близости этих ветвей. Список генетически значимых признаков был получен путем наложения на контент БД двух фильтров:

1) частота (в абсолютном выражении) $N \leq 50$ (т.е. не более чем в два раза от числа романских языков в БД);

2) частота в относительном измерении в романской ветви $F \geq 0,5$.

Также из таблицы мы видим, что других кандидатов на родство с таким набором признаков в ИЕ-семье нет.

Алтайская макросемья. В БД имеется 14 общих признаков среди семей: монгольской, тунгусо-манчжурской, тюркской, среди которых 10 являются фактографическими (см. табл. 6.4). В качестве базового фильтра были установлены следующие параметры: частота в алтайской макросемье больше или равна 0,5; общая частота распространения признака в БД не должна превышать 100 (в абсолютных единицах, всего в БД 76 языков из алтайской макросемьи). Это является очень сильным аргументом в пользу существования такой макросемьи.

В то же время гипотеза о принадлежности японского (и старояпонского) языка к алтайской группе по данным количественных расчетов представляется малоубедительной. Из четырнадцати общих признаков для алтайской макросемьи старояпонский разделяет только 1.

Кстати, хочется обратить внимание на то, что большинство общих генетически значимых признаков алтайских языков по своей частотной природе относятся к фреквенталиям, которые попадают в 25 % наиболее частотных признаков в БД. Рассмотрим, какие из языковых семей, представленных в БД «Языки Мира» тяготеют к алтайским (табл. 6.5).

Таблица 6.3

Генетически значимые раритари и фреквенталии, совпадающие в германской и романской ветвях ИЕ-семьи

Частота (N ≤ 50)	№ в БД, признак, иерархия	индоевропейские (133)	анаголийские (1)	армянские (1)	германские (14)	дардские (13)	индоарийские (9)	иранские (38)	италийские (4)	кельтские (9)	нуристанские (5)	романские (25) (F >= 0,5)	славянские (14)	дамели (1)
25	(1171)...в артикле <=синтаксический <=способ выражения согласоват. классов <=согласовательные классы <=2.3.2.ИМЕННЫЕ КЛАССИФИКАЦИИ	0,18	0	0	0,50	0	0,11	0,03	0	0,11	0	0,56	0	0
50	(1399)имен и местоимений <=субъекта <=различное падежное оформление <=актантные падежные значения <=2.3.4.ПАДЕЖНЫЕ ЗНАЧЕНИЯ	0,28	0	0	0,14	0,38	0,11	0,26	0	0,11	0,2	0,56	0,21	1
50	(2744) ...свободные <=артикли <=определенность/неопределенность имени <=2.3.6.ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ	0,35	0	0	0,79	0,23	0	0,21	0	0,11	0	0,96	0	0
39	(2838)...герундий <=нефинитные формы <=глагол <=знаменательные <=2.3.7.ЧАСТИ РЕЧИ	0,23	0	0	0,07	0	0,22	0,02	0,25	0	0,4	0,96	0	0
34	(2914) ..род <=артикуль <=2.4.0.ПАРАДИГМЫ	0,25	0	0	0,64	0	0,11	0,00	0	0	0	0,92	0	0
34	(2916) ..число <=артикуль <=2.4.0.ПАРАДИГМЫ	0,25	0	0	0,57	0	0	0,05	0	0,11	0	0,88	0	0

Генетически значимые признаки для алтайской макросемьи

Частота (N >=100)	Признак, иерархия, № в БД	Факт./Клас.	алтайские (76) (F>=0,5)	монгольские (11)	тунгусо-маньч- журские (10)	тюркские (54)	японо- рюкюские (1)
99	...передний/задний <=ряд <=2.1.1.ФОНЕМНЫЙ СОСТАВ (17)	Факт.	0,5	0,45	0	0,61	0
76передний/задний <=в нижнем подъеме <=различия по ряду <=общая схема монофтонгов <=2.1.1.ФОНЕМНЫЙ СОСТАВ (48)	Факт.	0,5	0,36	0	0,63	0
87губные/переднеязычные/заднеязычные <=назальные <=локальный ряд <=сонорные <=шумные <=согласные <=2.1.1.ФОНЕМНЫЙ СОСТАВ (342)	Факт.	0,51	0,82	0,1	0,52	1
98	..область гармонического уподобления <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (523)	Клас.	0,84	0,64	1	0,87	0
77	...слово <=область гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (527)	Факт.	0,71	0,55	0,7	0,76	0
98	..признак гармонического уподобления <=область гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (529)	Клас.	0,84	0,64	0,8	0,91	0
74	...ряд <=признак гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (531)	Факт.	0,62	0,64	0,2	0,7	0
55	...огубленность <=признак гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (536)	Факт.	0,59	0,27	0,5	0,69	0
79	..уподобляющий элемент <=признак гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (538)	Клас.	0,68	0,64	1	0,65	0
47	...первый гласный слова <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (539)	Факт.	0,5	0,55	0,5	0,5	0

Частота (N >= 100)	Признак, иерархия, № в БД	Факт./Клас.	алтайские (76) (F >= 0,5)	монгольские (11)	тунгусо-маньч- журские (10)	тюркские (54)	японо- рюкюские (1)
60	...гармония <=комбинаторные позиции <=гласные <=2.1.3.ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ (613)	Факт.	0,57	0,36	1	0,54	0
73	...элатив <=локативные формы имени <=падежное оформление локативных отношений <=2.3.4.ПАДЕЖНЫЕ ЗНАЧЕНИЯ (1673)	Факт.	0,59	0,91	0,7	0,52	0
92	...рефлексив <=категориальный состав кроме активного <=залог <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (1784)	Факт.	0,58	0	0,3	0,76	0
73	...реципрок <=категориальный состав кроме активного <=залог <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (1790)	Факт.	0,62	0,73	1	0,54	0

Кроме перечисленных закономерностей были отмечены следующие.

Дардские и нуристанские имеют три общих раритарных признака, которые подчеркивают их особенную генетическую близость по сравнению с остальными ветвями ИЕ-семьи (табл. 6.6). Выбраны признаки, у которых $F \geq 0,5$ для дардских языков.

Шумерский язык. Шумерский считается языком-изолятом. С помощью количественных расчетов было обнаружено, что шумерский имеет 6 генетических признаков (3 фактографических и 3 классифицирующих) из тех 11, что являются доминирующими для семитской ветви афразийской семьи языков (табл. 6.7).

Ни с какой из других семей шумерский язык такого количества общих генетически значимых признаков не имеет.

Нивхский язык. Также как и шумерский, нивхский считается языком-изолятом. Нивхский имеет 5 раритарных признаков из 25 генетических маркеров для чукотско-камчатской ветви. Соотношение 5 к 25 свидетельствует все же о весьма дальнем родстве этих языков.

Даже с учетом того, что данные, представленные в таблицах 6.3-6.8, отражают предварительные результаты, которые можно рассматривать пока как гипотезы, они заслуживают самого серьезного внимания со стороны специалистов, типологов и компаративистов. На наш взгляд, БД «Языки мира» имеет большие перспективы в части верификации генетических гипотез.

Таблица 6.5

Генетические признаки алтайской макросемьи на фоне остальных макросемей, представленных в БД «Языки мира»

Признак, иерархия, № в БД	австронезий- ские (1)	алтайские (76)	аустроазиат- ские (2)	афразийские (9)	индоевропей- ские (133)	кавказские (38)	палеоазиат- ские (18)	сино- тибетские (2)	уральские (22)	хуррито- урартские (2)
...передний/задний <=ряд <=2.1.1.ФОНЕМНЫЙ СОСТАВ (17)	0	0,5	0	0,11	0,22	0,37	0,17	0,5	0,45	1
....передний/задний <=в нижнем подъеме <=различия по ряду <=общая схема монофтонгов <=2.1.1.ФОНЕМНЫЙ СОСТАВ (48)	0	0,5	0	0	0,14	0,21	0,06	0	0,45	0
.....губные/переднеязычные/заднеязычные <=назальные <=локальный ряд <=сонорные <=шумные <=согласные <=2.1.1.ФОНЕМНЫЙ СОСТАВ (342)	0	0,51	0	0,11	0,21	0,03	0,5	0,5	0,23	0
..область гармонического уподобления <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (523)	0	0,84	0	0,22	0,06	0,18	0,17	0	0,55	0
...слово <=область гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (527)	0	0,71	0,5	0,22	0,02	0,11	0,11	0	0,41	0
..признак гармонического уподобления <=область гармонического уподобления <=сингармонизм	0	0,84	0,5	0,22	0,06	0,16	0,17	0	0,55	0

<=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (529)										
...ряд <=признак гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (531)	0	0,62	0	0	0,05	0,16	0,06	0	0,55	0
...огубленность <=признак гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (536)	0	0,59	0	0	0,02	0,08	0	0	0,14	0
..уподобляющий элемент <=признак гармонического уподобления <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (538)	0	0,68	0,5	0,11	0,05	0,18	0	0	0,5	0
...первый гласный слова <=сингармонизм <=2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ (539)	0	0,5	0	0	0,01	0,05	0	0	0,27	0
...гармония <=комбинаторные позиции <=гласные <=2.1.3.ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ (613)	0	0,57	0,5	0,11	0,02	0,11	0,06	0	0,27	0
...элатив <=локативные формы имени <=падежное оформление локативных отношений <=2.3.4.ПАДЕЖНЫЕ ЗНАЧЕНИЯ (1673)	0	0,59	0	0	0	0,34	0,06	0	0,59	0
...рефлексив <=категориальный состав кроме активного <=залог <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (1784)	0	0,58	0,5	0,44	0,18	0,05	0,22	0	0,5	0
...реципрок <=категориальный состав кроме активного <=залог <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (1790)	0	0,62	0,5	0,22	0,11	0,08	0,11	0	0,09	0

Общие генетические признаки дардской и нуристанской ветвей

Частота (N <=22)	Признак, иерархия, № в БД	Факт./ Клас.	индоевропей- ские (133)	анатолийские (1)	армянские (1)	германские (14)	дардские (13)	индоарийские (9)	иранские (38)	италийские (4)	кельтские (9)	нуристанские (5)	романские (25)	славянские (14)	дамели (1)
14переднеязычные <=дополнительные признаки <=шумные <=согласные <=2.1.1.ФОНЕМНЫЙ СОСТАВ (315)	Ф	0,11	0	0	0	0,69	0	0,08	0	0	0,4	0	0	0
22церебральные/ нецеребральные <=фарингализованные/ нефарингализованные <=дополнительные признаки <=сонорные <=шумные <=согласные <=2.1.1.ФОНЕМНЫЙ СОСТАВ (419)	К	0,14	0	0	0	0,77	0	0,13	0	0	0,8	0	0	1
17	..вигезимальная <=двадцатеричная <=система счисления <=2.3.3.ЧИСЛО (1305)	Ф	0,1	0	0	0	0,69	0	0	0	0,11	0,6	0	0	1

Общие генетически значимые признаки в семитских языках и в шумерском

Частота (N <=14)	Признак, иерархия, № в БД	Факт./Клас.	австронезийские (1)	алтайские (76)	аустроазиатские (2)	афразийские (9)	семитские (7) (F >= 0,5)	шумерский
6	...эмфатизованные/неэмфатизованные <=церебральные/нецеребральные <=дополнительные признаки <=шумные <=согласные <=2.1.1.ФОНЕМНЫЙ СОСТАВ (320)	К	0	0	0	0,56	0,57	0
11	.порода <=способ выражения <=версия <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2480)	К	0	0	0	1	1	1
11	..основные значения <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2481)	К	0	0	0	1	1	1
7	...интенсив <=основные значения <=порода <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2486)	Ф	0	0	0	0,56	0,71	1
8	...каузатив <=основные значения <=порода <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2487)	Ф	0	0	0	0,89	0,86	0
7	...рефлексив <=основные значения <=порода <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2496)	Ф	0	0	0	0,67	0,71	0
10	..способ выражения <=основные значения <=порода <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2501)	К	0	0	0	0,89	1	1
9	...аффикс <=порода	Ф	0	0	0	0,89	1	0

Частота (N ≤14)	Признак, иерархия, № в БД	Факт./Клас.	австронезийские (1)	алтайские (76)	аустроазиатские (2)	афразийские (9)	семитские (7) (F >= 0,5)	шумерский
	<=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2502)							
9	...операция над основой <=способ выражения <=порода <=2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ (2503)	Ф	0	0	0	0,78	1	1
8порода <=словоизменение <=финитные формы <=глагол <=2.4.0.ПАРАДИГМЫ (2942)	Ф	0	0	0	0,67	0,71	1
8	...состояние <=словоизменение <=существительное <=2.4.0.ПАРАДИГМЫ (3359)	Ф	0	0	0	0,67	0,86	0

Т а б л и ц а 6.8

Нивхский и чукотско-камчатские: общие генетические маркеры

Частота (N ≤ 8)	Признак, иерархия, № в БД		камчатские (4) (F >= 0,5)	НИВХСКИЙ
2шипящие <=фрикативные и аффрикаты <=переднеязычные <=локальный ряд <=шумные <=согласные <=2.1.1.ФОНЕМНЫЙ СОСТАВ (214)	Ф	0,5	0
8	...имплозивные/эксплозивные <=дополнительные признаки <=шумные <=согласные <=2.1.1.ФОНЕМНЫЙ СОСТАВ (247)	К	0,5	0
6	...геминация	Ф	0,5	0

	<=структурные позиции <=согласные <=2.1.3.ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ (650)			
7удвоение <=стыковые явления <=ограничения на фонемную структуру слова <=фонологическая структура слова <=2.2.1.ФОНОЛОГИЧЕСКАЯ СТРУКТУРА (1021)	Ф	0,5	0
7	..инкорпорация <=синтетические языки с элементами аналитизма <=степень спаянности морфем <=2.3.0.МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА (1135)	К	0,75	1
7	..языки с элементами инкорпорации <=инкорпорация <=2.3.0.МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА (1137)	К	0,75	1
8	...абсолютив <=актантные падежные значения <=2.3.4.ПАДЕЖНЫЕ ЗНАЧЕНИЯ (1355)	Ф	0,5	1
3	..относительная частица <=падежное оформление посессивного отношения <=2.3.4.ПАДЕЖНЫЕ ЗНАЧЕНИЯ (1540)	Ф	0,5	0
2	...ориентатив <=локативные формы имени <=падежное оформление локативных отношений <=2.3.4.ПАДЕЖНЫЕ ЗНАЧЕНИЯ (1649)	Ф	0,5	0
7	..предикатив (прилагательное + глагол) <=знаменательные <=2.3.7.ЧАСТИ РЕЧИ (2870)	Ф	0,5	0
3монопersonальное/полиперсональное <=типы спряжения <=финитные формы <=глагол <=2.4.0.ПАРАДИГМЫ (2959)	Ф	0,5	0
6	...принадлежность <=словоизменение <=имя <=2.4.0.ПАРАДИГМЫ (3196)	Ф	0,5	1
8	...число <=словоизменение <=имя <=2.4.0.ПАРАДИГМЫ (3200)	Ф	0,5	1
3	..классифицирующие категории классифицирующие категории <=словоизменение <=имя <=2.4.0.ПАРАДИГМЫ (3201)	К	0,5	0

3	...личность/неличность <=имя <=2.4.0.ПАРАДИГМЫ (3202)	Ф	0,5	0
3	.предикатив <=послелог <=2.4.0.ПАРАДИГМЫ (3258)	К	0,5	0
2	.."быть" <=2.4.0.ПАРАДИГМЫ (3259)	Ф	0,5	0
2	.."иметь" <=предикатив <=2.4.0.ПАРАДИГМЫ (3260)	Ф	0,5	0
2	..комитатив <=предикатив <=2.4.0.ПАРАДИГМЫ (3261)	Ф	0,5	0
2	..антикомитатив <=предикатив <=2.4.0.ПАРАДИГМЫ (3262)	Ф	0,5	0
2	..предикатив обладания <=предикатив <=2.4.0.ПАРАДИГМЫ (3263)	Ф	0,5	0
7	...циркумфиксы <=аффиксация <=деривация <=2.5.2.СЛОВООБРАЗОВАНИЕ (3606)	Ф	0,5	0
6	.инкорпорация <=коммуникативно значимые средства <=2.5.3.ПРОСТОЕ ПРЕДЛОЖЕНИЕ (3761)	К	0,75	0
5	..синтаксические отношения между компонентами <=инкорпорация <=2.5.3.ПРОСТОЕ ПРЕДЛОЖЕНИЕ (3763)	К	0,75	0
3	...инкорпорация атрибута <=инкорпорация <=2.5.3.ПРОСТОЕ ПРЕДЛОЖЕНИЕ (3764)	Ф	0,75	0

6.6. Резюме по главе 6

1. Предложена классификация признаков по частоте встречаемости, уточняющая ранее предлагавшиеся.
2. Описана методика применения введенного аппарата для автоматического выявления генетических маркеров – признаков, имеющих сильную корреляцию с родством языков.
3. Предложен новый метод установления общности происхождения языков, основанный на сопоставлении редких признаков.
4. Метод верифицирован проверкой его на группах языков с точно установленным родством (романские и германские языки, алтайская макросемья).
5. Метод применен к нескольким языкам-изолятам (шумерский, нивхский), полученные результаты подтверждают некоторые из ранее высказанных гипотез.

Глава 7. ФЕНОМЕН ТИПОЛОГИЧЕСКОГО СДВИГА

Квантитативные языковые портреты по семьям, ветвям и группам позволили выявить новые статистические закономерности в динамике грамматических признаков на евразийском пространстве. Исследованию этих не известных ранее явлений посвящена настоящая глава.

7.1. Природа языковых универсалий

Построение квантитативных языковых карт по разделам реферата БД позволило обнаружить весьма любопытную корреляцию между признаками древних мертвых языков и статистическими универсалиями, то есть наиболее частотными признаками в БД.

Эти закономерности было решено проверить на более представительном материале. БД «Языки мира» представляет в этом смысле широкие возможности, так как включает описание 54 мертвых языков. Были построены частотные графики по всем разделам реферата, на которые были наложены диаграммы, показывающие наличие признаков в некоторых мертвых языках. В итоге, путем поиска в БД и подбора соответствующих языков, удалось сформировать выборку из 11 древних языков (Авесты, латинский, древнееврейский, древнеперсидский, старописьменный монгольский, арабский классический – аналог древнеарабского, тюрки, урартский, половецкий, хеттский, древнегузский), которые покрыли все 100 % пиков на графиках частотного распределения признаков. На рис. 7.1 показан пример фрагмента диаграмм для раздела «2.3.6. Дейктические категории». На верхней диаграмме показаны частотные графики распределения признаков по БД³⁸, на нижней диаграмме столбцы, указывающие наличие признака в том или ином мертвом языке. Расшифровка первых 40 признаков из раздела 2.3.6. дана в таблице 7.1.

Т а б л и ц а 7.1

Расшифровка признаков для диаграммы на рис. 7.1

Номер на графике рис. 6	Порядковый номер в БД (версия от 26.03.06)	Номер раздела	Число языков	Наименование признака
1	2520	2.3.6.	316	2.3.6.ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ
2	2521	2.3.6.	12	.О
3	2522	2.3.6.	299	.местоимения и местоименные слова
4	2523	2.3.6.	296	..разряды местоимений
5	2524	2.3.6.	3	...адвербиальные
6	2525	2.3.6.	17	...анафорические
7	2526	2.3.6.	3	...атрибутивно-указательные
8	2527	2.3.6.	6	...безличные

³⁸ Диаграмма включает 6 графиков: частота признака по всей БД, частота в выборке из 52 мертвых языков (МЯ), частота в тестовой группе из 52 живых языков (ТГ), генетически эквивалентных мертвым, сдвиг частоты БД-МЯ, сдвиг частоты ТГ-МЯ.

Номер на графике рис. 6	Порядковый номер в БД (версия от 26.03.06)	Номер раздела	Число языков	Наименование признака
9	2528	2.3.6.	16	...взаимные
10	2529	2.3.6.	2	...взаимно-возвратные
11	2530	2.3.6.	0	...взаимно-личные
12	2531	2.3.6.	1	...взаимно-притяжательные
13	2532	2.3.6.	124	...возвратные
14	2533	2.3.6.	1	...инклюзивность/экслюзивность
15	2534	2.3.6.	1	...субъектно/объектные
16	2535	2.3.6.	1	...возвратно-выделительные
17	2536	2.3.6.	8	...возвратно-определяющие
18	2537	2.3.6.	1	...возвратно-относительно-неопределенные
19	2538	2.3.6.	1	...возвратно-отрицательные
20	2539	2.3.6.	13	...возвратно-притяжательные
21	2540	2.3.6.	2	...возвратно-указательные
22	2541	2.3.6.	3	...возвратно-усилительные
23	2542	2.3.6.	201	...вопросительные
24	2543	2.3.6.	30	...вопросительно-относительные
25	2544	2.3.6.	1	...выделительные
26	2545	2.3.6.	1	...глагольные
27	2546	2.3.6.	1	...качественные
28	2547	2.3.6.	2	...кванторные
29	2548	2.3.6.	1	...количественные
30	2549	2.3.6.	1	...косвенно-объектные
31	2550	2.3.6.	292	...личные
32	2551	2.3.6.	37	...инклюзивность/экслюзивность
33	2552	2.3.6.	2	...различия по полу во всех лицах
34	2553	2.3.6.	6	...субъектные/объектные
35	2554	2.3.6.	7	...лично-возвратные
36	2555	2.3.6.	0	...личные единичности
37	2556	2.3.6.	2	...лично-определенные
38	2557	2.3.6.	2	...лично-определяющие
39	2558	2.3.6.	5	...лично-притяжательные
40	2559	2.3.6.	1	...личные косвенного объекта

Ситуация, подобная рис. 7.1, наблюдается по всем без исключения разделам реферата в БД. Полный состав статистических универсалий приведен в приложении Ж.

Для дальнейшей проверки указанных закономерностей, были проведены частотные расчеты по 52 мертвым языкам в БД³⁹. Графики частотного распределения по мертвым языкам также представлены на рис. 7.1 (и в Приложении Ж). Кроме того, была составлена тестовая группа из 52 живых языков, генетически эквивалентных мертвым⁴⁰, по которым также были проведены расчеты частоты распределения признаков. В подавляющем большинстве случаев частотные пики группы из 52 мертвых языков совпали с частотными пиками из 52 живых языков и совпали с частотными всплесками по всей БД. Такая хорошая корреляция позволила выдвинуть гипотезу о том, что большинство языковых универсалий на евразийском пространстве сформировались достаточно давно, то есть до момента возникновения самых древних письменных источников. Кроме того, было обнаружено, что универсалии в тестовой группе живых языков в большинстве случаев (67 %) имеют большую частоту, чем универсалии в мертвых.

Еще один вопрос, который представляет интерес: «Как распределены универсалии по языковым семьям?» В таблице 7.2 представлены данные о частоте встречаемости наиболее частотных фактографических признаков в языковых семьях, представленных в БД. Данные получены при наложении фильтра на контент БД: $N \geq 158$, т.е. признак должен охватывать не менее 50 % языков. Всего таких признаков набралось 107.

Можно видеть, что статистические универсалии распределены практически равномерно по всем языковым семьям Евразии. Дальнейший анализ таблицы 7.2 все же позволяет выделить отклонения в распределении, которые могут указывать на возможный источник и возможные приемники тех или иных заимствований. Так, языковые семьи, у которых частота распространения признака меньше 0,5, вероятнее всего являются приемниками, то есть заимствующей стороной. Такие признаки выделены в таблице курсивом и подчеркиванием одновременно. Языковая семья с максимальным значением частоты распространения признака, скорее всего, является источником его распространения, то есть тем, у кого идет заимствование. Эти признаки выделены более крупным шрифтом. Нельзя исключать и спонтанного появления признаков одновременно в двух и более семьях. Об этом могут свидетельствовать общие высокие частоты распространения признака по всем семьям. Таких признаков также в таблице немало. Общие итоги по количеству заимствований в качестве источника/приемника для языковых семей приведены в конце таблицы 7.2. Если в строке оказывались два одинаковых максимальных значения, то такая строка в итогах не учитывалась, так как нет возможности определить источник заимствования. Эти данные можно интерпретировать по-разному: с точки зрения времени существования языковых семей, с позиций интенсивности контактов/изолированности, с позиций цивилизационного влияния этносов и т.д.

Наконец, из таблиц Ж.1-Ж.5 в приложении Ж можно сделать вывод о том, что примерно 68% от числа статистических универсалий составляют классифицирующие признаки, и только 32% – фактографические.

³⁹ Два языка (пиктский и саргулямский) в выборку включены не были из-за небольшого объема информации в реферате.

⁴⁰ Список языков тестовой группы приведен в разделе 5.4.

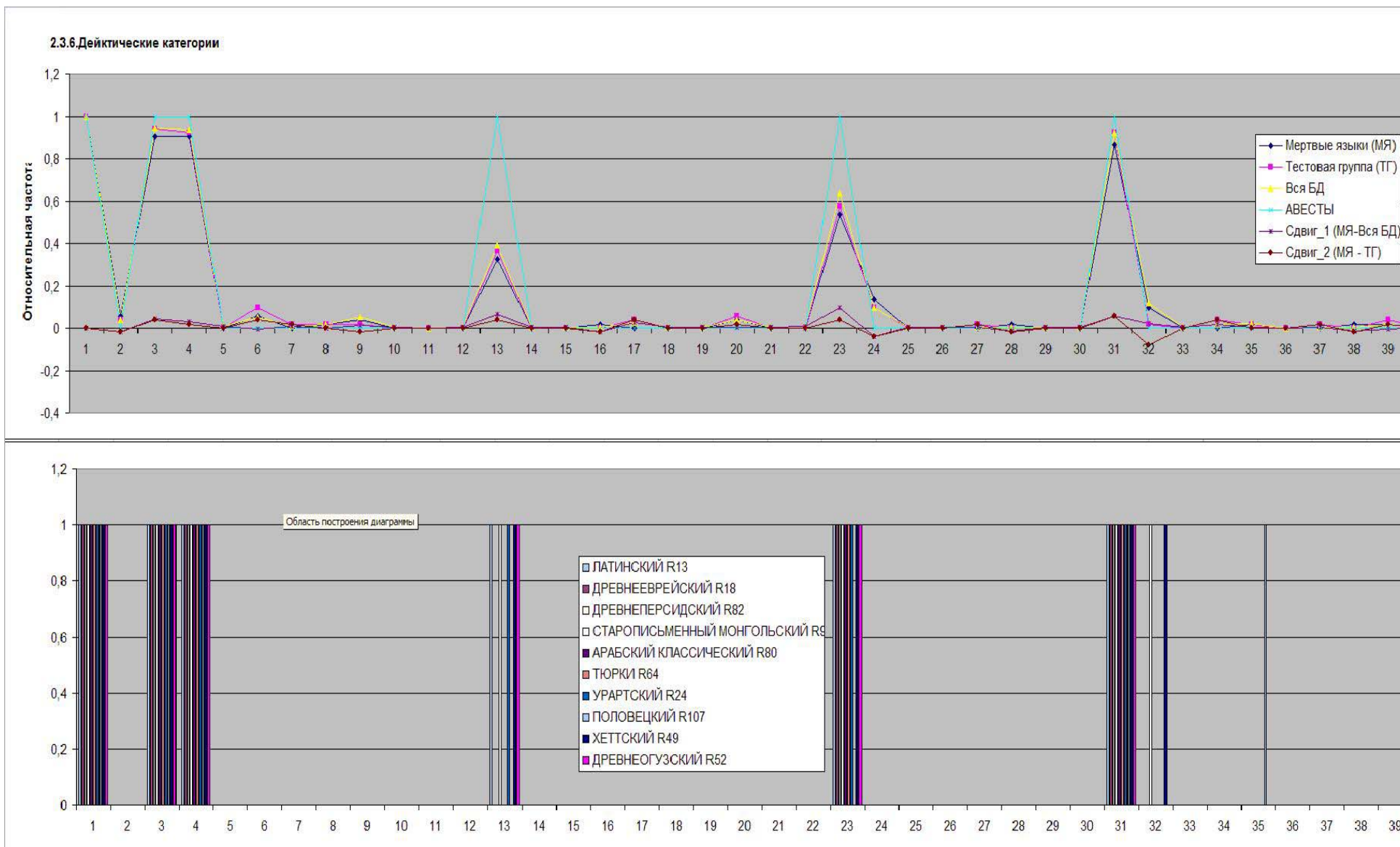


Рис. 7.1. Фрагмент диаграммы для раздела «2.3.6. Дейктические категории».

Таблица 7.2.

Данные о частоте встречаемости наиболее частотных фактографических признаков в языковых семьях

Порядковый номер в БД (версия от 26.03.06)	Номер раздела	Частота	Наименование признака	алтайские (76)	афразийские (9)	индоевропейские (133)	кавказские (38)	палеоазиатские (18)	уральские (22)
8	2.1.1.	231	...верхний/средний/нижний	0,63	0,67	0,74	0,92	0,89	0,86
19	2.1.1.	183	...передний/средний/задний	0,39	0,89	0,68	0,53	0,72	0,55
44	2.1.1.	166отсутствуют	0,24	0,67	0,65	0,68	0,83	0,36
52	2.1.1.	191передний/задний	0,64	0,67	0,63	0,71	0,39	0,55
145	2.1.1.	206	...звонкие/глухие	0,79	0,89	0,80	0,00	0,50	0,77
181	2.1.1.	244фрикативные/смычные/аффрикаты	0,79	0,44	0,81	0,95	0,44	0,77
185	2.1.1.	180единый ряд губных	0,55	0,78	0,55	0,63	0,33	0,86
216	2.1.1.	168свистящие/шипящие	0,63	0,11	0,48	0,79	0,44	0,50
232	2.1.1.	179единый ряд собственно среднеязычных	0,70	0,44	0,65	0,29	0,11	0,64
335	2.1.1.	183назальные/плавные/вibrанты/глайды	0,42	1,00	0,69	0,50	0,44	0,73
364	2.1.1.	266переднеязычные	0,88	1,00	0,83	0,97	0,56	1,00
370	2.1.1.	199глайды	0,42	1,00	0,73	0,55	0,89	0,73
446	2.1.2.	198	...динамическое	0,53	0,67	0,71	0,79	0,33	0,64
454	2.1.2.	229	...слог	0,66	0,78	0,74	0,89	0,67	0,95
518	2.1.2.	176	...фонема	0,55	0,78	0,58	0,53	0,56	0,64
719	2.1.3.	188	...нейтрализация	0,64	0,56	0,59	0,61	0,44	0,73
1145	2.3.1.	239	..морфолого-синтаксические	0,75	0,89	0,75	0,89	0,61	0,86
1245	2.3.3.	262	..единственное/множественное	0,97	0,44	0,88	0,95	0,44	0,68
1257	2.3.3.	291	..аффиксация	0,96	1,00	0,94	1,00	0,94	0,95
1294	2.3.3.	204	..актантно-предикатное	0,50	1,00	0,71	0,74	0,39	0,95
1431	2.3.4.	252	...падежные аффиксы	0,93	0,56	0,74	0,92	0,89	0,95
1562	2.3.4.	178падеж	0,91	0,22	0,35	0,53	0,83	1,00
1768	2.3.5.	160	...каузатив	0,86	0,44	0,23	0,76	0,56	0,68
1795	2.3.5.	212	...морфологическая категория	0,87	0,67	0,59	0,74	0,50	0,95
1822	2.3.5.	201	...настоящее/прошедшее/будущее	0,79	0,56	0,59	0,82	0,61	0,36
2084	2.3.5.	268	...морфологическая категория	0,89	0,89	0,91	0,92	0,50	0,86
2220	2.3.5.	211императив	0,57	0,44	0,83	0,45	0,72	0,86
2338	2.3.5.	269	...морфологическая категория	0,91	0,89	0,83	0,95	0,89	1,00
2542	2.3.6.	201	...вопросительные	0,70	0,89	0,60	0,84	0,39	0,55

Порядковый номер в БД (версия от 26.03.06)	Номер раздела	Частота	Наименование признака	алтайские (76)	афразийские (9)	индоевропейские (133)	кавказские (38)	палеоазиатские (18)	уральские (22)
2595	2.3.6.	276	...указательные	0,86	0,89	0,91	0,92	0,78	0,95
2615	2.3.6.	177	...ближний план/ дальний план	0,76	1,00	<u>0,49</u>	<u>0,47</u>	<u>0,22</u>	0,82
2697	2.3.6.	158	...притяжательные аффиксы	0,84	<u>0,44</u>	<u>0,32</u>	<u>0,26</u>	0,67	0,82
2736	2.3.6.	163	...пространственные падежи	0,79	<u>0,11</u>	<u>0,29</u>	0,82	<u>0,33</u>	1,00
2809	2.3.6.	159	...отрицательные аффиксы	0,71	<u>0,11</u>	<u>0,36</u>	0,87	<u>0,44</u>	<u>0,45</u>
2815	2.3.6.	162	...препозиция	<u>0,34</u>	0,67	0,57	<u>0,42</u>	0,50	0,91
2835	2.3.7.	297	...финитные формы	0,96	1,00	0,97	1,00	0,94	1,00
2854	2.3.7.	253	...причастие	0,89	0,78	0,80	0,92	<u>0,39</u>	1,00
2866	2.3.7.	299	..местоимение	0,93	0,89	0,95	1,00	0,94	1,00
2867	2.3.7.	261	..наречие	0,86	1,00	0,79	1,00	0,56	1,00
2872	2.3.7.	272	..прилагательное	0,92	0,56	0,92	0,95	<u>0,22</u>	1,00
2875	2.3.7.	286	..существительное	0,93	0,67	0,98	0,95	<u>0,44</u>	1,00
2876	2.3.7.	278	..числительное	0,92	1,00	0,83	1,00	0,78	1,00
2885	2.3.7.	201	..послелог	0,91	<u>0,11</u>	<u>0,40</u>	0,89	0,72	0,95
2894	2.3.7.	271	..союз	0,84	0,89	0,85	0,97	0,72	0,95
2899	2.3.7.	215	..междометия	0,71	0,89	0,59	0,84	0,78	0,95
2926	2.4.0.	287время	0,93	0,78	0,95	0,97	0,89	1,00
2928	2.4.0.	210залог	0,82	<u>0,33</u>	0,68	0,68	0,56	0,64
2932	2.4.0.	253лицо	0,87	0,89	0,87	<u>0,45</u>	0,83	1,00
2935	2.4.0.	282наклонение	0,89	0,89	0,92	0,97	0,89	1,00
2952	2.4.0.	273число	0,87	0,89	0,90	0,87	0,83	1,00
3334	2.4.0.	254	...падеж	0,91	0,78	0,75	0,95	0,72	1,00
3367	2.4.0.	282	...число	0,92	1,00	0,95	0,95	0,67	0,95
3414	2.5.1.	268	..словообразовательные аффиксы	0,88	1,00	0,82	0,95	0,83	1,00
3415	2.5.1.	273	..словоизменятельные аффиксы	0,88	1,00	0,84	0,97	0,94	1,00
3604	2.5.2.	283	...суффиксы	0,91	1,00	0,90	0,95	1,00	1,00
3657	2.5.3.	184	...SOV	0,86	<u>0,33</u>	<u>0,41</u>	0,84	0,78	<u>0,36</u>
3715	2.5.3.	158по лицу	0,51	0,78	0,52	<u>0,29</u>	<u>0,28</u>	0,95
3720	2.5.3.	173по числу	0,50	0,78	0,56	0,61	<u>0,22</u>	0,95
3809	2.5.4.	248	..сочинение/подчинение	0,70	1,00	0,77	0,97	0,67	0,95
3813	2.5.4.	208	..союзная/бессоюзная	0,63	0,67	0,65	0,87	<u>0,33</u>	0,77
3820	2.5.4.	216	...самостоятельные служебные элементы	0,54	0,67	0,79	0,68	0,61	0,68
Соотношение источник/приемник				6/5	8/12	1/9	10/8	1/20	25/4

7.2. Природа языковых раритарий

В предыдущем разделе была исследована природа статистических универсалий. Аналогичные вопросы вставали и о природе языковых раритарий, то есть сравнительно редких при-

знаках в БД. Может ли информация в БД помочь понять причины их появления и закономерности дальнейшего существования? Оказалось, что в БД есть ответ и на этот вопрос.

Были проведены расчеты и сравнительный анализ выборки из 52 мертвых языков и 52 живых языков на множестве признаков в низкочастотной сфере.

Для этого были построены так называемые ЯП-графики⁴¹. ЯП-графики показывают, сколько признаков принадлежит одному языку, двум, трем и т.д. По оси абсцисс откладывается количество признаков, по оси ординат – количество языков. На рис. 6.2 показан ЯП-график для выборки из мертвых языков (МЯ) и тестовой группы (ТГ).

Даже при поверхностном анализе графиков бросается в глаза следующее. График очень напоминает гиперболу (то есть функцию $y = 1/x$)⁴². На графике отчетливо проявляется три зоны:

- зона I – от 0 до 4 по оси ординат (языки),
- зона II – от 5 до 48,
- зона III от 49 до 52.

Они характеризуются следующими особенностями. Зона I включает удельно гораздо большее количество признаков на каждый пункт оси x, чем зоны II и III. В этой зоне сосредоточены уникалии и раритари. При этом, как можно предположить из графика, очень большое количество уникалий и редких признаков (распространенных в 2-4 языках) скорее всего не является артефактом, а отражает закономерности динамических процессов в языковом пространстве. В зонах II и III график ведет себя практически как прямая линия с небольшими всплесками в зоне III. В зонах I и III мы наблюдаем превышение значений графика ТГ над графиком МЯ. Обратим на последний факт особое внимание, так как к его интерпретации мы вернемся позже.

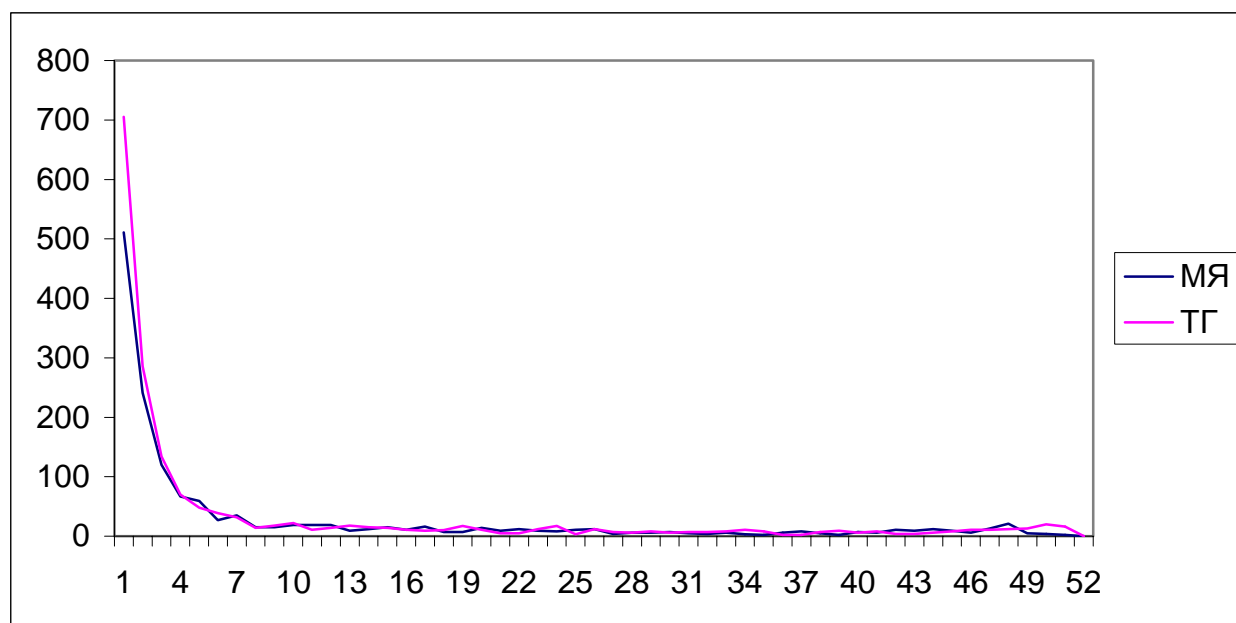


Рис. 7.2. ЯП-графики для выборки из мертвых языков (МЯ) и тестовой группы (ТГ)

⁴¹ Аббревиатура ЯП означает «язык-признак». ЯП-графики уже применялись нами для анализа закономерностей эволюции языков Евразии в главе 5.

⁴² Это не удивительно, если учесть, что этот график в каком-то смысле есть результат преобразования над распределением Ципфа, которое также подчиняется гиперболическому закону.

Однако нас интересуют раритари, то есть признаки в низкочастотной сфере. Для более детального анализа закономерностей обратимся к рис. 7.3.

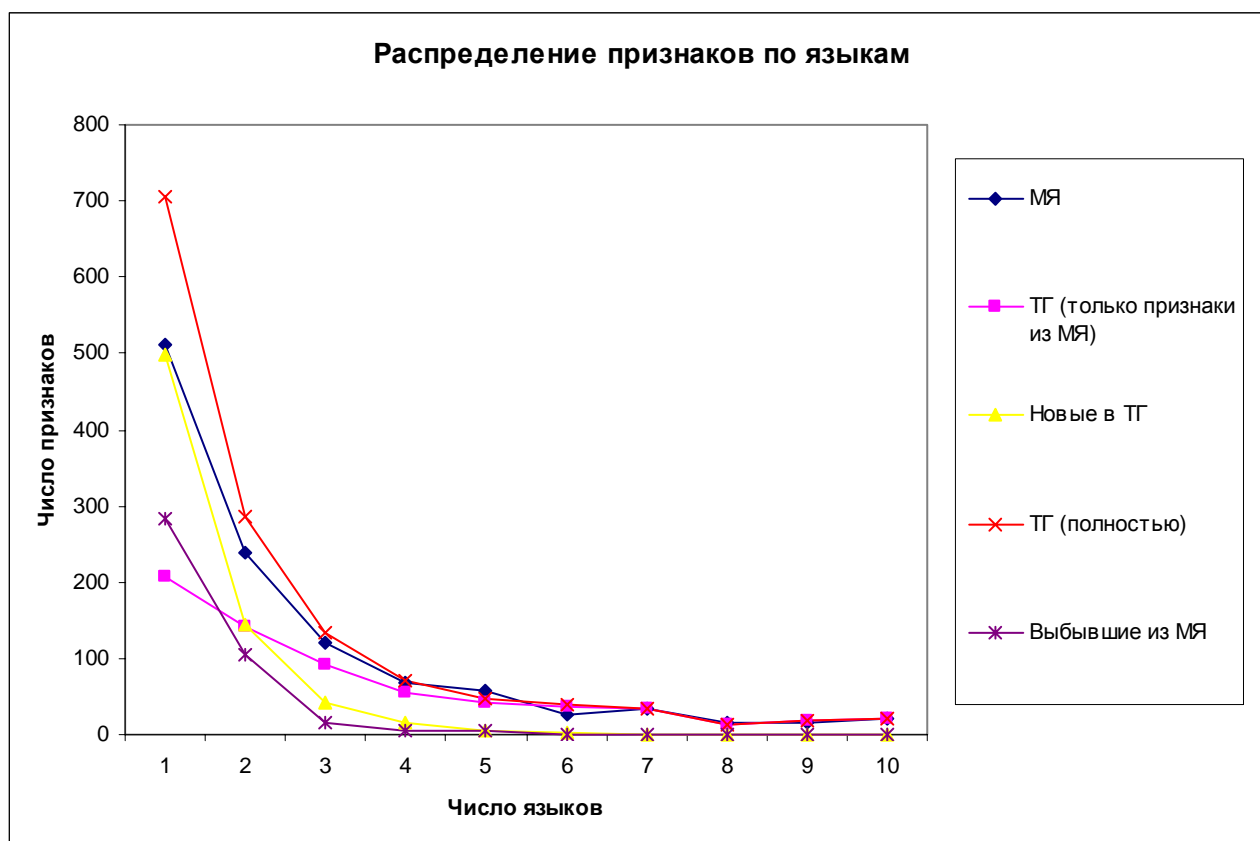


Рис. 7.3. Распределение признаков в низкочастотной сфере

Здесь показаны ЯП-графики (МЯ и ТГ) в начальной области (зона I и часть зоны II). Также представлен график признаков из мертвых языков, перешедших в языки тестовой группы, график новых признаков в тестовой группе, которых не было в мертвых языках, и график признаков, выбывших из выборки по мертвым языкам. Мы видим, что, несмотря на то, что количество уникалий в тестовой группе увеличилось, число уникалий в ТГ, доставшихся в наследство от мертвых языков резко сократилось (свыше чем на 300 признаков). Вместе с тем появилось 500 новых уникалий, до этого в МЯ не зарегистрированных вообще. Очевидно, что эти цифры не случайные и характеризуют чрезвычайно высокую динамику языковой структуры в низкочастотной сфере. Часть раритарных признаков из МЯ выбыло, часть осталась, часть переместилась в более частотную область (вправо по оси ординат на графике).

Процесс перераспределения признаков заслуживает отдельного внимания. На рис. 7.4 показаны диаграммы перераспределения признаков из группы мертвых языков в низкочастотной сфере. То есть каждый график показывает, какое положение занимают признаки в тестовой группе живых языков, если они занимали в группе мертвых языков соответственно: положение 1 (уникалии), 2 (были распространены в двух языках), 3 (были распространены в трех языках). Ромбом помечен график перераспределения признаков с первого уровня (уникалии), квадратом – с уровня 2, треугольником – с уровня 3.

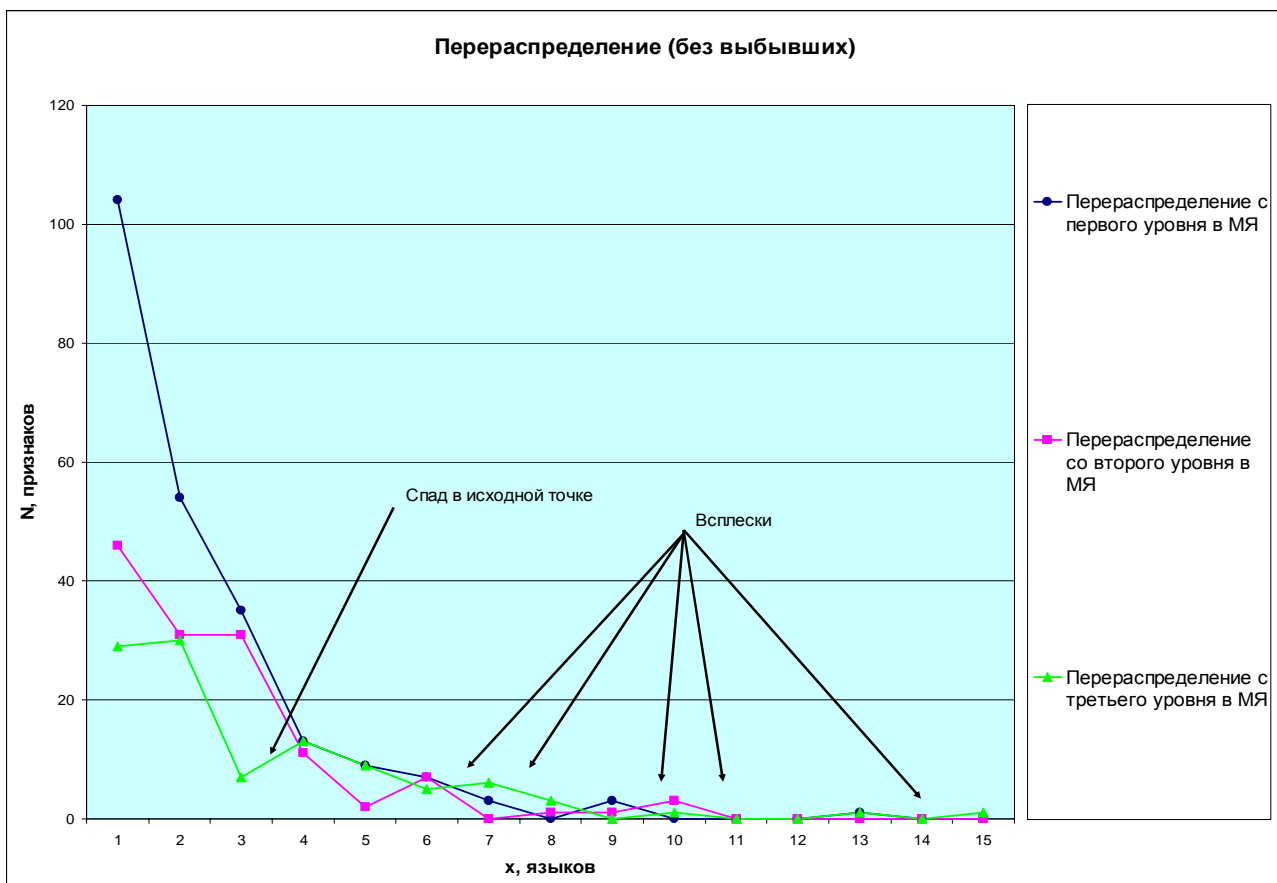


Рис. 7.4. Перераспределение признаков в низкочастотной сфере

Рисунок 7.4 показывает довольно сильную динамику раритарных признаков. При этом можно отметить несколько закономерностей.

1. Чем больше стартовая позиция смещена вправо, тем дальше вправо и в большем количестве происходит перераспределение признаков. Например, для признаков первого уровня в МЯ (уникалии) максимальным уровнем в ТГ оказался десятый, в то время как для второго и третьего уровней в МЯ максимальным уровнем перераспределения оказался уровень 15, при этом с третьего уровня на пятнадцатый уровень перераспределилось больше признаков.

2. На графиках (особенно на графиках перераспределения с уровнями 2 и 3) совершенно очевидно обозначились всплески, напоминающие затухающую периодическую функцию. Возможно, что эти всплески являются свидетельством гармонических составляющих, которые могут быть связаны с колебаниями не только в правую сторону, но и в левую. Таким образом, налицо колебательные явления, которые сопутствуют процессам перераспределения признаков. О перераспределении признаков влево очевидным образом свидетельствуют графики с уровнями 2 и 3.

3. На графике перераспределения с третьего уровня в исходной точке обозначился четкий спад, который также можно интерпретировать в пользу колебательных процессов.

Много интересного также можно получить из анализа диаграммы рис. 7.5. Здесь показано количество выбывших из оборота признаков для каждого уровня в мертвых языках и количество новых признаков в тестовой группе языков, которые продвинулись на соответствующий уровень в ТГ.

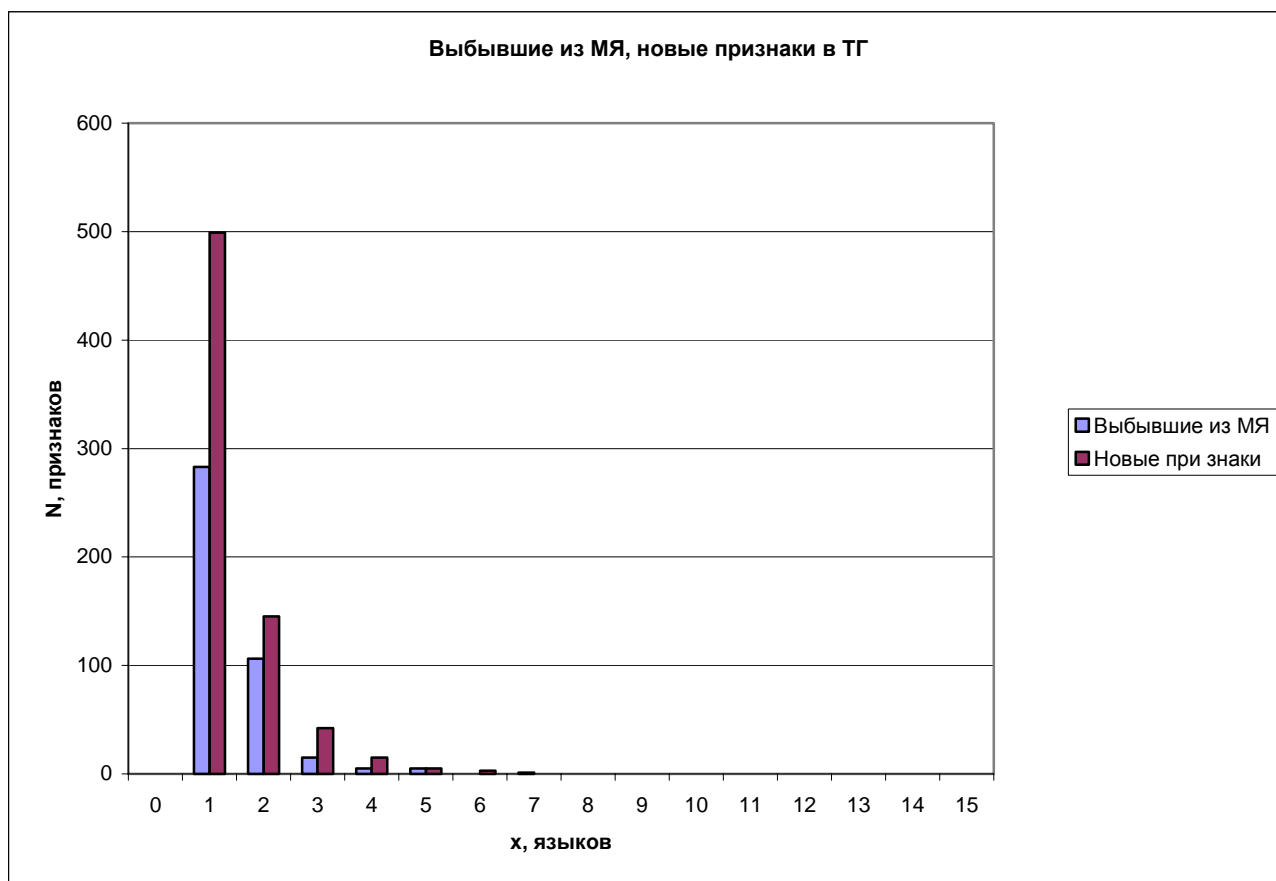


Рис. 7.5. Выбывшие признаки из МЯ, новые признаки в ТГ

Относительно выбывших признаков мы видим, что чем дальше вправо по оси частоты распространения признаков, тем меньшая доля признаков выбывает (см. табл. 7.3). Так, с уровня 1 (уникалии) выбыло около 55 процентов признаков, в то время как с уровня 2 только около 44 процентов и т.д. по убывающей. Примерно такая же картина наблюдается и в части распространения вновь «прибывших» признаков в тестовой группе. Обратим внимание на то, что дальше шестого уровня новые признаки не распространились. Процесс вымывания признаков за исследуемый период не распространился дальше седьмого уровня.

Т а б л и ц а 7.3

Численные данные к диаграммам на рис. 7.4 и 7.5

x	Число признаков в МЯ	Выбывшие из МЯ (в %)	Новые признаки (в % от остатка в МЯ)
1	512	283 (55 %)	499 (218 %)
2	239	106 (44 %)	145 (137 %)
3	120	15 (13 %)	42 (40 %)
4	67	5 (7 %)	15 (24 %)
5	59	5 (8 %)	5 (9 %)
6	27	0 (0 %)	3 (11 %)

7	34	1 (3 %)	0 (0 %)
---	----	---------	---------

Фактически это свидетельствует о том, что в тестовой группе живых языков относительно группы МЯ эволюция прошла сравнительно небольшой путь, несопоставимый с полным объемом эволюционных процессов в группе мертвых языков.

7.3. Феномен типологического сдвига

Можно резюмировать результаты исследования признакового пространства (природы статистических универсалий и раритарий), описанные в разделах 7.1. и 7.2.

1. В БД «Языки мира» описано 54 мертвых языка, 52 из них составили группу для исследований. Обнаружено, что универсалии в БД хорошо коррелируют с признаками в мертвых языках.

2. Было выбрано 11 языков (Авесты, латинский, древнееврейский, древнеперсидский, старописьменный монгольский, арабский классический – аналог древнеарабского, тюрки, урартский, половецкий, хеттский, древнегузский), которые покрыли 100 процентов статистических универсалий в БД.

3. Составлена тестовая группа из 52 живых языков, генетически близких мертвым. Было обнаружено, что универсалии в тестовой группе живых языков в большинстве случаев (67 %) имеют большую частоту, чем универсалии в мертвых. То есть универсалии становятся более универсальными.

4. Относительно раритарий было показано, что часть раритарий полностью выбывает из оборота. Эта доля уменьшается (и в количественном, и в процентном отношении) по мере увеличения частоты распространения признака на евразийском пространстве. Другая часть раритарий смещается вправо по диаграмме частотного распределения и становится более частотной. Перераспределение признаков имеет явно выраженный колебательный характер (наблюдается наличие гармонических составляющих).

Все эти явления в совокупности получили название **«феномен типологического сдвига на евразийском пространстве»**.

Авторы понимают, что правомерность применения представленной методики исследования динамики языков требует существенного обоснования. В качестве обоснования методики можно привести следующий аргумент. На диаграмме в приложении 3 показано распределение используемых в одном из расчетов 11 мертвых языков по оси времени⁴³. По этой диаграмме можно оценить период существования и эволюции тестовой группы языков. Очевидно, что этот период не превышает возраста самого «старого» из мертвых языков (хеттского) и соответствует интервалу времени около 30 веков (от XI в. до н.э до XIX в.н.э). За это время эволюции в тестовой группе новые признаки (уникалии в мертвых языках) не продвинулись дальше уровня 6 (то есть распространились только в 6 языках и меньше). В то же время распространение и количество универсалий в МЯ и устойчивость состава универсалий за 30 веков (сравнение МЯ и ТГ) показывает, что эволюционный период для мертвых языков во много раз превышает указанный отрезок.

Таким образом, вполне закономерно предположить, что меры близости, предложенные ранее (см. главу 2), кроме генетической информации содержат информацию о типологической близости, которая создает своеобразный типологический шум в рас-

⁴³ Составлено по данным из энциклопедических источников. Период существования языка датируется временем составления письменных памятников.

четах и не позволяет получить стопроцентно достоверные результаты⁴⁴. Одновременно, это позволяет по-новому взглянуть на явление языкового разнообразия и построить новые модели эволюции языков.

Дальнейшие усилия исследовательской группы предполагается направить на верификацию полученных количественных результатов и выводов. Для этого планируется привлечь альтернативные источники информации и методики расчетов.

7.4. Резюме по главе 7

1. Количественные методы наиболее эффективно могут быть применены к описанию макроэволюции языков и языковых признаков.
2. Наличие в БД достаточно полных описаний 52 мертвых языков позволяет провести систематическое сравнение распространения свойств в языках, существовавших в среднем 2 тыс. лет назад, и современных. Это дает новый инструмент анализа языковой эволюции.
3. Обнаружено явление, названное «типологическим сдвигом» и состоящее в том, что за указанный промежуток времени частотные признаки (статистические универсалии) стали еще более частотными, а редкие признаки (раритариум) подверглись вымыванию.

⁴⁴ Однако это не означает, что качественную меру близости нельзя создать в принципе.

ЗАКЛЮЧЕНИЕ

В книге рассмотрены новые методики количественных исследований на материале базы данных «Языки мира», в том числе методы расчета парных мер близости, методы кластерного анализа и интеллектуального анализа данных.

Были рассчитаны количественные портреты по всему генетическому указателю БД. С помощью этих расчетов удалось выявить новые статистические закономерности эволюции языков. Обнаружен феномен типологического сдвига на евразийском пространстве, который также объясняет наличие сильного типологического фона, препятствующего выявлению генетической близости языков прямыми компьютерными методами.

Обнаруженные факты и закономерности позволили по-новому взглянуть на явление языкового разнообразия и построить новую модель структурной эволюции языков.

Была предложена новая методика верификации генетических гипотез, основанная на использовании раритатных признаков. Методика дает хорошее соответствие с традиционными представлениями для языков романско-германской группы, алтайской макросемьи. Были получены данные в поддержку генетической близости шумерского и семитских языков, нивхского и чукотско-камчатских. Принадлежность старояпонского к алтайской макросемье не подтвердилась.

В целом можно сказать, что база данных «Языки мира» является ценным компьютерным лингвистическим ресурсом и значение ее для научного сообщества со временем будет только возрастать.

Литература

- (Арапов, Херц 1974) Арапов А.Е., Херц М.М. Математические методы в исторической лингвистике. – М., 1974.
- (Атлас 1998) Атлас языков мира. / Ред.: Комри Б., Мэттьюс С., Полински М. – Словакия: Лик пресс, 1998. – 224 с.
- (Балалыкина 2001) Балалыкина Э.А. Русско-немецкий диалог XIX-XX века по балтославянским языковым отношениям // Немецко-русский диалог в филологии. – Peter Lang, 2001. – С. 39-51.
- (Бодуэн 1903) Бодуэн де Куртенэ И.А. Лингвистические заметки и афоризмы. – ЖМНП. – Т.346. – 1903.
- (БЭС 1997) Большой энциклопедический словарь. – Изд. 2-е. – М.: Изд-во «Большая российская энциклопедия», 1997. – 1434 с.
- (Бугаева 2006) Бугаева А. Частное сообщение. 2006.
- (Бурлак, Старостин 2001) Бурлак С.А., Старостин С.А. Введение в лингвистическую компаративистику. – М.: УРСС, 2001. – 272 с.
- (Виноградов, Новиков, Ярославцева 2003) Виноградов В.А., Новиков А.И., Ярославцева Е.И. База данных «Языки мира» как инструмент лингвистического исследования // Вопросы языкознания. – 2003. – №3.
- (Виноградов, Новиков, Ярославцева, Поляков, Логунов 2003) Виноградов В.А., Новиков А.И., Ярославцева Е.И., Поляков В.Н., Логунов В.В.. База данных «Языки мира» и лингвистические исследования // II Международные Бодуэновские чтения (Казань, 11-13 декабря 2003 г.). Труды и материалы. Т.1. – С. 39-41.
- (Глас 2006) Глас Р. Защита от ротавируса // В мире науки. – № 7. – 2006.
- (Дресслер 1988) Дресслер В. К вопросу о реконструкции индоевропейского синтаксиса // Новое в зарубежной лингвистике. – Вып. XXI. – М., 1988.
- (Жамбю 1988) Жамбю М. Иерархический кластерный анализ и соответствия. – М.: Финансы и статистика, 1988. – 343 с.
- (Журинская, Новиков, Ярославцева 1986) Журинская М. А., Новиков А.И., Ярославцева Е.И. Энциклопедическое описание языков. – М.: «Наука», 1986.
- (Климов 1980) Климов Г.А. К взаимоотношению генеалогической, типологической и ареальной классификаций языков // Теоретические основы классификации языков мира. – М., 1980. – С. 6-23.
- (Лингвистический 1990) Лингвистический энциклопедический словарь / Ред. Ярцева В.Н. – М.: Советская энциклопедия, 1990. – 685 с.
- (Новиков, Ярославцева 1985) Новиков А.И., Ярославцева Е.И. База лингвотипологических данных и принципы ее функционирования // Вести АН СССР. – 1985. – № 3.
- (Пенроуз 2004) Пенроуз Р. Новый ум короля. – М.: УРСС, 2004. – 400 с.
- (Поляков, Савельев, Соловьев 2006-1) Поляков В.Н., Савельев А.А., Соловьев В.Д. Опыт применения методов интеллектуального анализа данных в компаративистских и типологических исследованиях (на материале созданной в ИЯ РАН БД «Языки мира») // КИИ-2006: десятая национальная конференция по искусственному интеллекту с международным участием: труды конференции, Т.1. – М.: Физматлит, 2006. – С.217-224.

- (Поляков, Савельев, Соловьев 2006-2) Поляков В.Н., Савельев А.А., Соловьев В.Д. Методы интеллектуального анализа данных по типологии языков // Труды межд. конф. по искусственному интеллекту. – Таганрог, 2006.
- (Поляков, Соловьев 2006-1) Поляков В.Н., Соловьев В.Д. База данных «ЯЗЫКИ МИРА»: текущее состояние и концепция развития // Труды школы по компьютерной и когнитивной лингвистике. – Казань: Отечество, 2006.
- (Поляков, Соловьев 2006-2) Поляков В.Н., Соловьев В.Д. База данных «ЯЗЫКИ МИРА» и статистические исследования в типологии // III Межд. Бодуэновские чтения. Труды, т.2. – Казань: КГУ, 2006. – С. 225-227.
- (Поляков, Соловьев 2006-3) Поляков В.Н., Соловьев В.Д. Сложное предложение в базе данных «Языки мира» // Труды международного симпозиума «Языки Европы, Северной и Центральной Азии». – Томск: ТГПУ, 2006. – С. 80-82.
- (Поляков, Соловьев 2006-4) Поляков В.Н., Соловьев В.Д. Синтаксис простого предложения в базе данных «Языки мира» // Теория перевода. Типология языков. Межкультурная коммуникация. – М.: МГУ, 2006 (в печати).
- (Поляков, Соловьев 2006-5) Поляков В.Н., Соловьев В.Д. База данных «Языки мира»: текущее состояние (на 01.12.2006). Препринт. – Казань: КГУ, 2006.
- (Поляков, Соловьев 2006-6) Поляков В.Н., Соловьев В.Д. О новых возможностях количественных исследований (на материале БД «Языки мира») // Труды Бодуэновской конф. – Казань: КГУ, 2006 (в печати).
- (Поляков, Соловьев 2006-7) Поляков В.Н., Соловьев В.Д. Подходы к классификации языков. Анализ положения дел // Труды Казанской школы по когнитивной и компьютерной лингвистике. – Казань: КГУ, 2007 (в печати).
- (Поляков, Соловьев, Ахтямов 2006) Поляков В.Н., Соловьев В.Д., Ахтямов Р.Б. Сложное предложение в базе данных «Языки мира»: статистика и методы исследования // Труды лаборатории языков народов Сибири. – Томск: ТГПУ, 2006 (в печати).
- (Раушенбах 1985) Раушенбах Г.В. Меры близости и сходства // Анализ нечисловой информации в социологических исследованиях. – М.: Наука, 1985. – С. 169-203.
- (Савельев 2004) Савельев А.А. Моделирование пространственной структуры растительного покрова (геоинформационный подход) / Казань: Казанский государственный университет, 2004. – 247с.
- (Сводеш 1999) Сводеш М. Лексикостатистическое датирование доисторических этнических контактов // Зарубежная лингвистика. – Вып.1. Ред.: Звягинцева В.А., Чемоданова Н.С. – М.: Прогресс, 1999. – С. 7-36.
- (Сильницкий 2004) Сильницкий Г. Г. Количественная грамматико-фонетическая типология языков и языковых признаков. – Смоленск, 2004. – 280 с.
- (Соловьев, Поляков 2006) Соловьев В.Д., Поляков В.Н. Модель распространения языковых признаков // Web Journal of Formal, Computational & Cognitive Linguistics. – 2006. – <http://fccl.ksu.ru>.
- (Табасаранские 1982) Табасаранские этюды. – М.: МГУ, 1982.
- (Феллер 1984) Феллер В. Введение в теорию вероятностей и ее приложения. – М.: Мир, 1984. – 752 с.
- (Шахматов 1915) Шахматов А.А. Очерк древнейшего периода истории русского языка // Энциклопедия славянской филологии. – Вып. II. – Петроград, 1915.

- (Якобсон 1963) Якобсон Я. Типологические исследования и их вклад в сравнительно-историческое языкознание // Новое в лингвистике. – Вып. 3. – М., 1963. – С. 95-105.
- (Языки мира: Уральские 1993) Языки мира: Уральские языки. – М., 1993.
- (Языки мира: Тюркские 1997) Языки мира: Тюркские языки. – М.: Издательство «Индрик», 1997. – 544 с.
- (Языки мира: Монгольские 1997) Языки мира: Монгольские языки. Тунгусо-маньчжурские языки. Японский язык. Корейский язык. / ред.: Кибрик А.А., Рогова Н.В., Романова О.И. – М.: Издательство «Индрик», 1997. – 408 с.
- (Языки мира. Палеоазиатские 1996) Языки мира. Палеоазиатские языки. – М.: Издательство «Индрик», 1996. – 231 с.
- (Языки мира: Иранские 1997) Языки мира: Иранские языки. I. Юго-западные иранские языки. – М.: Индрик, 1997. – 207 с.
- (Языки мира: Иранские 1999-1) Языки мира: Иранские языки. II. Северо-западные иранские языки. – М.: Индрик, 1999. – 302 с.
- (Языки мира: Иранские 1999-2) Языки мира: Иранские языки. III. Восточноиранские языки. – М.: Индрик, 1999. – 343 с.
- (Языки мира: Дардские 1998) Языки мира: Дардские и нуристанские языки. – М.: Индрик, 1998. – 143 с.
- (Языки мира: Кавказские 2001) Языки мира: Кавказские языки. – М.: Academia, 2001. – 480с.
- (Языки мира: Германские 1999) Языки мира: Германские языки. Кельтские языки. – М.: Academia, 1999. – 472 с.
- (Языки мира: Романские 2001) Языки мира: Романские языки. – М: Academia, 2001. – 720 с.
- (Языки мира: Индоарийские 2004) Языки мира: Индоарийские языки древнего и среднего периода. – М.: Academia, 2004. – 160 с.
- (Языки мира: Славянские 2005) Языки мира: Славянские языки / РАН. Институт языкознания / Ред. колл.: А. М. Молдован, С. С. Скорвид, А. А. Кибрик и др. — М.: Academia, 2005. – 656 с.
- (Языки мира: Балтийские 2006) Языки мира: Балтийские языки / РАН. Институт языкознания / Ред. Коллегия.: В.Н.Топоров, М.В.Завьялова, А.А.Кибрик и др. – М.: Academia, 2006. – 224 с.
- (Ярославцева 1999) Ярославцева Е.И. Грамматикон, его виды и аналоги // Язык, сознание, коммуникация. – Вып.10. – МГУ. – М., 1999.
- (Ярославцева 2001)Ярославцева Е.И. Грамматикон и база данных «Языки мира» // Проблемы прикладной лингвистики 2001. – М., 2002.
- (Ярославцева 2005) Ярославцева Е.И. Компьютерная база данных «Языки Мира» и ее возможные применения. Автореф. ... д-ра фил. наук по специальности 10.02.21 – прикладная лингвистика. – ИЯз РАН, 2005.
- (Ярцева 1977) Ярцева В.Н. Научно-техническая революция и развитие языка // Научно-техническая революция и функционирование языков мира. – М., 1977.
- (Bishop 1998) Bishop С.М. GTM: the Generative Topographic Mapping / С.М.Bishop, М. Svensen, С.К.I.Williams // Neural Computation. – 1998. – №10 (1). – P. 215-234.
- (Croft 1990) Croft W. Typology and Universals. – Cambridge: Cambridge University Press, 1990. – 310 p.

- (Dahl Ö., 2004) Dahl Ö. The growth and maintenance of linguistic complexity. – Amsterdam: John Benjamins, 2004.
- (Janhunen 2004) Janhunen Ju. On the Hierarchy of Structural convergence in the Amdo Sprachbund // International symposium «The Typology of Argument Structure and Grammatical Relations». – Kazan: KSU, 2004.
- (Givón 2002) Givón T. Bio-linguistics. – Amsterdam: Benjamins Publ. Co, 2002.
- (Haspelmath 2005) Haspelmath, Martin & Matthew S. Dryer & David Gil & Bernard Comrie (eds.) The World Atlas of Language Structures. – Oxford: Oxford University Press, 2005. – 695 p.
- (Keenan 1985) Keenan E. Passive in the World's languages // Linguistic typology and syntactic description. – V.1, ed. T. Shopen. – Cambridge: Cambridge University Press, 1985. – P. 243-281.
- (Kruskal 1964) Kruskal J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis / J.B.Kruskal // Psychometrika. – 1964. – №29. – P.1-27.
- (Kohonen 1997) Kohonen T. Self-organization and Associative Memory / T. Kohonen. – New York: Springer-Verlag, 1997. – 428 p.
- (Laudan 1976) Laudan L. Progress and its problems: towards a theory of scientific growth. – Berkeley: University of California Press, 1976.
- (Legendre 1998) Legendre L. Numerical ecology / L.Legendre, P.Legendre. – Amsterdam: Elsevier Science, 1998. – 853 p.
- (Mithun 2004) Mithun M. The potential instability of grammatical relations // International symposium «The Typology of Argument Structure and Grammatical Relations». – Kazan: KSU, 2004.
- (Nichols 1992) Nichols J. Linguistic Diversity in Space and Time. – Chicago: University of Chicago Press, 1992. – 358 p.
- (Novikov, Yaroslavtseva 1986) Novikov A., Yaroslavtseva E. Linguotypological Data Bank // Social Sciences. – USSR Academy of Sciences. – Vol. XVII. – No. 3. – 1986.
- (R Development Core Team 2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>).
- (Rosch 1975) Rosch E. Cognitive representations of semantic categories // Journal of Experimental Psychology: General, V. 104, 1975.
- (Schleicher 1871) Schleicher A. Compendium der vergleichenden Grammatik der indogermanischen Sprachen. Auf. III. – Weimar, 1871.

Генетическая принадлежность языков БД

Язык	Классификация, принятая в издании «Языки мира»	Классификация, принятая в (Бурлак, Старостин 2001)	Кол-во признаков в реферате	Прим.
Абазинский	Кавказские > абхазо-адыгские > абхазо-абазинские	Синокавказские > северно-кавказские > абхазо-адыгские > абхазские	288	
Абхазский	Кавказские > абхазо-адыгские > абхазо-абазинские	Синокавказские > северно-кавказские > абхазо-адыгские > абхазские	290	
Аварский	Кавказские > нахско-дагестанские > аваро-андо-цезские > аварский	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские	333	
Авесты	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > иранские	399	
Агульский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские > восточно-лезгинские	320	
Адыгейский	Кавказские > абхазо-адыгские > адыгские	Синокавказские > северно-кавказские > абхазо-адыгские > адыгские	303	
Азербайджанский	Алтайские > тюркские > огузские	Ностратические > алтайские > тюркские > огузские	330	
Азиатских эскимосов	Палеоазиатские > эскимосско-алеутские > эскимосские > юпикские	Ностратические > эскалеутские > эскимосские > группа юпик	251	
Айнский	Палеоазиатские. Изолят	Изолят	292	
Аккадский	Афразийские > семитские > восточnoseмитские	Афразийские > семитские > северосемитские > северо-восточные	330	
Алабугатских татар	Алтайские > тюркские > кыпчакские > кыпчакско-ногайские		315	
Аланский	Индоевропейские > иранские > восточные иранские > скифские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские >	106	

Язык	Классификация, принятая в издании «Языки мира»	Классификация, принятая в (Бурлак, Старостин 2001)	Кол-во признаков в реферате	Прим.
		восточно-иранские		
Алеутский	Палеоазиатские > эскимосско-алеутские > алеутский	Ностратические > эскалеутские > алеутский	284	
Алтайский	Алтайские > тюркские > хакасско-алтайские > алтайские	Ностратические > алтайские > тюркские > сибирские	288	
Аляскинский юпик	Палеоазиатские > эскимосско-алеутские > эскимосские > юпикские	Ностратические > эскалеутские > эскимосские > группа юпик	235	
Английский	Индоевропейские > германские > западногерманские > англо-фризские	Ностратические > индоевропейские > собственно-индоевропейские > германские > западногерманские	391	
Андийский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	316	
Арабский классический	Афразийские > семитские > западносемитские > североаравийские	Афразийские > семитские > северосемитские > северо-западные > центральные > южно-центральные	370	
Арабский современный	Афразийские > семитские > западносемитские > североаравийские	Афразийские > семитские > северосемитские > северо-западные > центральные > южно-центральные	348	
Арагонский	Индоевропейские > романские > иберо-романские		244	
Армянский	Индоевропейские > армянские	Ностратические > индоевропейские > собственно-индоевропейские > армянские	364	
Арумынский	Индоевропейские > романские > балкано-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > балканские	285	
Арчинский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские	376	

Язык	Классификация, принятая в издании «Языки мира»	Классификация, принятая в (Бурлак, Старостин 2001)	Кол-во признаков в реферате	Прим.
Ассамский	Индоевропейские > индоарийские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > индийские	338	
Астраханских ногайцев-карагашей	Алтайские > тюркские > кыпчакские > кыпчакско-ногайские		310	
Астурийский	Индоевропейские > романские > иберо-романские		291	
Аунги	Афразийские > кушитские > агавские	Афразийские > кушитские > центральнокушитские	370	
Афганский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	366	
Африкаанс	Индоевропейские > германские > западногерманские > нижненемецкие	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	274	
Ахвахский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	313	
Ашкун	Индоевропейские > нуристанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > нуристанские	246	
Багвалинский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	309	
Бактрийский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	232	м
Балканских тюрков	Алтайские > тюркские > огузские		174	

Барабинских татар	Алтайские > тюркские > кыпчакские > центральнокыпчакские		294	
Бартагский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские > шугнано-рушанские		255	
Бацбийский	Кавказские > нахско-дагестанские > нахские	Синокавказские > северно-кавказские > нахско-дагестанские > нахские	255	
Башкарди диалек. гр.	Индоевропейские > иранские > юго-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	135	
Башкирский	Алтайские > тюркские > кыпчакские > башкирский	Ностратические > алтайские > тюркские > кыпчакские	328	
Бежтинский	Кавказские > нахско-дагестанские > аваро-андо-цезские > цезские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > цезские > восточно-цезские	334	
Белорусский	Индоевропейские > славянские > восточнославянские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > южнославянские	406	
Белуджский	Индоевропейские > иранские > северо-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	371	
Бенгальский	Индоевропейские > индоарийские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > индийские	344	
Бирманский	Сино-тибетские > тибето-бирманские > лоло-бирманские	Синокавказские > сино-тибетские > тибето-бирманские	297	
Боджнурди	Алтайские > тюркские > огузские		236	
Болгарский	Индоевропейские > славянские > южнославянские > болгаро-македонские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > восточнославянские	380	
Ботлихский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	282	
Бретонский	Индоевропейские > кельтские > бриттские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	314	

Будухский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские > южно-лезгинские	285	
Булгарский	Алтайские > тюркские > болгарские	Ностратические > алтайские > тюркские > болгарские	190	м, к
Бурушаски	Изолят	Изолят	371	
Бурятский	Алтайские > монгольские	Ностратические > алтайские > монгольские > северные	352	
Вайгали	Индоевропейские > нуристанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > нуристанские	236	
Валлийский	Индоевропейские > кельтские > британские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	358	
Ванджский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские		56	
Ванеци	Индоевропейские > иранские > восточные иранские		227	
Ваханский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	258	
Венгерский	Уральские > финно-угорские > угорские	Ностратические > уральские > финно-угорские > угорские	327	
Вепский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > финно-саамские > прибалтийско-финские	346	
Водский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > финно-саамские > прибалтийско-финские	295	
Вьетнамский	Аустроазиатские > мон-кхмерские > вьетийские	Мяо-аустроазиатские > аустроазиатские > вьетские	232	
Гавар	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	209	
Гагаузский	Алтайские > тюркские > огузские	Ностратические > алтайские > тюркские > огузские	318	

Галисийский	Индоевропейские > романские > иберо-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	334	
Галльский	Индоевропейские > кельтские > континентальные кельтские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	191	м
Гарви	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	211	
Гасконский	Индоевропейские > романские > галло-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	246	
Гилянский	Индоевропейские > иранские > северо-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	328	
Гинухский	Кавказские > нахско-дагестанские > аваро-андо-цезские > цезские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > цезские > западно-цезские	358	
Глангали	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	246	
Годоберинский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	309	
Горномарийский литературный	Уральские > финно-угорские > финно-пермские > марийские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > марийские	326	
Готский	Индоевропейские > германские > восточногерманские	Ностратические > индоевропейские > собственно-индоевропейские > германские > восточногерманские	331	м
Гренландский	Палеоазиатские > эскимосско-алеутские > эскимосские > инуитские	Ностратические > эскалеутские > эскимосские > группа инуит	237	
Грузинский	Кавказские > картвельские	Ностратические > картвельские > грузинско-занские	343	
Гуджарати	Индоевропейские > индоарийские	Ностратические > Индоевропейские > Собственно-индоевропейские > индоиранские > индийские	387	

Гунзибский	Кавказские > нахско-дагестанские > аваро-андо-цезские > цезские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > цезские > восточно-цезские	341	
Гуннов	Алтайские > тюркские		86	м
Давани диалект	Индоевропейские > иранские > юго-западные иранские		145	
Дагурский	Алтайские > монгольские	Ностратические > алтайские > монгольские > южные	300	
Далматинский	Индоевропейские > романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > балканские	257	
Дамели	Индоевропейские > дардские (?)	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские (?)	219	
Даргинский	Кавказские > нахско-дагестанские > даргинские	Синокавказские > северно-кавказские > нахско-дагестанские	321	
Дари	Индоевропейские > иранские > юго-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	349	
Датский	Индоевропейские > германские > скандинавские	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	389	
Долганский	Алтайские > тюркские > якутские	Ностратические > алтайские > тюркские > якутские	276	
Древнееврейский	Афразийские > семитские > западносемитские > ханаанейские	Афразийские > семитские > северосемитские > северо-западные > центральные > северо-центральные > левантийские > ханаанейские	346	м
Древнекитайский	Сино-тибетские > синитские	Сино-кавказские > сино-тибетские > китайские	214	м
Древнеогузский	Алтайские > тюркские > огузские		276	м, к
Древнеперсидский	Индоевропейские > иранские > юго-западные > иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	352	м
Древнеуйгурский	Алтайские > тюркские > древнетюркские	Ностратические > алтайские > тюркские > карлукские	269	м

Дунсянский	Алтайские > монгольские	Ностратические > алтайские > монгольские > южные	270	
Еврейский	Афразийские > семитские > западносемитские > ханаанейские	Афразийские > семитские > северосемитские > северо-западные > центральные > северо-центральные > левантийские > ханаанейские	322	
Еврейско-испанский	Индоевропейские > романские > иберо-романские > еврейско-романские		348	
Земиаки	Индоевропейские > нуристанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > нуристанские	212	
Зенага	Афразийские > берберо-канарские > западноберберские	Афразийские > берберо-канарские > западноберберские	247	
Идиш	Индоевропейские > германские > западногерманские > средненемецкие	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	305	
Ижорский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > финно-саамские > прибалтийско-финские	319	
Ингушский	Кавказские > нахско-дагестанские > нахские	Синокавказские > северно-кавказские > нахско-дагестанские > нахские	283	
Иннуитов аляски	Палеоазиатские > эскимосско-алеутские > эскимосские > инуитские	Ностратические > эскалеутские > эскимосские > группа инуит	247	
Иннуитов канады ⁴⁵	Палеоазиатские > эскимосско-алеутские > эскимосские > инуитские	Ностратические > эскалеутские > эскимосские > группа инуит	280	
Ирландский	Индоевропейские > кельтские > гойдельские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	347	
Исландский	Индоевропейские > германские > скандинавские	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	389	

⁴⁵ Незавершен

Испанский	Индоевропейские > романские > иберо-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	359	к
Иссык-кульских калмыков	Алтайские > монгольские		259	
Истророманский	Индоевропейские > романские > балкано-романские		227	
Истрорумынский	Индоевропейские > романские > балкано-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > балканские	297	
Итальянский	Индоевропейские > романские > итало-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	349	
Ительменский	Палеоазиатские > чукотско-камчатские	Чукотско-камчатские	274	
Ишкашимский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	300	
Кабардинский	Кавказские > абхазо-адыгские > адыгские	Синокавказские > северно-кавказские > абхазо-адыгские > адыгские	326	
Казахский	Алтайские > тюркские > кыпчакские > кыпчакско-ногайские	Ностратические > алтайские > тюркские > кыпчакские	331	
Калмыцкий	Алтайские > монгольские	Ностратические > алтайские > монгольские > северные	351	
Камасинский	Уральские > самодийские	Ностратические > уральские > самодийские > северно-самодийские	282	м
Караимский	Алтайские > тюркские > кыпчакские > центральнокыпчакские	Ностратические > алтайские > тюркские > кыпчакские	294	
Кара-калпакский	Алтайские > тюркские > кыпчакские > кыпчакско-ногайские	Ностратические > алтайские > тюркские > кыпчакские	288	
Каратинский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	306	
Караханидо-уйгурский	Алтайские > тюркские > карлукские		238	м

Карачаево-балкарский	Алтайские > тюркские > кыпчакские > кыпчакско-половецкие	Ностратические > алтайские > тюркские > кыпчакские	294	
Карельский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > ульские > финно-угорские > финно-пермские > финно-волжские > финно-саамские > прибалтийско-финские	338	
Картакалай	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	277	
Каталанский	Индоевропейские > романские > иберо-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	304	
Кати	Индоевропейские > нуристанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > нуристанские	362	
Кашмири	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	386	
Кашубский	Индоевропейские > славянские > западнославянские > лехитские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > западнославянские	343	
Кельтиберский	Индоевропейские > кельтские > континентальные > кельтские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	196	м
Керекский	Палеоазиатские > чукотско-камчатские	Чукотско-камчатские > чукотско-корякские	214	
Кетский	Палеоазиатские > енисейские	Синокавказские > енисейские	345	
Киргизский	Алтайские > тюркские > хакасско-алтайские > киргизский	Ностратические > алтайские > тюркские > кыпчакские	299	
Коми-зырянский	Уральские > финно-угорские > финно-пермские > пермские	Ностратические > ульские > финно-угорские > финно-пермские > пермские	333	
Коми-пермяцкий	Уральские > финно-угорские > финно-пермские > пермские	Ностратические > ульские > финно-угорские > финно-пермские > пермские	319	
Корнский	Индоевропейские > кельтские > британские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	212	м
Корсиканский	Индоевропейские > романские > итало-романские		232	

Корякский	Палеоазиатские > чукотско-камчатские	Чукотско-камчатские > чукотско-корякские > корякские	336	
Коттский	Палеоазиатские > енисейские	Синокавказские > енисейские	304	м
Крызский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские > южно-лезгинские	293	
Крымскотатарский	Алтайские > тюркские > кыпчакские > кыпчакско-половецкие	Ностратические > алтайские > тюркские > кыпчакские	349	
Крымчакский	Алтайские > тюркские > кыпчакские > кыпчакско-половецкие	Ностратические > алтайские > тюркские > кыпчакские	244	
Кудрари	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	59	
Кумзари	Индоевропейские > иранские > юго-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	163	
Кумыкский	Алтайские > тюркские > кыпчакские > кыпчакско-половецкие	Ностратические > алтайские > тюркские > кыпчакские	305	
Курдшули	Индоевропейские > иранские > юго-западные иранские		134	
Кхмерский	Аустроазиатские > мон-кхмерские > Кхмерские	Мяо-аустроазиатские > аустроазиатские	274	
Кховар	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	243	
Ладинский	Индоевропейские > романские > галло-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	296	
Лазский	Кавказские > картвельские	Ностратические > картвельские > грузинско-занские > занские	280	
Лакский	Кавказские > нахско-дагестанские > лакский	Синокавказские > северно-кавказские > нахско-дагестанские	279	
Ларская диал. гр.	Индоевропейские > иранские > юго-западные иранские		179	

Латинский	Индоевропейские > италийские > латино-фалисские	Ностратические > индоевропейские > собственно-индоевропейские > италийские > древние	409	м
Лезгинский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские > восточно-лезгинские	298	
Лепонтийский	Индоевропейские > кельтские > континентальные кельтские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	131	м
Ливский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > финно-саамские > прибалтийско-финские	325	
Лурские и бахтиярские диалекты	Индоевропейские > иранские > юго-западные иранские		215	
Люксембургский	Индоевропейские > германские > западногерманские > средненемецкие	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	291	
Мазандеранский	Индоевропейские > иранские > северо-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	237	
Майан	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	206	
Македонский	Индоевропейские > славянские > южнославянские > болгаро-македонские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > южнославянские	345	
Мамлюкско-кыпчакский	Алтайские > тюркские > кыпчакские		197	м, к
Маньчжурский	Алтайские > тунгусо-маньчжурские > южные	Ностратические > алтайские > тунгусо-маньчжурские	313	м
Марийский	Уральские > финно-угорские > финно-пермские > марийские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > марийские	334	
Маторско-тайгийско-карагасский	Уральские > самодийские	Ностратические > уральские > самодийские > южно-самодийские	219	м
Мегленорумынский	Индоевропейские > романские > балкано-романские	Ностратические > индоевропейские > собственно-индоевропейские > италийские > романские > балканские	311	

Мегрельский	Кавказские > картвельские	Ностратические > картвельские > грузинско-занские > занские	287	
Медновских алеутов	Палеоазиатские > эскимосско-алеутские > алеутский		232	
Могольский	Алтайские > монгольские	Ностратические > алтайские > монгольские > юго-западные	286	м
Мокшанский	Уральские > финно-угорские > финно-пермские > мордовские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > мордовские	375	
Молдавский	Индоевропейские > романские > балкано-романские	Ностратические > индоевропейские > собственно-индоевропейские > италийские > романские > балканские	314	
Монголов внутренней Монголии	Алтайские > монгольские		301	
Монгольский	Алтайские > монгольские	Ностратические > алтайские > монгольские > северные	338	
Монгорский	Алтайские > монгольские	Ностратические > алтайские > монгольские > южные	277	
Мунджанский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	333	
Мэнкский	Индоевропейские > кельтские > гойдельские	Ностратические > индоевропейские > собственно-индоевропейские > кельтские	292	
Нанайский	Алтайские > тунгусо-маньчжурские > амурские	Ностратические > алтайские > тунгусо-маньчжурские > южные	363	
Нганасанский	Уральские > самодийские	Ностратические > уральские > самодийские > северно-самодийские	310	
Негидальский	Алтайские > тунгусо-маньчжурские > сибирские	Ностратические > алтайские > тунгусо-маньчжурские > южные	351	
Немецкий	Индоевропейские > германские > западногерманские > средненемецкие	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	423	
Немецко-швейцарский диалект	Индоевропейские > германские > западногерманские > верхненемецкие		288	

Ненецкий	Уральские > самодийские	Ностратические > уральские > самодийские > северно-самодийские	370	
Нивхский	Палеоазиатские. Изолят	Изолят	343	
Нидерландский	Индоевропейские > германские > западногерманские > нижненемецкие	Ностратические > индоевропейские > собственноиндоевропейские > германские > северогерманские	385	
Новоарамейские диалекты	Афразийские > семитские > западносемитские > арамейские	Афразийские > семитские > северосемитские > северо-западные > центральные > левантийские > арамейские	271	
Ногайский	Алтайские > тюркские > кыпчакские > кыпчакско-ногайские	Ностратические > алтайские > тюркские > кыпчакские	310	
Норвежский	Индоевропейские > германские > скандинавские	Ностратические > индоевропейские > собственноиндоевропейские > германские > северогерманские	415	
Огузский X-XI вв.	Алтайские > тюркские > огузские	Ностратические > алтайские > тюркские > огузские	280	м, к
Орокский	Алтайские > тунгусо-маньчжурские > амурские	Ностратические > алтайские > тунгусо-маньчжурские > южные	366	
Орочский	Алтайские > тунгусо-маньчжурские > амурские	Ностратические > алтайские > тунгусо-маньчжурские > южные	334	
Орхоно-енисейских надписей	Алтайские > тюркские > древнетюркские	Ностратические > алтайские > тюркские > сибирские	316	м,к
Оскский	Индоевропейские > италийские > оскско-умбрские	Ностратические > индоевропейские > собственноиндоевропейские > италийские > древние	302	м
Пали	Индоевропейские > индоарийские	Ностратические > индоевропейские > собственноиндоевропейские > индоиранские	320	
Панджаби	Индоевропейские > индоарийские	Ностратические > индоевропейские > собственноиндоевропейские > индоиранские	386	
Парачи	Индоевропейские > иранские > северо-западные иранские	Ностратические > индоевропейские > собственноиндоевропейские > индоиранские > иранские > западно-иранские	315	

Парфянский ⁴⁶	Индоевропейские > иранские > северо-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	53	м
Пашаи	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	257	
Персидский	Индоевропейские > иранские > юго-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	332	
Печенежский	Алтайские > тюркские		111	м
Пиктский	Индоевропейские > кельтские		47	м
Полабский	Индоевропейские > славянские > западнославянские > лехитские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > западнославянские	340	м
Половецкий	Алтайские > тюркские > кыпчакские > кыпчакско-половецкие		245	м, к
Польский	Индоевропейские > славянские > западнославянские > лехитские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > западнославянские	363	
Португальский	Индоевропейские > романские > иберо-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	362	к
Прасун	Индоевропейские > нуристанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > нуристанские	211	
Провансальский	Индоевропейские > романские > галло-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	328	
Пхалура	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	205	

⁴⁶ Незавершен

Раджастхани	Индоевропейские, индоарийские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские	300	
Ретороманский	Индоевропейские, романские, галло-романские	Ностратические > индоевропейские > собственно-индоевропейские > италийские > романские > западнороманские	334	
Ровшорвский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские > шугнано-рушанские		259	
Ротума	Австронезийские > восточно-малайско-полинезийские > восточно-океанийские	Австронезийские > малайско-полинезийские > центрально-восточные > восточно-малайско-полинезийские > океанийские > языки отдаленной Океании > центрально-тихоокеанские > Фиджи-ротума	253	
Румынский	Индоевропейские > романские > балкано-романские	Ностратические > индоевропейские > собственно-индоевропейские > италийские > романские > балканские	363	
Русский	Индоевропейские > славянские > восточнославянские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > восточнославянские	526	
Рутульский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские > западно-лезгинские	303	
Рушанский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские > шугнано-рушанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские > шугнано-рушанские	340	
Саамский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > финно-саамские	340	
Саларский	Алтайские > тюркские > огузские	Ностратические > алтайские > тюркские > огузские	373	
Сангличский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	331	
Саргулямский	Индоевропейские > иранские > восточные иранские		53	М

Сардинский	Индоевропейские > романские > островные	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > сардинский	335	
Сарыг-югурский	Алтайские > тюркские > хакасско-алтайские	Ностратические > алтайские > тюркские > сибирские	323	
Сарыкольский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские > шугнано-рушанские		247	
Сванский	Кавказские > картвельские	Ностратические > картвельские	297	
Селькупский	Уральские > самодийские	Ностратические > уральские > самодийские > южно-самодийские	373	
Серболужицкий	Индоевропейские > славянские > западнославянские > серболужицкие		393	
Сербскохорватский	Индоевропейские > славянские > южнославянские > юго-западно-славянские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > южнославянские	389	
Сингальский	Индоевропейские > индоарийские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские	401	
Сиреникских эскимосов	Палеоазиатские > эскимосско-алеутские > эскимосские > сиреникский	Ностратические > эскалеутские > эскимосские	241	
Скифский	Индоевропейские > иранские > восточные иранские > скифские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	126	м
Словацкий	Индоевропейские > славянские > западнославянские > чешско-словацкие	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > западнославянские	396	
Словенский	Индоевропейские > славянские > южнославянские > юго-западно-славянские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > южнославянские	359	
Согдийский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	394	м
Солонский	Алтайские > тунгусо-маньчжурские > сибирские	Ностратические > алтайские > тунгусо-маньчжурские > северные	338	

Сонкорско-тюркский	Алтайские > тюркские > огузские		105	
Среднеогузский	Алтайские > тюркские > огузские		243	м, к
Среднеперсидский	Индоевропейские > иранские > юго-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	287	м
Староанатолийско-тюркский	Алтайские > тюркские > огузские		300	м
Старописьменный монгольский	Алтайские > монгольские	Ностратические > алтайские > монгольские > северные	307	м
Старославянский	Индоевропейские > славянские > южно-славянские > болгаро-македонские		343	м
Старофранцузский	Индоевропейские > романские > галло-романские		333	м
Старояпонский	Алтайские > японо-рюкюские	Ностратические > алтайские	339	м
Табасаранский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские > восточно-лезгинские	294	
Таджикский	Индоевропейские > иранские > юго-западные > иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	341	
Татарский	Алтайские > тюркские > кыпчакские > центральнокыпчакские	Ностратические > алтайские > тюркские > кыпчакские	349	
Татский	Индоевропейские > иранские > юго-западные > иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	339	
Тиграйский	Афразийские > семитские > западносемитские > эфиосемитские	Афразийские > семитские > периферийные > северные	277	к
Тиндинский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	292	
Тирахи	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	182	

Торвали	Индоевропейские > дардские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > дардские	201	
Тофаларский	Алтайские > тюркские > тоба	Ностратические > алтайские > тюркские > сибирские	288	
Трегами	Индоевропейские > нуристанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > нуристанские	57	
Тувинский	Алтайские > тюркские > тоба	Ностратические > алтайские > тюркские > сибирские	323	
Турецкий	Алтайские > тюркские > огузские	Ностратические > алтайские > тюркские > огузские	378	
Туркменский	Алтайские > тюркские > огузские	Ностратические > алтайские > тюркские > огузские	328	
Тюрки	Алтайские > тюркские		208	
Убыхский	Кавказские > абхазо-адыгские	Синокавказские > северно-кавказские > абхазо-адыгские	314	
Удинский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские	300	
Удэгейский	Алтайские > тунгусо-маньчжурские > амурские	Ностратические > алтайские > тунгусо-маньчжурские > южные	352	
Узбекский	Алтайские > тюркские > карлукские	Ностратические > алтайские > тюркские > карлукские	300	
Уйгурский	Алтайские > тюркские > карлукские	Ностратические > алтайские > тюркские > карлукские	324	
Украинский	Индоевропейские > славянские > восточнославянские	Ностратические > индоевропейские > собственно-индоевропейские > балто-славянские > славянские > восточнославянские	396	
Ульчский	Алтайские > тунгусо-маньчжурские > амурские	Ностратические > алтайские > тунгусо-маньчжурские > южные	361	
Умбрский	Индоевропейские > итальянские > оскско-умбрские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > древние	280	м
Урартский	Хуррито-урартские	Синокавказские > северно-кавказские > хуррито-урартские	321	м

Урду	Индоевропейские > индоарийские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские	400	
Урумский	Алтайские > тюркские > огузские		171	
Фалисский	Индоевропейские > итальянские > латино-фалисские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > древние	225	м
Фарерский	Индоевропейские > германские > скандинавские	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	315	
Фарса диалекты	Индоевропейские > иранские > юго-западные иранские		299	
Финский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > финно-саамские > прибалтийско-финские	385	
Франкопровансальский	Индоевропейские > романские > галло-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	272	
Французский	Индоевропейские > романские > галло-романские	Ностратические > индоевропейские > собственно-индоевропейские > итальянские > романские > западнороманские	338	
Фризский	Индоевропейские > германские > западногерманские > англо-фризские	Ностратические > индоевропейские > собственно-индоевропейские > германские > северогерманские	315	к
Фриульский	Индоевропейские > романские > галло-романские		290	
Фуйюйских кыргызов	Алтайские > тюркские > хакасско-алтайские > хакасские		141	
Хазара	Индоевропейские > иранские > юго-западные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > западно-иранские	317	
Хакасский	Алтайские > тюркские > хакасско-алтайские > хакасские	Ностратические > алтайские > тюркские > сибирские	339	
Халаджский	Алтайские > тюркские > огузские	Ностратические > алтайские > халаджский	247	

Хантыйский	Уральские > финно-угорские > угорские > обско-угорские	Ностратические > уральские > финно-угорские > угорские > обско-угорские	367	
Хваршинский	Кавказские > нахско-дагестанские > аваро-андо-цезские > цезские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > цезские > западно-цезские	329	
Хеттский	Индоевропейские > анатолийские	Ностратические > индоевропейские > анатолийские	333	м
Хиналугский	Кавказские > нахско-дагестанские > хиналугский	Синокавказские > северно-кавказские > нахско-дагестанские	271	
Хинди	Индоевропейские > индоарийские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские	387	
Хорасано-тюркский	Алтайские > тюркские > огузские		139	
Хорезмско-тюркский	Алтайские > тюркские		291	м
Хотаносакский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	289	м
Хурритский	Хуррито-урартские	Синокавказские > северно-кавказские > хуррито-урартские	327	м
Хуфский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские > шугнано-рушанские		275	
Цахурский	Кавказские > нахско-дагестанские > лезгинские	Синокавказские > северно-кавказские > нахско-дагестанские > лезгинские > западно-лезгинские	294	
Цезский	Кавказские > нахско-дагестанские > аваро-андо-цезские > цезские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > цезские > западно-цезские	365	
Чагатайский	Алтайские > тюркские > карлукские	Ностратические > алтайские > тюркские > карлукские	214	м
Чамалинский	Кавказские > нахско-дагестанские > аваро-андо-цезские > андийские	Синокавказские > северно-кавказские > нахско-дагестанские > аваро-андо-цезские > аваро-андийские > андийские	304	
Чар-аймаков диалектная группа	Индоевропейские > иранские > юго-западные иранские		40	

Чеченский	Кавказские > нахско-дагестанские > Нахские	Синокавказские > северно-кавказские > нахско-дагестанские > нахские	361	
Чешский ⁴⁷	Индоевропейские > славянские > западнославянские > чешско-словацкие	Ностратические > индоевропейские > собственноиндоевропейские > балто-славянские > славянские > западнославянские	409	
Чувашский	Алтайские > тюркские > болгарские	Ностратические > алтайские > тюркские > болгарские	261	
Чукотский	Палеоазиатские > чукотско-камчатские	Чукотско-камчатские > чукотско-корякские	292	
Чулымско-тюркский	Алтайские > тюркские > хакасско-алтайские > хакасские		277	
Шведский	Индоевропейские > германские > скандинавские	Ностратические > индоевропейские > собственноиндоевропейские > германские > северогерманские	374	
Шина	Индоевропейские > дардские	Ностратические > индоевропейские > собственноиндоевропейские > индоиранские > иранские > дардские	322	
Шира югуров	Алтайские > монгольские		243	
Шорский	Алтайские > тюркские > хакасско-алтайские > хакасские	Ностратические > алтайские > тюркские > сибирские	314	
Шотландский	Индоевропейские > кельтские > гойдельские		177	
Шугнанский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские > шугнано-рушанские	Ностратические > индоевропейские > собственноиндоевропейские > индоиранские > иранские > восточно-иранские	396	
Шумашти	Индоевропейские > дардские	Ностратические > индоевропейские > собственноиндоевропейские > индоиранские > иранские > дардские	182	
Шумерский	Изолят	Изолят	261	м
Эвенкийский	Алтайские > тунгусо-маньчжурские > сибирские	Ностратические > алтайские > тунгусо-маньчжурские > северные	402	

⁴⁷ Незавершен

Эвенский	Алтайские > тунгусо-маньчжурские > сибирские	Ностратические > алтайские > тунгусо-маньчжурские > северные	389	
Эламский	Изолят	Изолят	225	м
Энецкий	Уральские > самодийские	Ностратические > уральские > самодийские > южно-самодийские	314	
Эрзянский	Уральские > финно-угорские > финно-пермские > мордовские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > мордовские	346	
Эскимосов Берингова пролива	Палеоазиатские > эскимосско-алеутские > эскимосские > инуитские	Ностратические > эскалеутские > эскимосские > группа инуит	237	
Эстонский	Уральские > финно-угорские > финно-пермские > балтийско-финские	Ностратические > уральские > финно-угорские > финно-пермские > финно-волжские > финно-саамские > прибалтийско-финские	337	
Югский	Палеоазиатские > енисейские		307	м
Юкагирский	Палеоазиатские > юкагиро-чуванские	Изолят	350	
Юртовских татар	Алтайские > тюркские > кыпчакские > кыпчакско-ногайские		301	
Ягнобский	Индоевропейские > иранские > восточные иранские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	364	
Язгулямский	Индоевропейские > иранские > восточные иранские > шугнано-язгулямские > шугнано-рушанские	Ностратические > индоевропейские > собственно-индоевропейские > индоиранские > иранские > восточно-иранские	368	
Якутский	Алтайские > тюркские > якутские	Ностратические > алтайские > тюркские > якутские	372	

Примечание:

м - мертвый

к - язык межплеменного или междиалектного общения, койне.

Перечень классифицирующих признаков, представленных в БД

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
1	2.1.1.		2.1.1. ФОНЕМНЫЙ СОСТАВ
2	2.1.1.	13	.О
3	2.1.1.	303	.гласные
4	2.1.1.	2	.О
5	2.1.1.	293	..подъем
15	2.1.1.	299	..ряд
41	2.1.1.	292	..общая схема монофтонгов
42	2.1.1.	292	...различия по ряду
43	2.1.1.	291в нижнем подъеме
49	2.1.1.	290в верхнем подъеме
56	2.1.1.	187	...различия по лабиализации
65	2.1.1.	1	...различия по напряженности
67	2.1.1.	12	...различия по назализации
71	2.1.1.	3	...различия по фарингализации
73	2.1.1.	87	..различие схем
83	2.1.1.	283	..число монофтонгов
109	2.1.1.	94	.дифтонги
117	2.1.1.	78	.общее число дифтонгов и трифтонгов
141	2.1.1.	302	.согласные
142	2.1.1.	301	..шумные
143	2.1.1.	296ларингальные признаки
176	2.1.1.	1губные
178	2.1.1.	300	...способ образования
182	2.1.1.	302	...локальный ряд
183	2.1.1.	298губные
187	2.1.1.	302переднеязычные
188	2.1.1.	301смычные неаффрицированные
189	2.1.1.	215единый ряд переднеязычных
203	2.1.1.	300фрикативные и аффрикаты
230	2.1.1.	299среднеязычные
236	2.1.1.	299заднеязычные
240	2.1.1.	285постувулярные
245	2.1.1.	144	...дополнительные признаки
247	2.1.1.	8импловзивные/эксплозивные
252	2.1.1.	31лабиализованные/нелабиализованные
269	2.1.1.	2назализованные/неназализованные
276	2.1.1.	60палатализованные/непалатализованные
293	2.1.1.	1твердые/мягкие
295	2.1.1.	1палатализ./полупалатализ./непалатализ.
297	2.1.1.	16простые/геминированные
305	2.1.1.	2ретрофлексные/неретрофлексные
308	2.1.1.	1сильные/слабые
310	2.1.1.	10фарингализованные/нефарингализованные
314	2.1.1.	26церебральные/нецеребральные

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
320	2.1.1.	6эмфатизованные/неэмфатизованные
326	2.1.1.	2однофокусные/двухфокусные
328	2.1.1.	299	...сонорные
329	2.1.1.	297способ образования
337	2.1.1.	299локальный ряд
338	2.1.1.	295назальные
350	2.1.1.	293плавные
352	2.1.1.	229латеральные
362	2.1.1.	286вибранты
383	2.1.1.	114дополнительные признаки
384	2.1.1.	8глухие/звонкие
391	2.1.1.	1дентальные/ретрофлексные
393	2.1.1.	52палатализованные/непалатализованные
400	2.1.1.	1палатализованные/полупалатализованные/непалатализованные
402	2.1.1.	6лабиализованные/нелабиализованные
407	2.1.1.	5интенсивные/неинтенсивные
409	2.1.1.	10ретрофлексные/неретрофлексные
413	2.1.1.	11фарингализованные/нефарингализованные
419	2.1.1.	22церебральные/нецеребральные
425	2.1.1.	7придыхательные/непридыхательные
428	2.1.1.	15простые/геминированные
435	2.1.1.	1простые/геминированные
438	2.1.2.		2.1.2. ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ
440	2.1.2.	261	.ударение
442	2.1.2.	195	..тип
445	2.1.2.	234	..вид
451	2.1.2.	244	..носитель ударения
456	2.1.2.	241	..фиксированность
458	2.1.2.	175	...связанное
479	2.1.2.	39	...подвижное
480	2.1.2.	76	..главное/второстепенное ударение
482	2.1.2.	30	.тон
483	2.1.2.	21	..число тонов
487	2.1.2.	12	..регистровые признаки
490	2.1.2.	19	..контурные признаки
507	2.1.2.	23	..носитель тона
511	2.1.2.	196	.долгота
512	2.1.2.	192	..характер долготы
517	2.1.2.	195	..носитель долготы
519	2.1.2.	30	...слог
522	2.1.2.	102	.сингармонизм
523	2.1.2.	98	..область гармонического уподобления
529	2.1.2.	98	..признак гармонического уподобления
538	2.1.2.	79	..уподобляющий элемент
551	2.1.2.	17	.прочие просодические явления
555	2.1.2.	4	..ларингализация

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
562	2.1.2.	4	..фонация
568	2.1.3.		2.1.3. ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ
569	2.1.3.	49	.О
570	2.1.3.	222	.гласные
571	2.1.3.	162	..структурные позиции
592	2.1.3.	91	...редукция
603	2.1.3.	173	..комбинаторные позиции
608	2.1.3.	47	...ассимиляция
617	2.1.3.	2	...диссимиляция
629	2.1.3.	15	..дифтонги
637	2.1.3.	191	..фонологическое содержание чередований
640	2.1.3.	244	.согласные
641	2.1.3.	3	..О
642	2.1.3.	161	..структурные позиции
681	2.1.3.	216	..комбинаторные позиции
685	2.1.3.	174	...ассимиляция
694	2.1.3.	21	...диссимиляция
717	2.1.3.	222	..фонологическое содержание чередований
720	2.1.3.	5	.тоны
721	2.1.3.	4	..структурные позиции
726	2.1.3.	1	..комбинаторные позиции
728	2.1.3.	0	...ассимиляция
732	2.1.3.	5	..фонологическое содержание чередований
735	2.1.4.		2.1.4. СЛОГ
736	2.1.4.	46	.О
737	2.1.4.	3	.открытые/закрытые слоги
738	2.1.4.	261	.элемент, предшествующий гласному
743	2.1.4.	260	..состоит, как правило, из
745	2.1.4.	66	...одного или двух согласных
748	2.1.4.	38	...от одного до трех согласных
751	2.1.4.	1	.медиаль
752	2.1.4.	2	..состоит, как правило, из
754	2.1.4.	262	.элемент, следующий за гласным
759	2.1.4.	256	..состоит, как правило, из
761	2.1.4.	140	...одного или двух согласных
765	2.1.4.	20	...от одного до трех согласных
768	2.1.4.	2	.слабый слог
772	2.1.4.	24	.прочие запрещения
784	2.1.4.	1	..исход слога внутри слова
786	2.1.4.	3	..начало слога
800	2.1.4.	1	.тенденция
802	2.1.4.	247	.конкретная (или преобладающая) структура слога
889	2.1.4.	4	.состоит из мор
892	2.1.4.	1	..конкретная структура моры
894	2.1.4.	0	...начальной

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
896	2.1.4.	0	...конечной
898	2.1.4.	1	.правило слогового равновесия
899	2.2.1.		2.2.1. ФОНОЛОГИЧЕСКАЯ СТРУКТУРА
900	2.2.1.	102	.О
901	2.2.1.	49	.фонологическая структура морфемы
904	2.2.1.	1	..глагольные корни не содержат гласных
909	2.2.1.	11	...ограничения в зависимости от морфологической категории
910	2.2.1.	1	...корневые морфемы
912	2.2.1.	5	..морфологическая структура
913	2.2.1.	4	...ограничения в зависимости от морфологической категории
914	2.2.1.	3	...корневые морфемы
920	2.2.1.	26	..ограничения на фонемную структуру сегм. морфем
921	2.2.1.	18	...начальная позиция
931	2.2.1.	0	...срединная позиция
933	2.2.1.	18	...конечная позиция
937	2.2.1.	2	...как правило, согласные
939	2.2.1.	1	...как правило, гласные
948	2.2.1.	4	...корневые морфемы
956	2.2.1.	4	...служебные морфемы
961	2.2.1.	200	.фонологическая структура слова
962	2.2.1.	3	..особенности слогового состава слова
967	2.2.1.	172	..ограничения на фонемную структуру слова
969	2.2.1.	115	...начальная позиция
980	2.2.1.	9	...срединная позиция
986	2.2.1.	95	...конечная позиция
1001	2.2.1.	137	...стыковые явления
1030	2.2.2.		2.2.2. ФОНОЛОГИЧЕСКИЕ ПРОТИВОПОСТАВЛЕНИЯ МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ
1031	2.2.2.	122	.О
1032	2.2.2.	46	.знаменательные/служебные слова
1039	2.2.2.	64	.исконные/заимствованные слова
1045	2.2.2.	13	.именные/глагольные корни
1051	2.2.2.	72	.корневые/аффиксальные морфемы
1059	2.2.3.		2.2.3. МОРФОЛОГИЧЕСКИ ОБУСЛОВЛЕННЫЕ ЧЕРЕДОВАНИЯ
1060	2.2.3.	91	.О
1061	2.2.3.	15	.отсутствие чередований
1062	2.2.3.	209	.функциональный тип
1064	2.2.3.	138	..смыслоразличительные
1065	2.2.3.	79	...словообразовательные
1068	2.2.3.	119	...словоизменительные
1071	2.2.3.	210	.вид чередований
1072	2.2.3.	175	..вокалические
1077	2.2.3.	6	..тоновые

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
1081	2.3.0.		2.3.0. MORFOЛОГИЧЕСКИЙ ТИП ЯЗЫКА
1082	2.3.0.	5	.О
1083	2.3.0.	303	.способ соединения морфем в слове
1085	2.3.0.	54	..агглютинативные языки с элементами флексии
1091	2.3.0.	25	..флективные языки с элементами агглютинации
1097	2.3.0.	25	..флективно-агглютинативные языки
1098	2.3.0.	13	...агглютинация
1102	2.3.0.	16	...флексия
1105	2.3.0.	25	..признаки флективного строя
1111	2.3.0.	18	..признаки агглютинативного строя
1117	2.3.0.	265	.степень спаянности морфем
1120	2.3.0.	17	..аналитические языки с элементами синтетизма
1123	2.3.0.	5	..аналитико-синтетические языки
1124	2.3.0.	2	...элементы аналитизма
1126	2.3.0.	2	...элементы синтетизма
1132	2.3.0.	110	..синтетические языки с элементами аналитизма
1135	2.3.0.	7	.инкорпорация
1137	2.3.0.	7	..языки с элементами инкорпорации
1139	2.3.1.		2.3.1. КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ
1140	2.3.1.	51	.О
1142	2.3.1.	262	.семантические и формальные
1146	2.3.2.		2.3.2. ИМЕННЫЕ КЛАССИФИКАЦИИ
1147	2.3.2.	16	.О
1149	2.3.2.	161	.согласовательные классы
1150	2.3.2.	122	..род
1154	2.3.2.	156	..способ выражения согласовательных классов
1156	2.3.2.	27	...лексический
1162	2.3.2.	108	...морфологический
1164	2.3.2.	22аффиксация
1169	2.3.2.	134	...синтаксический
1184	2.3.2.	141	..мотивация согласовательных классов
1215	2.3.2.	61	.личность/неличность
1220	2.3.2.	48	.одушевленность/неодушевленность
1225	2.3.2.	1	.референтность/нереферентность
1230	2.3.2.	3	.собирательность/несобирательность
1240	2.3.3.		2.3.3. ЧИСЛО
1241	2.3.3.	8	.О
1243	2.3.3.	302	.структура числа
1255	2.3.3.	298	.способы выражения числа
1268	2.3.3.	23	..служебные слова
1287	2.3.3.	49	..особое выражение собирательности

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
1290	2.3.3.	2	.гонорифические формы числа
1293	2.3.3.	224	.согласование по числу
1296	2.3.3.	213	.система счисления
1297	2.3.3.	178	..десятеричная
1302	2.3.3.	15	..двадцатеричная
1308	2.3.3.	132	.классы числительных, отличные от порядковых и количественных
1351	2.3.4.		2.3.4. ПАДЕЖНЫЕ ЗНАЧЕНИЯ
1352	2.3.4.	12	.О
1353	2.3.4.	295	.актантные падежные значения
1354	2.3.4.	280	..субъектно-объектные
1396	2.3.4.	119	..различное падежное оформление
1397	2.3.4.	60	...субъекта
1409	2.3.4.	110	...объекта
1426	2.3.4.	291	..способ выражения субъектно-объектных значений
1433	2.3.4.	106	...служебные слова
1437	2.3.4.	200	..прочие актантные падежи
1473	2.3.4.	55	.падежное оформление именного сказуемого
1492	2.3.4.	165	.падежное оформление атрибутивного отношения
1518	2.3.4.	284	.падежное оформление посессивного отношения
1535	2.3.4.	24	..особенности в оформлении
1551	2.3.4.	19	..притяжательное прилагательное
1559	2.3.4.	253	.падежное оформление локативных отношений
1560	2.3.4.	251	..способ выражения
1561	2.3.4.	196	...аффиксы именные
1573	2.3.4.	189	..локативные формы имени
1681	2.3.4.	144	.обстоятельственные падежи
1748	2.3.4.	19	.возможность наложения двух падежных форм
1757	2.3.5.		2.3.5. ГЛАГОЛЬНЫЕ КАТЕГОРИИ
1758	2.3.5.	11	.О
1759	2.3.5.	253	.залог
1760	2.3.5.	244	..категориальный состав кроме активного
1777	2.3.5.	183	...пассив
1792	2.3.5.	243	..способ выражения
1797	2.3.5.	5	..двойные залоговые формы
1803	2.3.5.	17	..совпадение форм
1814	2.3.5.	303	.видо-временные категории
1815	2.3.5.	294	..основные временные противопоставления
1827	2.3.5.	59	..дополнительные временные противопоставления
1830	2.3.5.	155	..основные видовые противопоставления
1846	2.3.5.	133	..выражение вида и времени
1849	2.3.5.	1	..виды будущего времени

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
1853	2.3.5.	1	..виды будущего времени
1855	2.3.5.	99	..виды настоящего времени
1897	2.3.5.	10	..виды настояще-будущего времени
1904	2.3.5.	242	..виды прошедшего времени
2081	2.3.5.	278	..способ выражения
2086	2.3.5.	144	.способ действия
2214	2.3.5.	139	..способ выражения
2217	2.3.5.	291	.наклонения кроме изъявительного
2218	2.3.5.	291	..основные значения
2219	2.3.5.	281	...повелительное наклонение
2246	2.3.5.	266	...косвенные наклонения
2297	2.3.5.	24потенциалис
2337	2.3.5.	277	..способ выражения
2340	2.3.5.	103	.модальность
2341	2.3.5.	101	..основные значения
2435	2.3.5.	98	..способ выражения
2439	2.3.5.	60	.синкретичное выражение нескольких глагольных значений одной категорией
2449	2.3.5.	206	.переходность
2450	2.3.5.	199	..маркер переходности
2459	2.3.5.	24	..лабильные глаголы
2467	2.3.5.	15	.версия
2468	2.3.5.	15	..основные значения
2477	2.3.5.	14	..способ выражения
2480	2.3.5.	11	.порода
2481	2.3.5.	11	..основные значения
2501	2.3.5.	10	..способ выражения
2504	2.3.5.	1	.глагольный падеж
2505	2.3.5.	1	..основные значения
2520	2.3.6.		2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ
2521	2.3.6.	12	.О
2522	2.3.6.	299	.местоимения и местоименные слова
2523	2.3.6.	296	..разряды местоимений
2532	2.3.6.	124	...возвратные
2550	2.3.6.	292	...личные
2572	2.3.6.	86	...определительные
2583	2.3.6.	107	...притяжательные
2613	2.3.6.	240	.дейктический статус
2614	2.3.6.	233	..указательные местоимения
2630	2.3.6.	2	...личные местоимения 3 лица
2632	2.3.6.	3	..глагольные превербы
2636	2.3.6.	10	..именные аффиксы
2648	2.3.6.	10	..местоименные наречия
2659	2.3.6.	9	..послелогии
2669	2.3.6.	3	..пространственные наречия
2673	2.3.6.	3	..темпоральные наречия
2677	2.3.6.	2	..наречия образа действия и прочие

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
2680	2.3.6.	8	..предлоги
2690	2.3.6.	2	..частица
2692	2.3.6.	277	.грамматическая категория лица
2694	2.3.6.	168	..в имени
2699	2.3.6.	269	..в глаголе
2702	2.3.6.	225	...одноличное спряжение
2708	2.3.6.	15	.эмфаза
2709	2.3.6.	10	..в имени
2713	2.3.6.	3	..в указательных местоимениях
2716	2.3.6.	261	.пространственная ориентация действия
2717	2.3.6.	260	..выражение
2720	2.3.6.	26	...глагольные превербы
2740	2.3.6.	199	.определенность/неопределенность имени
2742	2.3.6.	81	..артикли
2756	2.3.6.	125	..синтаксическое выражение
2772	2.3.6.	34	.формы вежливости
2773	2.3.6.	9	..в именной группе
2779	2.3.6.	17	..в глагольной группе
2780	2.3.6.	15	...адрессив
2783	2.3.6.	3	...гоноратив
2789	2.3.6.	3	..в сентенциальном синтаксисе
2792	2.3.6.	3	.фокус
2793	2.3.6.	2	..сфера применения
2796	2.3.6.	3	..способ выражения
2798	2.3.6.	2	...служебные слова
2801	2.3.6.	249	.отрицание
2805	2.3.6.	181	..лексическое
2807	2.3.6.	169	..морфологическое
2811	2.3.6.	236	..позиция отрицательного показателя
2816	2.3.6.	1	.единичность/множественность
2818	2.3.6.	1	.единичность/общность
2820	2.3.6.	1	.контрастивность
2821	2.3.6.	1	..в имени
2823	2.3.6.	1	..в глаголе
2825	2.3.6.	2	.референтность/нереферентность имени
2828	2.3.7.		2.3.7. ЧАСТИ РЕЧИ
2829	2.3.7.	6	.О
2831	2.3.7.	303	.знаменательные
2834	2.3.7.	308	..глагол
2836	2.3.7.	267	...нефинитные формы
2861	2.3.7.	3	..качественный глагол
2863	2.3.7.	1	...нефинитные формы
2877	2.3.7.	297	.служебные
2902	2.4.0.		2.4.0. ПАРАДИГМЫ
2903	2.4.0.	14	.О
2904	2.4.0.	1	.адъектив
2905	2.4.0.	1	..словоизменение

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
2907	2.4.0.	66	.артиклъ
2917	2.4.0.	298	.глагол
2918	2.4.0.	297	..финитные формы
2919	2.4.0.	295	...словоизменение
2953	2.4.0.	92	...типы спряжения
2964	2.4.0.	114	...синкретичное выражение нескольких категорий одним аффиксом
2989	2.4.0.	192	..нефинитные формы
2990	2.4.0.	3	...абсолютив
2995	2.4.0.	2	...глагольное имя
3000	2.4.0.	48	...инфинитив
3023	2.4.0.	3	...герундий
3027	2.4.0.	79	...деепричастие
3101	2.4.0.	6	...имя действия
3106	2.4.0.	2принадлежность
3108	2.4.0.	3	...масдар
3113	2.4.0.	179	...причастие
3183	2.4.0.	1	...причастно-деепричастная форма
3185	2.4.0.	3	...супин
3191	2.4.0.	9	.имя
3192	2.4.0.	9	..словоизменение
3201	2.4.0.	3	..классифицирующие категории
3203	2.4.0.	1	..синкретическое выражение
3205	2.4.0.	3	.качественный глагол
3206	2.4.0.	3	..финитные формы
3207	2.4.0.	3	...словоизменение
3210	2.4.0.	2смысловые группы
3219	2.4.0.	1	..нефинитные формы
3220	2.4.0.	1	...деепричастие
3227	2.4.0.	96	.местоимение
3229	2.4.0.	1	.местоименный предлог
3233	2.4.0.	53	.наречие
3243	2.4.0.	3	.отрицание
3249	2.4.0.	1	.отрицательная частица
3251	2.4.0.	17	.послелог
3258	2.4.0.	3	.предикатив
3266	2.4.0.	219	.прилагательное
3267	2.4.0.	216	..словоизменение
3281	2.4.0.	7	...предикативность/атрибутивность
3288	2.4.0.	1	...репрезентация
3303	2.4.0.	22	..синкретическое выражение
3310	2.4.0.	1	.служебное имя
3311	2.4.0.	1	..словоизменение
3314	2.4.0.	291	.существительное
3315	2.4.0.	286	..словоизменение
3338	2.4.0.	21	...предикативность/непредикативность
3345	2.4.0.	148	...принадлежность

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
3350	2.4.0.	1	...репрезентация
3369	2.4.0.	157	..классифицирующие категории
3377	2.4.0.	54	..синкретическое выражение
3391	2.4.0.	7	.частица
3398	2.4.0.	52	.числительное
3400	2.4.0.	5	.междометие
3404	2.4.0.	4	.предлог
3409	2.5.1.		2.5.1. СТРУКТУРА СЛОВОФОРМЫ
3410	2.5.1.	32	.О
3413	2.5.1.	281	.типы морфем, присоединяемых к корневым
3416	2.5.1.	273	.средства словоизменения
3424	2.5.1.	100	..префиксы
3429	2.5.1.	271	..суффиксы
3436	2.5.1.	1	.возможность перестановки морфем
3439	2.5.1.	220	.модели словоформы
3441	2.5.1.	8	..преимущественно префиксальная
3443	2.5.1.	216	..преимущественно суффиксальная
3445	2.5.1.	142	.конкретная модель словоформы
3449	2.5.1.	113	..глагольная
3524	2.5.1.	119	..именная
3581	2.5.1.	1	..идеофоны
3583	2.5.1.	1	.модель квазисловоформы для изолирующих языков
3584	2.5.1.	1	..глагольная
3588	2.5.1.	1	..именная
3592	2.5.1.	1	..основа+афф. мн. ч.+падежный афф.+лично-притяжный афф.
3593	2.5.2.		2.5.2. СЛОВООБРАЗОВАНИЕ
3594	2.5.2.	23	.О
3595	2.5.2.	290	.деривация
3596	2.5.2.	290	..аффиксация
3610	2.5.2.	14	..несовпадение типов деривации
3619	2.5.2.	65	.редупликация
3622	2.5.2.	219	.сложение основ
3625	2.5.2.	3	..синтаксич. отношения между соединяемыми основами
3633	2.5.3.		2.5.3. ПРОСТОЕ ПРЕДЛОЖЕНИЕ
3634	2.5.3.	16	.О
3635	2.5.3.	271	.строй
3639	2.5.3.	198	..номинативный
3643	2.5.3.	6	..смешанный
3648	2.5.3.	44	..эргативный
3652	2.5.3.	282	.линейный порядок членов предложения
3655	2.5.3.	263	..основной
3662	2.5.3.	231	.линейный порядок в группах
3663	2.5.3.	218	..в именной группе
3674	2.5.3.	101	..в глагольной группе

Порядковый номер в БД	Номер раздела	Число языков	Наименование и уровень классифицирующего признака
3682	2.5.3.	56	.маргинальные конструкции
3711	2.5.3.	208	.согласование
3712	2.5.3.	193	..актант определяет форму предиката
3713	2.5.3.	184	...субъектное согласование
3721	2.5.3.	46	...объектное согласование
3727	2.5.3.	9	...субъектно-объектное согласование
3733	2.5.3.	4	..предикат определяет форму актанта
3734	2.5.3.	0	...объектное согласование
3736	2.5.3.	3	...субъектное согласование
3740	2.5.3.	1	...субъектно-объектное согласование
3742	2.5.3.	95	..атрибутивное согласование
3752	2.5.3.	183	.коммуникативно значимые средства
3761	2.5.3.	6	.инкорпорация
3763	2.5.3.	5	..синтаксические отношения между компонентами
3768	2.5.3.	1	..класс инкорпорируемой основы
3773	2.5.4.		2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ
3774	2.5.4.	41	.О
3775	2.5.4.	40	.линейный порядок компонентов
3781	2.5.4.	144	.особенности подчиненного компонента
3785	2.5.4.	137	..оформление сказуемого
3796	2.5.4.	0	.особенности построения дискурса
3799	2.5.4.	14	.структура относительного предложения
3807	2.5.4.	259	.тип построения
3812	2.5.4.	264	.тип связи
3816	2.5.4.	223	..союзы
3821	2.5.4.	2	...союзные формы глагола

Реферат шведского языка

- 0.0.0. *НАЗВАНИЕ ЯЗЫКА
 - .шведский
- 1.1.0. *СОКРАЩЕННОЕ НАЗВАНИЕ ЯЗЫКА
 - .швед.
- 1.1.1. ВАРИАНТЫ НАЗВАНИЯ
 - .0
- 1.1.2. ГЕНЕТИЧЕСКАЯ ПРИНАДЛЕЖНОСТЬ
 - .шведский<скандинавские< германские<индоевропейские
- 1.1.3. РАСПРОСТРАНЕНИЕ
 - .Швеция (госуд.)
 - .Финляндия (госуд.)
 - .Аландские острова
 - .Соединенные Штаты Америки (я.п.м.)
 - .Канада (я.п.м.)
 - .число говорящих
 - ..ок. 9 млн.чел.
- 1.2.1. ДИАЛЕКТНЫЙ СОСТАВ
 - .южношведские диал.
 - .етские диал.
 - .свейские диал.
 - ..центральношведские диал.
 - ..верхнешведские диал.
 - .норландские диал.
 - .Готланд острова диал.
 - .восточношведские диал.
 - .тенденция
 - ..диалектный континуум
- 1.3.1. СТАТУС ЯЗЫКА
 - .государственный
 - ..Швеция
 - ..Финляндия
 - .разговорно-обиходный
 - ..Аландские острова
- 1.3.2. ЛИТЕРАТУРНЫЙ ЯЗЫК
 - .результат смешения диалектов
 - .две литературные нормы
- 1.3.3. ПРЕПОДАВАНИЕ
 - .язык преподавания
 - ..начальная школа
 - ..средняя школа
- ..высшая школа
- 1.4.0. ПИСЬМО
 - .тип письма
 - ..алфавитное письмо
 - ...латинский алфавит
 -с диакритиками
- 1.5.0. ИСТОРИЧЕСКИЕ ПЕРИОДЫ
 - .древнешведский (ок. 800 г. - до 1525 г.)
 - ..рунический (ок. 800 г. - ок. 1225 г.)
 - ..классический (ок. 1225 г. - ок. 1375 г.)
 - ..поздний древнешведский (ок. 1375 г. - 1525 г.)
 - .новошведский (1525 г. - н.вр.)
 - ..ранний новошведский (1525 г. - 1732 г.)
 - ..современный (1732 г. - н.вр.)
- 1.6.0. ВНУТРИСТРУКТУРНЫЕ ИЗМЕНЕНИЯ ИЗ-ЗА КОНТАКТОВ
 - .нижненемецкий
 - ..словообразовательные аффиксы
 - .верхненемецкий
 - ..словообразовательные аффиксы
 - .латинский
 - ..словообразовательные аффиксы
 - .греческий
 - ..словообразовательные аффиксы
- 2.6.0. ЛЕКСИЧЕСКИЕ ЗАИМСТВОВАНИЯ
 - .латинский
 - ..термины торговли и строительства
 - ..бытовая лексика
 - ..лексика церковного обихода
 - ..научная терминология
 - .нижненемецкий
 - ..социальные термины

- ..термины ремесел и торговли
- .верхненемецкий
- ..термины горнорудного и военного дела
- ..торговая лексика
- .французский
- ..политическая лексика
- ..бытовая лексика
- ..терминология искусства
- .английский
- ..техническая терминология
- ..экономические термины
- ..спортивная терминология
- ..бытовая лексика
- 2.1.1. ФОНЕМНЫЙ СОСТАВ
- .гласные
- ..подъем
- ...верхний/пониженный
- верхний/средний/пониженный
- ..ряд
- ...передний/средний/задний
- ..долгие/краткие
- ..лабиализованные/нелабиализованные
- ..общая схема монофтонгов
- ..различия по ряду
-в нижнем подъеме
-средний/задний
-в верхнем подъеме
-передний/задний
- ..различия по лабиализации
-в переднем ряду
- ..различие схем
- ...у кратких и долгих
- ..число монофтонгов
- ..восемнадцать
- .согласные
- ..шумные
- ...ларингальные признаки
-звонкие/глухие
- ...способ образования
-фрикативные/смычные
- ...локальный ряд
-губные
-билабиальные/лабиодентальные
-переднеязычные
-смычные
- неаффрицированные

-дентальные/ретрофлексные
-фрикативные и аффрикаты
-свистящие/шипящие/ретрофлексные
-среднеязычные
-единый ряд собственно среднеязычных
-заднеязычные
-единый ряд заднеязычных
-постувулярные
-только фарингальные
- ...дополнительные признаки
-простые/геминированные
-ретрофлексные/неретрофлексные
-переднеязычные
-свистящие
- ...сонорные
-способ образования
-назальные/плавные/вibrанты
-локальный ряд
-назальные
-губные/переднеязычные/заднеязычные
-плавные
-латеральные
-vibrанты
-переднеязычные
- ...дополнительные признаки
-ретрофлексные/неретрофлексные
-простые/геминированные
-назальные/плавные/vibrанты
- 2.1.2. ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ
- .ударение
- ..тип
- ...фонологическое
- ..вид
- ...динамическое
- ..носитель ударения
- ...слог
- ..фиксированность
- ...связанное
-начальный слог
- ..главное/второстепенное ударение

- .тон
- ..число тонов
- ...два
- ..контурные признаки
- ...акутный/грависный
- ..фонологически значимый тон
- ..носитель тона
- ...слово
- .долгота
- ..характер долготы
- ...фонологическая
- ..носитель долготы
- ...фонема

2.1.3. ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ

- .гласные
- ..структурные позиции
- ...редукция
-качественная
-количественная
- ..комбинаторные позиции
- ..расширение
- .согласные
- ..структурные позиции
- ...выпадение
- ...оглушение
- ..комбинаторные позиции
- ...аккомодация
- ...аспирация
- ...ассимиляция
-регрессивная
- ..фонологическое содержание чередований
- ...аллофоническое варьирование

2.1.4. СЛОГ

- .элемент, предшествующий гласному
- ..состоит, как правило, из
- ...от одного до трех согласных
- .элемент, следующий за гласным
- ..состоит, как правило, из
- ...одного или двух согласных
- .прочие запрещения
- ..ударный слог всегда
- содерж. один долгий элемент
- ..нельзя сочет. долгий гл.

- с таутосиллабич. согласным
- ..правило слогового равновесия
- ..ограниченное употребление стечений согласных
- ..CV/CVC/CVCC/CVCCC/CCV/CCVC/CCVCC/CCVCCC/ см.F4

2.2.1. ФОНОЛОГИЧЕСКАЯ СТРУКТУРА

- .фонологическая структура морфемы
- ..ограничения на фонемную структуру сегм. морфем
- ...конечная позиция
-ограниченные комбинации согласных
- .фонологическая структура слова
- ..ограничения на фонемную структуру слова
- ...стыковые явления
-оглушение
-стяжение

2.2.2. ФОНОЛОГИЧЕСКИЕ ПРОТИВОПОСТАВЛЕНИЯ МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ

- .знаменательные/служебные слова
- ..акцентные различия
- .исконные/заимствованные слова
- ..различия в фонетическом составе
- .корневые/аффиксальные морфемы
- ..различия в фонетическом составе

2.2.3. МОРФОЛОГИЧЕСКИ ОБУСЛОВЛЕННЫЕ ЧЕРЕДОВАНИЯ

- .функциональный тип
- ..сопутствующие
- ..смыслоразличительные
- ...словообразовательные
-в имени
-в глаголе
- ..словоизменительные
-в имени
-в глаголе
- .вид чередований

- ..вокалические
- ...количественные
- ...качественные
- 2.3.0. МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА
- .способ соединения морфем в слове
- ..флективные языки
- ..флективные языки с элементами агглютинации
- ..признаки флективного строя
- ...внутренняя флексия
- .степень спаянности морфем
- ..аналитические языки с элементами синтетизма
- ...в имени
- ...в глаголе
- 2.3.1. КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ
- .семантические и формальные
- ..морфолого-синтаксические
- 2.3.2. ИМЕННЫЕ КЛАССИФИКАЦИИ
- .согласовательные классы
- ..род
- ...общий/средний
- ..способ выражения согласовательных классов
- ...морфологический
- ...синтаксический
- ...в артикле
- ...в местоимении
- ...в прилагательных
- ...в причастии
- ..мотивация согласовательных классов
- ...одушевленность
- ...пол
- .одушевленность/неодушевленность
- ..выражение в анафорических местоимениях
- 2.3.3. ЧИСЛО
- .структура числа
- ..единственное/множественное
- .способы выражения числа
- ..аффиксация
- ..значимое чередование
- ..просодические явления
- ..супплетивизм

- .согласование по числу
- ..атрибутивное
- .система счисления
- ..десятеричная
- ...с элементами двенадцатеричной
- 2.3.4. ПАДЕЖНЫЕ ЗНАЧЕНИЯ
- .актантные падежные значения
- ..субъектно-объектные
- ...общий/родительный
- ...субъектный/объектный
- ..различное падежное оформление
- ...объекта
- ...имен и местоимений
- ..способ выражения субъектно-объектных значений
- ...падежные аффиксы
- ...порядок слов
- ...служебные слова
- ...в препозиции
- .падежное оформление атрибутивного отношения
- ..генитив
- ..предложная конструкция
- ..соположение имен
- .падежное оформление посессивного отношения
- ..генитив
- ..предложная конструкция
- ..притяжательное местоимение
- .падежное оформление локативных отношений
- ..способ выражения
- ...местоименные наречия
- ...предлоги
- 2.3.5. ГЛАГОЛЬНЫЕ КАТЕГОРИИ
- .залог
- ..категориальный состав кроме активного
- ...каузатив
- ...медий
- ...пассив
- ...статический/динамический
- ...рефлексив
- ...реципрок
- ..способ выражения
- ..морфологическая категория
- ...служебные слова

- ..совпадение форм
- ...медий/реципрок/пассив
- .видо-временные категории
- ..основные временные противопоставления
- ...настоящее/прошедшее/будущее
- ..дополнительные временные противопоставления
- ...таксис
- ..основные видовые противопоставления
- ...результативный/нерезультативный
- ..выражение вида и времени
- ...раздельное
- ..виды прошедшего времени
- ...претерит/перфект/плюсквамперфект
- ..виды будущего времени
- ...буд.1/буд.2/буд.в прошедшем1/буд.в прошедшем2
- ..способ выражения
- ...морфологическая категория
- ...служебные слова
- .способ действия
- ..основные значения
- ...итеративность
- ...инхоативность
- ...однократность
- .наклонения кроме изъявительного
- ..основные значения
- ...повелительное наклонение
- ...императив
- ...косвенные наклонения
- ...конъюнктив
- ..способ выражения
- ...морфологическая категория
- ...служебные слова
- .модальность
- ..основные значения
- ...возможность
- ...долженствование
- ...намерение
- ...необходимость
- ..способ выражения
- ...служебные слова

- .синкретичное выражение нескольких глагольных значений одной категорией
- ..время и наклонение
- .переходность
- ..маркер переходности
- ...отсутствует

2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ

- .местоимения и местоименные слова
- ..разряды местоимений
- ...взаимные
- ...возвратные
- ...возвратно-притяжательные
- ...вопросительные
- ...личные
- ...неопределенные
- ...неопределенно-личные
- ...определительные
- ...относительные
- ...отрицательные
- ...притяжательные
- ...указательные
- ..местоименные наречия
- .дейктический статус
- ..указательные местоимения
- ...ближний план/дальний план
- ...сфера говорящего/собеседника
- ..местоименные наречия
- ...сверху/снизу/внутри/снаружи/вдалеке
- .грамматическая категория лица
- ..в глаголе
- ...отсутствует
- .пространственная ориентация действия
- ..выражение
- ...наречия
- ...предлоги
- ...указательные местоимения
- .определенность/неопределенность имени
- ..артикли
- ...присоединяемые
- ...свободные
- ..синтаксическое выражение

- ...прилагательным
- ...причастием
- .формы вежливости
- ..в глагольной группе
- ...показатели вежливости
- ..в местоимениях
- ..в сентенциальном синтаксисе
- .отрицание
- ..именное >< глагольному
- ..лексическое
- ..морфологическое
- ...отрицательные аффиксы
- ..позиция отрицательного показателя
- ...постпозиция
- ...препозиция
- 2.3.7. ЧАСТИ РЕЧИ
- .знаменательные
- ..глагол
- ...финитные формы
- ...нефинитные формы
-инфинитив
-причастие
-супин
- ..местоимение
- ..наречие
- ..прилагательное
- ..существительное
- ..числительное
- .служебные
- ..артикль
- ..модальные слова
- ..предлог
- ..союз
- .междометия
- 2.4.0. ПАРАДИГМЫ
- .артикль
- ..определенность/неопределенность
- ..приименность/приадъективность
- ..род
- ..число
- .глагол
- ..финитные формы
- ...словоизменение
-время
-залог

-наклонение
- ...типы спряжения
- ..нефинитные формы
- ...инфинитив
-активный/пассивный
-неперфектный/перфектный/футуральный
- ...причастие
-настоящего времени/прошедшего времени
- .местоимение
- ..парадигма не совп. с именной по составу категорий
- .наречие
- ..степени сравнения
- .прилагательное
- ..словоизменение
- ...определенность/неопределенность
- ...род
- ...степени сравнения
- ...число
- ..типы склонения
- .существительное
- ..словоизменение
- ...определенность/неопределенность
- ...падеж
- ...число
- ..классифицирующие категории
- ...род
- .числительное
- ..парадигма не совп. с именной по составу категорий
- 2.5.1. СТРУКТУРА СЛОВОФОРМЫ
- .типы морфем, присоединяемых к корневым
- ..словообразовательные аффиксы
- ..словоизменяющие аффиксы
- .средства словоизменения
- ..внутренняя флексия
- ..суффиксы
- ...следуют за словообразовательными аффиксами
- .модели словоформы
- ..преимущественно суффиксальная

- . конкретная модель словоформы
- .. глагольная
- ... основа+афф. времени+афф. залога
- .. именная
- ... основа+афф. числа+афф. опред.+афф. падежа

2.5.2. СЛОВООБРАЗОВАНИЕ

- . деривация
- .. аффиксация
- ... префиксы
- ... суффиксы
- .. значимое чередование
- .. конверсия
- .. усечение/наращение основы
- . обратное словопроизводство
- . сложение основ
- .. количество основ может быть более двух
- .. прямое сложение
- .. соединительная морфема
- .. сращение
- . сложносокращенные слова

2.5.3. ПРОСТОЕ ПРЕДЛОЖЕНИЕ

- . строй
- .. номинативный
- . линейный порядок членов предложения
- .. жесткий
- .. основной
- ... SVO
- . линейный порядок в группах
- .. в именной группе
- ... адъективное опред. предшествует определяемому
- ... генитивное опред.

- предшествует определяемому
- ... нумеративное опред.
- предшествует определяемому
- .. в глагольной группе
- ... вспомогательный глагол предшествует главному
- ... фиксир. позиция обстоятельств и косв. дополнений
- . согласование
- .. атрибутивное согласование
- ... по определенности/неопределенности
- ... по роду
- ... по числу
- . коммуникативно значимые средства
- .. порядок слов

2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ

- . линейный порядок компонентов
- .. не фиксирован
- . особенности подчиненного компонента
- .. особый порядок слов
- .. оформление сказуемого
- ... нефинитные формы
- ... таксисные конструкции
- ... финитные формы
- . тип построения
- .. сочинение/подчинение
- . тип связи
- .. союзная/бессоюзная
- .. союзы
- ... самостоятельные служебные элементы

Результат сравнения рефератов датского и шведского языков

- 2.1.1. ФОНЕМНЫЙ СОСТАВ(1)
- .гласные(3)
 - ..подъем(5)
 - ..ряд(17)
 - ...передний/средний/задний(21)
 - ..долгие/краткие(24)
 - ..лабиализованные/нелабиализованные(28)
 - ..общая схема монофтонгов(43)
 - ...различия по ряду(44)
 -в нижнем подъеме(45)
 -в верхнем подъеме(51)
 -передний/задний(54)
 - ...различия по лабиализации(58)
 - ..число монофтонгов(85)
 - .согласные(143)
 - ..шумные(144)
 - ...ларингальные признаки(145)
 -звонкие/глухие(147)
 - ...способ образования(180)
 -фрикативные/смычные(181)
 - ..локальный ряд(184)
 -губные(185)
 -билабиальные/лабиодентальные(188)
 -переднеязычные(189)
 -смычные неаффрицированные(190)
 -фрикативные и аффрикаты(205)
 -среднеязычные(232)
 -единный ряд собственно среднеязычных(234)
 -заднеязычные(238)
 -постувулярные(242)
 -только фарингальные(246)
 - ...сонорные(330)
 -способ образования(331)
 -локальный ряд(339)
 -назальные(340)
 -губные/переднеязычные/заднеязычные(344)
 -плавные(352)
 -латеральные(354)
 -вибранты(364)
- 2.1.2. ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ(440)
- .ударение(442)
 - ..тип(444)
 - ..вид(447)
 - ...динамическое(448)
 - ..носитель ударения(453)
 - ...слог(456)
 - ..фиксированность(458)
 - ...связанное(460)
 -начальный слог(461)
 - ..главное/второстепенное ударение(482)
 - .долгота(513)
 - ..характер долготы(514)
 - ...фонологическая(515)
 - ..носитель долготы(519)
 - ...фонема(520)
- 2.1.3. ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ(570)
- .гласные(572)
 - ..структурные позиции(573)
 - ..комбинаторные позиции(605)
 - ..комбинаторные позиции(683)
 - ...ассимиляция(687)
 -регрессивная(688)
 - ..фонологическое содержание чередований(719)
 - ...аллофоническое варьирование(720)
- 2.1.4. СЛОГ(737)
- .элемент, предшествующий гласному(740)
 - ..состоит, как правило, из(745)
 - .элемент, следующий за гласным(756)
 - ..состоит, как правило, из(761)
 - ..прочие запрещения(774)

2.2.1. **ФОНОЛОГИЧЕСКАЯ СТРУКТУРА**(902)
.фонологическая структура морфемы(904)
..ограничения на фонемную структуру сегм. морфем(923)
...конечная позиция(936)
.фонологическая структура слова(964)
..ограничения на фонемную структуру слова(970)
2.2.2. **ФОНОЛОГИЧЕСКИЕ ПРОТИВОПОСТАВЛЕНИЯ МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ**(1033)
.знаменательные/служебные слова(1035)
..акцентные различия(1039)
.корневые/аффиксальные морфемы(1054)
..различия в фонетическом составе(1058)
2.2.3. **МОРФОЛОГИЧЕСКИ ОБУСЛОВЛЕННЫЕ ЧЕРЕДОВАНИЯ**(1062)
.функциональный тип(1065)
..сопутствующие(1066)
..смыслоразличительные(1067)
...словообразовательные(1068)
....в имени(1069)
....в глаголе(1070)
...словоизменительные(1071)
....в имени(1072)
....в глаголе(1073)
.вид чередований(1074)
..вокалические(1075)
...количественные(1076)
...качественные(1077)
2.3.0. **МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА**(1084)
.способ соединения морфем в слове(1086)
..флективные языки(1093)
..флективные языки с элементами агглютинации(1094)
..признаки флективного строя(1108)
...внутренняя флексия(1109)

.степень спаянности морфем(1120)
..аналитические языки с элементами синтетизма(1123)
...в имени(1124)
...в глаголе(1125)
2.3.1. **КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ**(1142)
.семантические и формальные(1145)
..морфолого-синтаксические(1148)
2.3.2. **ИМЕННЫЕ КЛАССИФИКАЦИИ**(1149)
.согласовательные классы(1152)
..род(1153)
...общий/средний(1156)
..способ выражения согласовательных классов(1157)
...морфологический(1165)
...синтаксический(1172)
....в артикле(1174)
....в местоимении(1179)
....в прилагательных(1182)
....в причастии(1183)
..мотивация согласовательных классов(1187)
...одушевленность(1200)
...пол(1203)
2.3.3. **ЧИСЛО**(1243)
.структура числа(1246)
..единственное/множественное(1248)
..способы выражения числа(1258)
..аффиксация(1260)
..значимое чередование(1262)
..просодические явления(1269)
..супплетивизм(1275)
..согласование по числу(1296)
..атрибутивное(1298)
..система счисления(1299)
..десятеричная(1300)
2.3.4. **ПАДЕЖНЫЕ ЗНАЧЕНИЯ**(1354)
.актантные падежные

значения(1356)
 .. субъектно-объектные(1357)
 ... общий/родительный(1381)
 ... субъектный/объектный
 (1392)
 .. различное падежное
 оформление(1399)
 ... объекта(1412)
 ... имен и местоимений(1413)
 .. способ выражения
 субъектно-объектных
 значений(1429)
 ... падежные аффиксы(1434)
 ... порядок слов(1435)
 ... служебные слова(1436)
 ... в препозиции(1437)
 . падежное оформление
 атрибутивного
 отношения(1495)
 .. генитив(1498)
 .. предложная
 конструкция(1510)
 .. соположение имен(1516)
 . падежное оформление
 посессивного отношения(1521)
 .. генитив(1526)
 .. предложная
 конструкция(1549)
 .. притяжательное
 местоимение(1553)
 . падежное оформление
 локативных отношений(1562)
 .. способ выражения(1563)
 ... местоименные
 наречия(1570)
 ... предлоги(1573)
 2.3.5. ГЛАГОЛЬНЫЕ
 КАТЕГОРИИ(1760)
 . залог(1762)
 .. категориальный состав
 кроме активного(1763)
 ... каузатив(1771)
 ... меди(1776)
 ... пассив(1780)
 ... статический/динамический
 (1781)
 ... рефлексив(1787)
 ... реципрок(1790)
 .. способ выражения(1795)
 ... морфологическая
 категория(1798)
 ... служебные слова(1799)
 .. совпадение форм(1806)
 ... меди/реципрок/пассив
 (1809)
 . видо-временные
 категории(1817)
 .. основные временные
 противопоставления(1818)
 ... настоящее/прошедшее/
 будущее(1825)
 .. основные видовые
 противопоставления(1833)
 ... результативный/
 нерезультативный(1844)
 .. выражение вида и
 времени(1849)
 ... раздельное(1850)
 .. виды прошедшего
 времени(1907)
 ... претерит/перфект/
 плюсквамперфект(1973)
 .. виды будущего
 времени(2039)
 ... буд.1/буд.2/буд.
 в прошедшем1/буд.
 в прошедшем2(2047)
 .. способ выражения(2084)
 ... морфологическая
 категория(2087)
 . способ действия(2089)
 .. основные значения(2090)
 ... однократность(2159)
 . наклонения кроме
 изъявительного(2220)
 .. основные значения(2221)
 ... повелительное
 наклонение(2222)
 ... императив(2223)
 ... косвенные
 наклонения(2249)
 .. способ выражения(2340)
 ... морфологическая
 категория(2341)
 ... служебные слова(2342)
 . модальность(2343)
 .. основные значения(2344)
 ... возможность(2351)

- ...долженствование(2358)
- ..способ выражения(2438)
- ...служебные слова(2441)
- .синкретичное выражение нескольких глагольных значений одной категорией(2442)
- ..время и наклонение(2446)
- .переходность(2452)
- ..маркер переходности(2453)
- ...отсутствует(2454)
- 2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ(2523)
- .местоимения и местоименные слова(2525)
- ..разряды местоимений(2526)
- ...взаимные(2531)
- ...возвратные(2535)
- ...возвратно-притяжательные(2542)
- ...вопросительные(2545)
- ...личные(2553)
- ...неопределенные(2566)
- ...неопределенно-личные(2567)
- ...относительные(2580)
- ...отрицательные(2582)
- ...притяжательные(2586)
- ...указательные(2598)
- ..местоименные наречия(2611)
- .дейктический статус(2616)
- ..указательные местоимения(2617)
- ...ближний план/дальний план(2618)
- ...сфера говорящего/собеседника(2631)
- ..местоименные наречия(2651)
- ...сверху/снизу/внутри/снаружи/вдалеке(2660)
- .грамматическая категория лица(2695)
- ..в глаголе(2702)
- ...отсутствует(2704)
- .пространственная ориентация действия(2719)
- ..выражение(2720)
- ...наречия(2733)
- ...предлоги(2738)

- ...указательные местоимения(2741)
- .определенность/неопределенность имени(2743)
- ..артикли(2745)
- ...присоединяемые(2746)
- ...свободные(2747)
- ..синтаксическое выражение(2759)
- ...прилагательным(2767)
- ...причастием(2769)
- .формы вежливости(2775)
- ..в глагольной группе(2782)
- ...показатели вежливости(2789)
- ..в местоимениях(2791)
- ..в сентенциальном синтаксисе(2792)
- .отрицание(2804)
- ..именное ><глагольному(2806)
- ..лексическое(2808)
- ..морфологическое(2810)
- ...отрицательные аффиксы(2812)
- ..позиция отрицательного показателя(2814)
- ...постпозиция(2817)
- ...препозиция(2818)
- 2.3.7. ЧАСТИ РЕЧИ(2831)
- .знаменательные(2834)
- ..глагол(2837)
- ...финитные формы(2838)
- ...нефинитные формы(2839)
- ...инфинитив(2852)
- ...причастие(2857)
- ..местоимение(2869)
- ..наречие(2870)
- ..прилагательное(2875)
- ..существительное(2878)
- ..числительное(2879)
- ..служебные(2880)
- ..артикль(2881)
- ..предлог(2890)
- ..союз(2897)
- ..междометия(2902)
- 2.4.0. ПАРАДИГМЫ(2905)
- .артикль(2910)
- ..определенность/

- неопределенность (2912)
 - ..приименность/
- приадъективность (2916)
 - ..род (2917)
 - ..число (2919)
 - .глагол (2920)
 - ..финитные формы (2921)
 - ...словоизменение (2922)
 -наклонение (2938)
 - ...типы спряжения (2956)
 - ..нефинитные формы (2992)
 - ...инфинитив (3003)
 -активный/пассивный (3004)
 -неперфектный/перфектный/футуральный (3011)
 - ...причастие (3116)
 -настоящего времени/прошедшего времени (3139)
 - .местоимение (3230)
 - ..парадигма не совп. с именной по составу категорий (3231)
 - .наречие (3236)
 - ..степени сравнения (3244)
 - .прилагательное (3269)
 - ..словоизменение (3270)
 - ...определенность/неопределенность (3281)
 - ...род (3294)
 - ...степени сравнения (3300)
 - ...число (3304)
 - ..типы склонения (3305)
 - .существительное (3317)
 - ..словоизменение (3318)
 - ...определенность/неопределенность (3336)
 - ...падеж (3337)
 - ...число (3370)
 - ..классифицирующие категории (3372)
 - ...род (3375)
 - .числительное (3401)
 - ..парадигма не совп. с именной по составу категорий (3402)
 - 2.5.1. СТРУКТУРА СЛОВОФОРМЫ (3412)

- .типы морфем, присоединяемых к корневым (3416)
 - ..словообразовательные аффиксы (3417)
 - ..словоизменительные аффиксы (3418)
 - .средства словоизменения (3419)
 - ..внутренняя флексия (3420)
 - ..суффиксы (3432)
 - ...следуют за словообразовательными аффиксами (3434)
 - .модели словоформы (3442)
 - ..преимущественно суффиксальная (3446)
 - .конкретная модель словоформы (3448)
 - ..глагольная (3452)
 - ...основа+афф. времени+афф. залога (3479)
 - ..именная (3528)
 - ...основа+афф. числа+афф. опред.+афф. падежа (3567)
 - 2.5.2. СЛОВООБРАЗОВАНИЕ (3598)
 - .деривация (3600)
 - ..аффиксация (3601)
 - ...префиксы (3607)
 - ...суффиксы (3609)
 - ..значимое чередование (3613)
 - ..конверсия (3614)
 - ..усечение/наращение основы (3621)
 - .обратное словопроизводство (3622)
 - .сложение основ (3627)
 - ..количество основ может быть более двух (3628)
 - ..прямое сложение (3629)
 - ..соединительная морфема (3634)
 - ..сращение (3635)
 - .сложносокращенные слова (3636)
 - 2.5.3. ПРОСТОЕ ПРЕДЛОЖЕНИЕ (3638)
 - .строй (3640)
 - ..номинативный (3644)

.линейный порядок членов предложения(3657)
 ..жесткий(3658)
 ..основной(3660)
 ...SVO(3661)
 .линейный порядок в группах(3667)
 ..в именной группе(3668)
 ...адъективное определение предшествует определяемому(3669)
 ...генитивное определение предшествует определяемому(3670)
 ...нумеративное определение предшествует определяемому(3673)
 ..в глагольной группе(3680)
 ...вспомогательный глагол предшествует главному(3681)
 ...фиксированная позиция обстоятельств и косвенных дополнений(3687)
 .согласование(3717)
 ..атрибутивное согласование(3748)

...по определенности/неопределенности(3752)
 ...по роду(3754)
 ...по числу(3757)
 .коммуникативно значимые средства(3758)
 ..порядок слов(3762)
 2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ(3779)
 .линейный порядок компонентов(3781)
 ..не фиксирован(3784)
 .особенности подчиненного компонента(3787)
 ..особый порядок слов(3790)
 ..оформление сказуемого(3791)
 ...таксисные конструкции(3798)
 .тип построения(3813)
 ..сочинение/подчинение(3815)
 .тип связи(3818)
 ..союзная/бессоюзная(3819)
 ..союзы(3822)
 ...самостоятельные служебные элементы(3826)

Классы:

2.1.1. ФОНЕМНЫЙ СОСТАВ(1)--->0.66
 2.1.2. ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ(440)--->0.69
 2.1.3. ФОНЕТИЧЕСКИ ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ(570)--->0.40
 2.1.4. СЛОГ(737)--->0.38
 2.2.1. ФОНОЛОГИЧЕСКАЯ СТРУКТУРА(902)--->0.36
 2.2.2. ФОНОЛ. ПРОТИВОП. МОРФОЛОГ. КАТЕГОРИЙ(1033)--->0.83
 2.2.3. МОРФОЛОГИЧЕСКИ ОБУСЛОВЛЕННЫЕ ЧЕРЕДОВАНИЯ(1062)--->1
 2.3.0. МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА(1084)--->0.80
 2.3.1. КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ(1142)--->1
 2.3.2. ИМЕННЫЕ КЛАССИФИКАЦИИ(1149)--->0.90
 2.3.3. ЧИСЛО(1243)--->0.89
 2.3.4. ПАДЕЖНЫЕ ЗНАЧЕНИЯ(1354)--->1
 2.3.5. ГЛАГОЛЬНЫЕ КАТЕГОРИИ(1760)--->0.78
 2.3.6. ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ(2523)--->0.95
 2.3.7. ЧАСТИ РЕЧИ(2831)--->0.94
 2.4.0. ПАРАДИГМЫ(2905)--->0.97
 2.5.1. СТРУКТУРА СЛОВОФОРМЫ(3412)--->1
 2.5.2. СЛОВООБРАЗОВАНИЕ(3598)--->1
 2.5.3. ПРОСТОЕ ПРЕДЛОЖЕНИЕ(3638)--->0.95
 2.5.4. СЛОЖНОЕ ПРЕДЛОЖЕНИЕ(3779)--->0.91
 S=0,820272762037506

СПЕЦКУРС

«Базы данных для типологических и компаративистских исследований»

Тема 1. БАЗЫ ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ (2 часа)

1.1. Структуры данных, слабоструктурированные данные. Цели использования БД в лингвистических исследованиях. Типы БД. Проблема совместимости и стандартизации форматов данных.

Тема 2. ЭНЦИКЛОПЕДИЧЕСКОЕ ИЗДАНИЕ СЕРИИ «ЯЗЫКИ МИРА», БД «ЯЗЫКИ МИРА» (2 часа)

2.1. История проекта. Обзор изданий серии «Языки мира». История создания БД «Языки мира». Концепция, положенная в основу БД. Персоналии издательского и электронного проектов «Языки мира». Обзор состояния дел в области БД и электронных описаний языков.

Тема 3. ХАРАКТЕРИСТИКИ БД «ЯЗЫКИ МИРА» (6 часов)

3.1. Общие характеристики БД «Языки мира». Представление данных в БД «Языки мира». Структура реферата. Иерархический принцип представления признаков. Текстовая часть реферата. Бинарная часть реферата. Обзор БД по группам признаков. Примеры описания языков.

3.2. Генетический указатель языков. Географический указатель языков. Обзор языков и групп языков, представленных в БД.

3.3. Функциональные возможности программного обеспечения. Обзор версий для DOS, WINDOWS, WEB. Демонстрация функций.

Тема 4. НАПОЛНЕНИЕ БД «ЯЗЫКИ МИРА» (4 часа)

4.1. Технологическая схема процесса подготовки и ввода данных в БД. Интерпретация текстового описания языков в терминах признаков. Ввод и корректировка данных. Проблема интерпретации спорных моментов. Проблема противоречивости данных. Проблема неполноты данных. Неполнота данных как неустранимое свойство описания языков. Проблема ошибочных данных. Ручные и автоматические методы верификации данных. Методы работы с неполными и противоречивыми данными. Проблема интерпретации и описания языково-специфических признаков. Проблема интерпретации целочисленных b и текстовых признаков в бинарных терминах. Выявление новых признаков. Пополнение шаблона описания реферата. Проблема толкования признаков при обращении к БД. Методы повышения валидности данных. Ссылки на первоисточники. Работа с рецензентами.

4.2. Пример наполнения БД для конкретного языка.

Тема 5. БД КАК ИНСТРУМЕНТ ТИПОЛОГИЧЕСКИХ И КОМПАРАТИВНЫХ ИССЛЕДОВАНИЙ (14 часов)

5.1. Типовые запросы к БД. Запрос: «Найти описание языка X». Запрос: «В какие языки входит признак Y». И/ИЛИ/НЕ – запросы по комбинации признаков. Запросы по подстроке. Запросы по генетическому указателю. Запросы по географическому указателю. Запросы с фильтрами (по признаку или комбинации признаков, генетическому указателю, географическому указателю). Количественные запросы. Комбинированные запросы. Выполнение простейших вычислительных процедур с применением электронных таблиц Excel. Примеры запросов.

5.2. Расчет мер близости языков. Виды мер близости. Аддитивные меры близости. Весовые меры близости по иерархии признаков. Весовые меры близости по экспертной оценке (рангу) признаков. Нормированные меры близости. Логарифмические меры близости. Меры близости на основе свертки иерархий. Нечеткие оценки мер близости. Меры близости на основе присутствующих/отсутствующих признаков. Формальная постановка задачи расчета меры близости. Примеры расчетов. Нерешенные проблемы.

5.3. Группирование и кластеризация языков. Простейшие способы группирования языков (по наличию/отсутствию/комбинации признаков, генетическому указателю, географическому указателю). Примеры запросов на группирование языков.

Кластеризация языков на основе мер близости: общие подходы. Формальная постановка задачи кластеризации. Алгоритмы кластеризации. Нечеткая кластеризация. Примеры расчетов по кластеризации языков.

Методы кластеризации Сильницкого Г. Анализ полученных результатов. Нерешенные проблемы.

5.4. Корреляционный анализ языков и визуализация.

Выявление признаков-универсалий и классифицирующих признаков. Примеры.

Выявление наследуемых признаков. Примеры.

Выявление приобретенных признаков. Примеры.

Выявление отклонений языков от группы. Примеры.

Выявление признаков, не входящих в группы универсалий и классифицирующих. Примеры. Нерешенные проблемы.

5.5. Возможности визуализации языков в пространстве признаков и признаков в пространстве языков. Примеры.

5.6. Постановка задачи для глоттохронологических исследований на основе БД «Языки мира». Постановка задачи для ностратических исследований на основе БД «Языки мира».

5.7. Перспективы развития БД. Наполнение БД новыми описаниями языков. Новая интерфейсная концепция. Расширенный функционал. Программы-мастера выполнения исследовательских расчетов. Визуализация результатов расчетов.

6. ЗНАКОМСТВО С ОСНОВНЫМИ ЗАПАДНЫМИ БД (4 часа)

6.1. Всемирный атлас лингвистических структур (Лейпциг, Германия). Сопоставление базовых архитектурных решений: графический интерфейс, язык запросов. Принципы выбора языков и параметров.

6.2. Базы данных по согласованию (Суррей, Великобритания), редупликаций (Грац, Австрия).

ОСНОВНАЯ ЛИТЕРАТУРА

- [1] Поляков В.Н., Соловьев В.Д. Компьютерные модели и методы в типологии и компаративистике (в печати).
- [2] Принципы описания языков мира. – М.: Наука, 1985.
- [3] Теоретические основы классификации языков. – М.: Наука. – Т. 1. – 1980, Т. 2. – 1982.
- [4] Языки и диалекты мира. – М.: Наука, 1982.
- [5] Журина М.А., Новиков А.И., Ярославцева Е.И. Энциклопедическое описание языков. – М.: Наука, 1986.
- [6] Ярославцева Е.И. Грамматикон и база данных «Языки мира»// Проблемы прикладной лингвистики. – 2001. – М., 2002.
- [7] Сильницкий Г.Г. Квантитативная грамматико-фонетическая типология языков и языковых признаков. – Смоленск, 2004.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

- [1] A.Novikov, E.Yaroslavtseva. Linguotypological Data Bank // Social Sciences. – USSR Academy of Sciences. – Vol. XVII. – No. 3. – 1986.
- [2] Бурлак С.А., Старостин С.А. Введение в лингвистическую компаративистику. – М.: УРСС, 2001. – 272 с.
- [3] В.А.Виноградов, А.И.Новиков, Е.И.Ярославцева, В.Н. Поляков, В.В. Логунов База данных «Языки мира» и лингвистические исследования. II Международные Бодуэновские чтения (Казань, 11-13 декабря 2003 г.). Труды и материалы. – Т.1. – С. 39-41.
- [4] Языки мира: Уральские языки. – М., 1993.
- [5] Языки мира: Тюркские языки. – М.: Издательство «Индрик», 1997. – 544 с.
- [6] Языки мира: Монгольские языки. Тунгусо-маньчжурские языки. Японский язык. Корейский язык. – М.: Издательство «Индрик», 1997. – 408 с.
- [7] Языки мира. Палеоазиатские языки. – М.: Издательство «Индрик», 1996. – 231 с.
- [8] Языки мира: Иранские языки. I. Юго-западные иранские языки. – М.: Индрик, 1997. – 207 с.
- [9] Языки мира: Иранские языки. II. Северо-западные иранские языки. – М.: Индрик, 1999.
- [10] Языки мира: Иранские языки. III. Восточноиранские языки. – М.: Индрик, 1999. – 343 с.
- [11] Языки мира: Дардские и нуристанские языки. – М.: Индрик, 1998. – 143 с.
- [12] Языки мира: Кавказские языки. – РАН. Ин-т языкознания. – М.: Academia, 2001. – 480 с.
- [13] Языки мира: Германские языки. Кельтские языки. – М.: Academia, 1999. – 472 с.
- [14] Языки мира: Романские языки. – М.: Academia, 2001. – 720 с.
- [15] Языки мира: Индоарийские языки древнего и среднего периода. – М.: Academia, 2004. – 160 с.
- [16] Языки мира: Славянские языки / РАН. Институт языкознания; Ред. колл.: А. М. Молдован, С. С. Скорвид, А. А. Кибрик и др. – М.: Academia, 2005. – 656 с.

Квантитативная карта признаков раздела «2.5.4. Сложное предложение» для индоевропейских языков

| Номер признака в БД | № бд | Наименование | Группа | индоевропейские (133) | анатолийские (1) | армянские (1) | германские (14) | дардские (13) | индоарийские (9) | иранские (38) | италийские (4) | кельтские (9) | нуристанские (5) | романские (25) | славянские (14) | дамели | АВЕСТЫ |
|---------------------|----------|--|----------|-----------------------|------------------|---------------|-----------------|---------------|------------------|---------------|----------------|---------------|------------------|----------------|-----------------|-------------|----------|
| | | | | F c | F B | | | | | | | | | | | | |
| 3773 | 312 | <i>2.5.4.СЛОЖНОЕ ПРЕДЛОЖЕНИЕ</i> | 1 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,96 | 1,00 | 1,00 | 1 |
| 3775 | 40 | ..линейный порядок компонентов | 1 | 0,14 | 0,00 | 0,00 | 0,50 | 0,00 | 0,22 | 0,08 | 0,00 | 0,00 | 0,00 | 0,20 | 0,14 | 0,00 | 0 |
| 3776 | 7 | ..главное предшествует придаточному | 4 | 0,04 | 0,00 | 0,00 | 0,07 | 0,00 | 0,00 | 0,05 | 0,00 | 0,00 | 0,00 | 0,08 | 0,00 | 0,00 | 0 |
| 3778 | 23 | ..не фиксирован | 1 | 0,10 | 0,00 | 0,00 | 0,43 | 0,00 | 0,11 | 0,03 | 0,00 | 0,00 | 0,00 | 0,12 | 0,14 | 0,00 | 0 |
| 3779 | 11 | ..придаточное предшествует главному | 1 | 0,01 | 0,00 | 0,00 | 0,00 | 0,00 | 0,11 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3780 | 1 | ..целевое предшествует главному | 6 | 0,01 | 0,00 | 0,00 | 0,00 | 0,00 | 0,11 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3781 | 144 | <i>..особенности подчиненного компонента</i> | 1 | 0,35 | 1,00 | 0,00 | 0,71 | 0,15 | 0,89 | 0,13 | 0,75 | 0,11 | 0,40 | 0,44 | 0,21 | 0,00 | 1 |
| 3782 | 5 | ..особое оформление именных групп | 5 | 0,02 | 0,00 | 0,00 | 0,00 | 0,08 | 0,00 | 0,00 | 0,00 | 0,00 | 0,20 | 0,00 | 0,00 | 0,00 | 0 |
| 3783 | 7 | ..особое оформление подлежащего | 5 | 0,02 | 0,00 | 0,00 | 0,07 | 0,00 | 0,00 | 0,00 | 0,25 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3784 | 7 | ..особый порядок слов | 4 | 0,05 | 0,00 | 0,00 | 0,43 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |

| | | | | | | | | | | | | | | | | | |
|-------------|----------|---|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|----------|
| 3785 | 137 | <i>...оформление сказуемого</i> | 1 | 0,34 | 1,00 | 0,00 | 0,57 | 0,15 | 1,00 | 0,13 | 0,75 | 0,11 | 0,40 | 0,44 | 0,21 | 0,00 | 1 |
| 3786 | 22 | <i>...зависимые личные формы</i> | 1 | 0,04 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,25 | 0,00 | 0,20 | 0,12 | 0,00 | 0,00 | 0 |
| 3787 | 6 | <i>...квотатив</i> | 5 | 0,01 | 0,00 | 0,00 | 0,00 | 0,08 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3788 | 104 | <i>...нефинитные формы</i> | 1 | 0,23 | <u>0,00</u> | 0,00 | 0,29 | 0,08 | 1,00 | 0,05 | 0,50 | 0,11 | 0,00 | 0,36 | 0,21 | 0,00 | 1 |
| 3789 | 1 | <i>...оборот с релятивной формой глагола</i> | 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3790 | 1 | <i>...особые глагольные показатели</i> | 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3791 | 1 | <i>...особый порядок слов</i> | 6 | 0,01 | 0,00 | 0,00 | 0,07 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3792 | 18 | <i>...таксисные конструкции</i> | 5 | 0,09 | 0,00 | 0,00 | 0,36 | 0,00 | 0,00 | 0,03 | 0,25 | 0,00 | 0,00 | 0,20 | 0,00 | 0,00 | 0 |
| 3793 | 76 | <i>...финитные формы</i> | 1 | 0,21 | <u>0,00</u> | 0,00 | 0,36 | 0,08 | 0,89 | 0,08 | 0,75 | 0,00 | 0,20 | 0,24 | 0,07 | 0,00 | 1 |
| 3794 | 2 | ...абсолютные обороты | 4 | 0,02 | 0,00 | 0,00 | 0,07 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,04 | 0,00 | 0,00 | 0 |
| 3795 | 1 | <i>..'тематическое' придаточное</i> | 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3799 | 14 | <i>.структура относительного предложения</i> | 4 | 0,07 | 1,00 | 0,00 | 0,07 | 0,00 | 0,33 | 0,03 | 0,25 | 0,00 | 0,00 | 0,04 | 0,07 | 0,00 | 1 |
| 3800 | 5 | <i>...главное предшествует придаточному</i> | 4 | 0,03 | <u>0,00</u> | 0,00 | 0,07 | 0,00 | 0,00 | 0,03 | 0,00 | 0,00 | 0,00 | 0,04 | 0,07 | 0,00 | 1 |
| 3801 | 1 | <i>...относительно-местоименные аффиксы в глаголе</i> | 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3802 | 1 | <i>...отсутствие союзного слова или его аналога</i> | 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3803 | 2 | ...придаточное предшествует главному | 4 | 0,02 | 0,00 | 0,00 | 0,00 | 0,00 | 0,22 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3804 | 1 | <i>...союзное слово в главном</i> | 6 | 0,01 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3805 | 10 | <i>...союзное слово в придаточном</i> | 5 | 0,05 | 0,00 | 0,00 | 0,07 | 0,00 | 0,33 | 0,00 | 0,25 | 0,00 | 0,00 | 0,00 | 0,07 | 0,00 | 0 |

| | | | | | | | | | | | | | | | | | |
|-------------|-----------|--|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|----------|
| 3806 | 1 | ..сказуемое придаточного следует за союзом | 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3807 | 259 | <i>.тип построения</i> | 1 | 0,83 | 1,00 | 0,00 | 1,00 | 0,46 | 1,00 | 0,76 | 1,00 | 0,89 | 0,40 | 0,92 | 1,00 | 1,00 | 1 |
| 3808 | 1 | ..сериализация | 6 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3809 | 248 | <i>..сочинение/подчинение</i> | 1 | 0,77 | 1,00 | 0,00 | 1,00 | 0,38 | 0,89 | 0,76 | 0,50 | 0,67 | 0,20 | 0,92 | 1,00 | 1,00 | 1 |
| 3810 | 7 | ..только подчинение | 5 | 0,03 | 0,00 | 0,00 | 0,00 | 0,08 | 0,11 | 0,00 | 0,25 | 0,00 | 0,20 | 0,00 | 0,00 | 0,00 | 0 |
| 3811 | 3 | ..только сочинение | 5 | 0,02 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,25 | 0,11 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3812 | 264 | <i>.тип связи</i> | 1 | 0,83 | 1,00 | 1,00 | 1,00 | 0,46 | 1,00 | 0,79 | 1,00 | 0,78 | 0,40 | 0,92 | 0,93 | 1,00 | 1 |
| 3813 | 208 | <i>..союзная/бессоюзная</i> | 1 | 0,65 | 1,00 | 1,00 | 1,00 | 0,31 | 1,00 | 0,53 | 0,50 | 0,11 | 0,20 | 0,84 | 0,86 | 1,00 | 1 |
| 3814 | 39 | ..преобладает бессоюзная | 1 | 0,02 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,03 | 0,25 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |
| 3815 | 40 | ..преобладает союзная | 3 | 0,20 | 0,00 | 0,00 | 0,14 | 0,15 | 0,22 | 0,26 | 0,25 | 0,67 | 0,20 | 0,08 | 0,07 | 0,00 | 0 |
| 3816 | 223 | <i>..союзы</i> | 1 | 0,80 | 1,00 | 1,00 | 1,00 | 0,38 | 1,00 | 0,79 | 0,75 | 0,56 | 0,40 | 0,92 | 0,93 | 1,00 | 1 |
| 3817 | 37 | ...знаменательные слова | 3 | 0,11 | 1,00 | 0,00 | 0,14 | 0,00 | 0,56 | 0,05 | 0,00 | 0,00 | 0,20 | 0,08 | 0,07 | 0,00 | 0 |
| 3819 | 14 | <i>...присоединяемые служебные элементы</i> | 5 | 0,03 | <u>0,00</u> | 0,00 | 0,00 | 0,08 | 0,00 | 0,05 | 0,00 | 0,00 | 0,20 | 0,00 | 0,00 | 0,00 | 1 |
| 3820 | 216 | <i>...самостоятельные служебные элементы</i> | 1 | 0,79 | 1,00 | 1,00 | 1,00 | 0,31 | 1,00 | 0,76 | 0,75 | 0,56 | 0,40 | 0,92 | 1,00 | 1,00 | 1 |
| 3821 | 2 | ...союзные формы глагола | 4 | 0,01 | 0,00 | 0,00 | 0,00 | 0,08 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0 |

Примечания.

1. В скобках после названия семьи (ветви) указано количество языков, имеющих валидные данные и принятых к расчету.
2. Жирным выделены раритарные признаки (группа IV).
3. Курсивом выделены универсалии и псевдо-раритарии, имеющие совпадение с языком Авесты.
4. Признаки с количеством языков равным нулю из таблицы исключены.
5. Порядковые номера признаков указаны по версии БД от 05.02.06.

Динамика статистических универсалий

Статистические универсалии в БД "Языки мира"

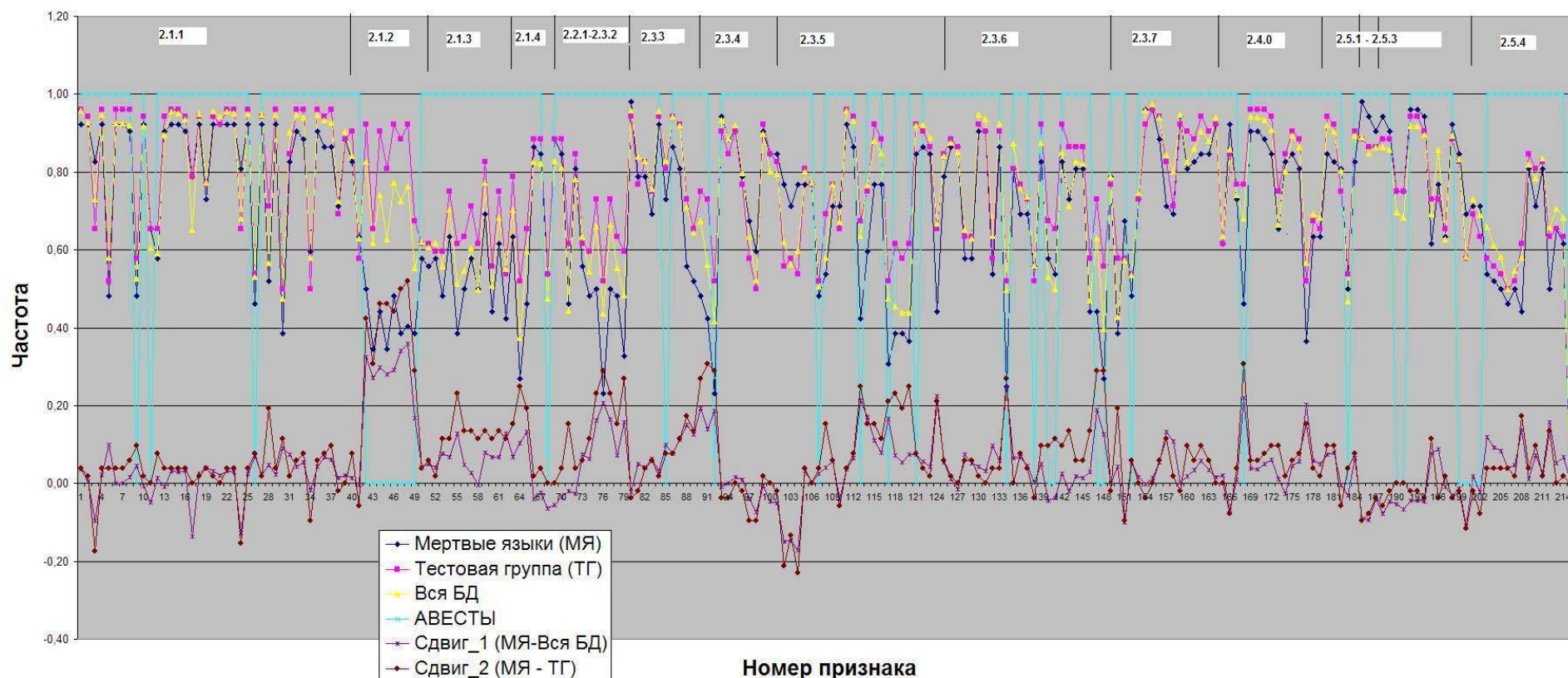


Рис. Ж.1. Графики частоты распространения универсальных признаков. На диаграмме показаны только те признаки, у которых частота в тестовой группе из 52 языков превышает или равна 0,5. Кроме того, показан сдвиг в частоте тестовой группы относительно частоты группы из мертвых языков ($F_{ТГ} - F_{МЯ}$) и сдвиг в частоте по всей БД относительно группы мертвых языков ($F_{БД} - F_{МЯ}$). Положительное значение сдвига означает, что распространенность признака со временем увеличилась. Также показана диаграмма универсальных признаков для языка Авесты. Номер по оси x соответствует номерам в таблицах Ж.1-Ж.5.

Статистические универсалии. Раздел 2.1.1. ФОНЕМНЫЙ СОСТАВ

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Клас./Факт. | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРОПИСЬМЕННЫЙ МОНГОЛЬСКИЙ | АРАБСКИЙ КЛАССИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ |
|----------------------|------------|-----------------|---------|-----------------------------------|-------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|------------------|-----------------------------|-----------------------|-------|----------|------------|----------|----------------|
| | 1 | 2.1.1. | 316 | 2.1.1.ФОНЕМНЫЙ СОСТАВ | | | | | | | | | | | | | | | | | |
| 1 | 3 | 2.1.1. | 303 | .гласные | К | 0,92 | 0,96 | 0,96 | 1 | 0,04 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 2 | 5 | 2.1.1. | 293 | ..подъем | К | 0,92 | 0,94 | 0,93 | 1 | 0,00 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 3 | 8 | 2.1.1. | 231 | ...верхний/средний/нижний | Ф | 0,83 | 0,65 | 0,73 | 1 | -0,10 | -0,17 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4 | 15 | 2.1.1. | 299 | ..ряд | К | 0,92 | 0,96 | 0,95 | 1 | 0,02 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 5 | 19 | 2.1.1. | 183 | ...передний/средний/задний | Ф | 0,48 | 0,52 | 0,58 | 1 | 0,10 | 0,04 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 6 | 41 | 2.1.1. | 292 | ..общая схема монофтонгов | К | 0,92 | 0,96 | 0,92 | 1 | 0,00 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 7 | 42 | 2.1.1. | 292 | ...различия по ряду | К | 0,92 | 0,96 | 0,92 | 1 | 0,00 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 8 | 43 | 2.1.1. | 291 |в нижнем подъеме | К | 0,90 | 0,96 | 0,92 | 1 | 0,02 | 0,06 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 9 | 44 | 2.1.1. | 166 |отсутствуют | Ф | 0,48 | 0,58 | 0,53 | 0 | 0,04 | 0,10 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 10 | 49 | 2.1.1. | 290 |в верхнем подъеме | К | 0,92 | 0,94 | 0,92 | 1 | -0,01 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 11 | 52 | 2.1.1. | 191 |передний/задний | Ф | 0,65 | 0,65 | 0,60 | 0 | -0,05 | 0,00 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 12 | 56 | 2.1.1. | 187 | ...различия по лабиализации | К | 0,58 | 0,65 | 0,59 | 1 | 0,01 | 0,08 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 13 | 83 | 2.1.1. | 283 | ..число монофтонгов | К | 0,90 | 0,94 | 0,90 | 1 | -0,01 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 14 | 141 | 2.1.1. | 302 | .согласные | К | 0,92 | 0,96 | 0,96 | 1 | 0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 15 | 142 | 2.1.1. | 301 | ..шумные | К | 0,92 | 0,96 | 0,95 | 1 | 0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 16 | 143 | 2.1.1. | 296 | ...ларингальные признаки | К | 0,90 | 0,94 | 0,94 | 1 | 0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 17 | 145 | 2.1.1. | 206 |звонкие/глухие | Ф | 0,79 | 0,79 | 0,65 | 1 | -0,14 | 0,00 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 18 | 178 | 2.1.1. | 300 | ...способ образования | К | 0,92 | 0,94 | 0,95 | 1 | 0,03 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 19 | 181 | 2.1.1. | 244 |фрикативные/смычные/аффрикаты | Ф | 0,73 | 0,77 | 0,77 | 1 | 0,04 | 0,04 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 20 | 182 | 2.1.1. | 302 | ...локальный ряд | К | 0,92 | 0,94 | 0,96 | 1 | 0,03 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Клас./Факт. | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРОПИСЬМЕННЫЙ МОНГОЛЬСКИЙ | АРАБСКИЙ КЛАССИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ |
|----------------------|------------|-----------------|---------|--|-------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|------------------|-----------------------------|-----------------------|-------|----------|------------|----------|----------------|
| 21 | 183 | 2.1.1. | 298 |губные | К | 0,92 | 0,92 | 0,94 | 1 | 0,02 | 0,00 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 22 | 187 | 2.1.1. | 302 |переднеязычные | К | 0,92 | 0,96 | 0,96 | 1 | 0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 23 | 188 | 2.1.1. | 301 |смычные неаффрицированные | К | 0,92 | 0,96 | 0,95 | 1 | 0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 24 | 189 | 2.1.1. | 215 |единый ряд переднеязычных | К | 0,81 | 0,65 | 0,68 | 1 | -0,13 | -0,15 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 25 | 203 | 2.1.1. | 300 |фрикативные и аффрикаты | К | 0,92 | 0,96 | 0,95 | 1 | 0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 26 | 216 | 2.1.1. | 168 |свистящие/шипящие | Ф | 0,46 | 0,54 | 0,53 | 0 | 0,07 | 0,08 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| 27 | 230 | 2.1.1. | 299 |среднеязычные | К | 0,92 | 0,94 | 0,95 | 1 | 0,02 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 28 | 232 | 2.1.1. | 179 |единый ряд собственно среднеязычных | Ф | 0,52 | 0,71 | 0,57 | 1 | 0,05 | 0,19 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 29 | 236 | 2.1.1. | 299 |заднеязычные | К | 0,92 | 0,96 | 0,95 | 1 | 0,02 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 30 | 238 | 2.1.1. | 150 |велярные/увулярные | Ф | 0,38 | 0,50 | 0,47 | 1 | 0,09 | 0,12 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 31 | 240 | 2.1.1. | 285 |постувулярные | К | 0,83 | 0,85 | 0,90 | 1 | 0,07 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 32 | 328 | 2.1.1. | 299 |сонорные | К | 0,90 | 0,96 | 0,95 | 1 | 0,04 | 0,06 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 33 | 329 | 2.1.1. | 297 |способ образования | К | 0,88 | 0,96 | 0,94 | 1 | 0,06 | 0,08 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 34 | 335 | 2.1.1. | 183 |назальные/плавные/вibrанты/глайды | Ф | 0,60 | 0,50 | 0,58 | 1 | -0,02 | -0,10 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 35 | 337 | 2.1.1. | 299 |локальный ряд | К | 0,90 | 0,96 | 0,95 | 1 | 0,04 | 0,06 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 36 | 338 | 2.1.1. | 295 |назальные | К | 0,87 | 0,94 | 0,93 | 1 | 0,07 | 0,08 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 37 | 350 | 2.1.1. | 293 |плавные | К | 0,87 | 0,96 | 0,93 | 1 | 0,06 | 0,10 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 38 | 352 | 2.1.1. | 229 |латеральные | К | 0,71 | 0,69 | 0,72 | 1 | 0,01 | -0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 39 | 362 | 2.1.1. | 286 |vibrанты | К | 0,88 | 0,88 | 0,91 | 1 | 0,02 | 0,00 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |

Статистические универсалии. Разделы 2.1.2-2.2.3

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Классе/Факт | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРОПИСЬМЕН-
НЫЙ МОНГОЛЬ-
СКИЙ | АРАБСКИЙ КЛАС-
СИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ |
|----------------------|------------|-----------------|---------|---|-------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|----------------|------------------|---------------------------------------|----------------------------|-------|----------|------------|----------|----------------|
| | 438 | 2.1.2. | 316 | 2.1.2.ПРОСОДИЧЕСКИЕ ЯВЛЕНИЯ | | | | | | | | | | | | | | | | | |
| 40 | 440 | 2.1.2. | 261 | ..ударение | К | 0,50 | 0,92 | 0,83 | 0 | 0,33 | 0,42 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 41 | 442 | 2.1.2. | 195 | ..тип | К | 0,35 | 0,65 | 0,62 | 0 | 0,27 | 0,31 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 42 | 445 | 2.1.2. | 234 | ..вид | К | 0,44 | 0,90 | 0,74 | 0 | 0,30 | 0,46 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 43 | 446 | 2.1.2. | 198 | ...динамическое | Ф | 0,35 | 0,81 | 0,63 | 0 | 0,28 | 0,46 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 44 | 451 | 2.1.2. | 244 | ..носитель ударения | К | 0,48 | 0,92 | 0,77 | 0 | 0,29 | 0,44 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 45 | 454 | 2.1.2. | 229 | ...слог | Ф | 0,38 | 0,88 | 0,72 | 0 | 0,34 | 0,50 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 46 | 456 | 2.1.2. | 241 | ..фиксированность | К | 0,40 | 0,92 | 0,76 | 0 | 0,36 | 0,52 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 47 | 458 | 2.1.2. | 175 | ...связанное | К | 0,38 | 0,67 | 0,55 | 0 | 0,17 | 0,29 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 48 | 511 | 2.1.2. | 196 | ..долгота | К | 0,58 | 0,62 | 0,62 | 1 | 0,04 | 0,04 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 49 | 512 | 2.1.2. | 192 | ..характер долготы | К | 0,56 | 0,62 | 0,61 | 1 | 0,05 | 0,06 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 50 | 517 | 2.1.2. | 195 | ..носитель долготы | К | 0,58 | 0,60 | 0,62 | 1 | 0,04 | 0,02 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 51 | 518 | 2.1.2. | 176 | ...фонема | Ф | 0,48 | 0,60 | 0,56 | 1 | 0,08 | 0,12 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 568 | 2.1.3. | 315 | 2.1.3.ФОНЕТИЧЕСКИ
ОБУСЛОВЛЕННЫЕ ПРОЦЕССЫ | | | | | | | | | | | | | | | | | |
| 52 | 570 | 2.1.3. | 222 | ..гласные | К | 0,63 | 0,75 | 0,70 | 1 | 0,07 | 0,12 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 53 | 571 | 2.1.3. | 162 | ..структурные позиции | К | 0,38 | 0,62 | 0,51 | 1 | 0,13 | 0,23 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 54 | 603 | 2.1.3. | 173 | ..комбинаторные позиции | К | 0,50 | 0,63 | 0,55 | 1 | 0,05 | 0,13 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 55 | 637 | 2.1.3. | 191 | ..фонологическое содержание
чередований | К | 0,58 | 0,71 | 0,60 | 1 | 0,03 | 0,13 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 56 | 639 | 2.1.3. | 157 | ...нейтрализация | Ф | 0,50 | 0,62 | 0,50 | 1 | 0,00 | 0,12 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 57 | 640 | 2.1.3. | 244 | ..согласные | К | 0,69 | 0,83 | 0,77 | 1 | 0,08 | 0,13 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 58 | 642 | 2.1.3. | 161 | ..структурные позиции | К | 0,44 | 0,56 | 0,51 | 1 | 0,07 | 0,12 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 59 | 681 | 2.1.3. | 216 | ..комбинаторные позиции | К | 0,62 | 0,75 | 0,68 | 1 | 0,07 | 0,13 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Класс/Факт | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ
СТАГОЛИСЬМЕН-
НЫЙ МОНГОЛЬ-
СКИЙ | АРАБСКИЙ КЛАС-
СИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ | |
|----------------------|------------|-----------------|---------|--|------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|---|----------------------------|-------|----------|------------|----------|----------------|---|
| 60 | 685 | 2.1.3. | 174 | ...ассимиляция | К | 0,42 | 0,54 | 0,55 | 1 | 0,13 | 0,12 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 61 | 717 | 2.1.3. | 222 | ..фонологическое содержание чередований | К | 0,63 | 0,79 | 0,70 | 1 | 0,07 | 0,15 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 62 | 718 | 2.1.3. | 118 | ...аллофоническое варьирование | Ф | 0,27 | 0,52 | 0,37 | 1 | 0,10 | 0,25 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 63 | 719 | 2.1.3. | 188 | ...нейтрализация | Ф | 0,46 | 0,65 | 0,59 | 1 | 0,13 | 0,19 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 735 | 2.1.4. | 315 | 2.1.4.СЛОГ | | | | | | | | | | | | | | | | | |
| 64 | 738 | 2.1.4. | 261 | ..элемент, предшествующий гласному | К | 0,87 | 0,88 | 0,83 | 1 | -0,04 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 65 | 743 | 2.1.4. | 260 | ..состоит, как правило, из | К | 0,85 | 0,88 | 0,82 | 1 | -0,02 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 66 | 744 | 2.1.4. | 150 | ...не более чем одного согласного | Ф | 0,54 | 0,54 | 0,47 | 0 | -0,06 | 0,00 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 67 | 754 | 2.1.4. | 262 | ..элемент, следующий за гласным | К | 0,88 | 0,88 | 0,83 | 1 | -0,06 | 0,00 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 68 | 759 | 2.1.4. | 256 | ..состоит, как правило, из | К | 0,85 | 0,88 | 0,81 | 1 | -0,04 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 69 | 761 | 2.1.4. | 140 | ...одного или двух согласных | К | 0,46 | 0,62 | 0,44 | 1 | -0,02 | 0,15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 70 | 802 | 2.1.4. | 247 | ..конкретная (или преобладающая) структура слога | К | 0,81 | 0,85 | 0,78 | 1 | -0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| | 899 | 2.2.1. | 315 | 2.2.1.ФОНОЛОГИЧЕСКАЯ СТРУКТУРА | | | | | | | | | | | | | | | | | |
| 71 | 961 | 2.2.1. | 200 | ..фонологическая структура слова | К | 0,56 | 0,62 | 0,63 | 1 | 0,08 | 0,06 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 72 | 967 | 2.2.1. | 172 | ..ограничения на фонемную структуру слова | К | 0,48 | 0,60 | 0,54 | 1 | 0,06 | 0,12 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| | 1030 | 2.2.2. | 315 | 2.2.2.ФОНОЛОГИЧЕСКИЕ ПРОТИВОП. МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ | | | | | | | | | | | | | | | | | |
| | 1059 | 2.2.3. | 315 | 2.2.3.МОРФОЛОГИЧЕСКИ ОБУСЛОВЛЕННЫЕ ЧЕРЕДОВАНИЯ | | | | | | | | | | | | | | | | | |
| 73 | 1062 | 2.2.3. | 209 | ..функциональный тип | К | 0,50 | 0,73 | 0,66 | 1 | 0,16 | 0,23 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 74 | 1064 | 2.2.3. | 138 | ..смыслоразличительные | К | 0,23 | 0,52 | 0,44 | 1 | 0,21 | 0,29 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 75 | 1071 | 2.2.3. | 210 | ..вид чередований | К | 0,50 | 0,73 | 0,66 | 1 | 0,16 | 0,23 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 76 | 1072 | 2.2.3. | 175 | ..вокалические | К | 0,48 | 0,63 | 0,55 | 1 | 0,07 | 0,15 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 77 | 1075 | 2.2.3. | 153 | ..консонантные | Ф | 0,33 | 0,60 | 0,48 | 1 | 0,16 | 0,27 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

Статистические универсалии. Разделы 2.3.0-2.3.5

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Класс/Факт | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРОПИСЬМЕННЫЙ МОНГОЛЬСКИЙ | АРАБСКИЙ КЛАССИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ | |
|----------------------|------------|-----------------|---------|---|------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|------------------|-----------------------------|-----------------------|-------|----------|------------|----------|----------------|---|
| | 1081 | 2.3.0. | 315 | 2.3.0.МОРФОЛОГИЧЕСКИЙ ТИП ЯЗЫКА | | | | | | | | | | | | | | | | | | |
| 77 | 1083 | 2.3.0. | 303 | .способ соединения морфем в слове | К | 0,98 | 0,94 | 0,96 | 1 | -0,02 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 78 | 1117 | 2.3.0. | 265 | .степень спаянности морфем | К | 0,79 | 0,77 | 0,84 | 1 | 0,05 | -0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | |
| | 1139 | 2.3.1. | 315 | 2.3.1.КРИТЕРИИ ВЫДЕЛЕНИЯ ЧАСТЕЙ РЕЧИ | | | | | | | | | | | | | | | | | | |
| 79 | 1142 | 2.3.1. | 262 | .семантические и формальные | К | 0,79 | 0,83 | 0,83 | 1 | 0,04 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |
| 80 | 1145 | 2.3.1. | 239 | ..морфолого-синтаксические | Ф | 0,69 | 0,75 | 0,76 | 1 | 0,06 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |
| | 1146 | 2.3.2. | 315 | 2.3.2.ИМЕННЫЕ КЛАССИФИКАЦИИ | | | | | | | | | | | | | | | | | | |
| | 1240 | 2.3.3. | 315 | 2.3.3.ЧИСЛО | | | | | | | | | | | | | | | | | | |
| 81 | 1243 | 2.3.3. | 302 | .структура числа | К | 0,92 | 0,94 | 0,96 | 1 | 0,03 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 82 | 1245 | 2.3.3. | 262 | ..единственное/множественное | Ф | 0,73 | 0,81 | 0,83 | 0 | 0,10 | 0,08 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | |
| 83 | 1255 | 2.3.3. | 298 | .способы выражения числа | К | 0,87 | 0,94 | 0,94 | 1 | 0,08 | 0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 84 | 1257 | 2.3.3. | 291 | ..аффиксация | Ф | 0,81 | 0,92 | 0,92 | 1 | 0,11 | 0,12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 85 | 1293 | 2.3.3. | 224 | .согласование по числу | К | 0,56 | 0,73 | 0,71 | 1 | 0,15 | 0,17 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | |
| 86 | 1294 | 2.3.3. | 204 | ..актантно-предикатное | Ф | 0,52 | 0,65 | 0,65 | 1 | 0,13 | 0,13 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | |
| 87 | 1296 | 2.3.3. | 213 | .система счисления | К | 0,48 | 0,75 | 0,67 | 1 | 0,19 | 0,27 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | |
| 88 | 1297 | 2.3.3. | 178 | ..десятеричная | К | 0,42 | 0,73 | 0,56 | 1 | 0,14 | 0,31 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | |
| 89 | 1308 | 2.3.3. | 132 | .классы числительных, отличные от порядковых и количественных | К | 0,23 | 0,52 | 0,42 | 0 | 0,19 | 0,29 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 1351 | 2.3.4. | 314 | 2.3.4.ПАДЕЖНЫЕ ЗНАЧЕНИЯ | | | | | | | | | | | | | | | | | | |
| 90 | 1353 | 2.3.4. | 295 | .актантные падежные значения | К | 0,94 | 0,90 | 0,93 | 1 | -0,01 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 91 | 1354 | 2.3.4. | 280 | ..субъектно-объектные | К | 0,88 | 0,85 | 0,89 | 1 | 0,00 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 92 | 1426 | 2.3.4. | 291 | ..способ выражения субъектно-объектных значений | К | 0,90 | 0,90 | 0,92 | 1 | 0,02 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| Номер на графике Ж.1 | Номер в БД | Раздел реферага | Частота | Наименование признака | Класс/Факт | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРОПИСЬМЕННЫЙ МОНГОЛЬСКИЙ | АРАБСКИЙ КЛАССИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ |
|----------------------|------------|-----------------|---------|--|------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|------------------|-----------------------------|-----------------------|-------|----------|------------|----------|----------------|
| 93 | 1431 | 2.3.4. | 252 | ...падежные аффиксы | Ф | 0,79 | 0,77 | 0,80 | 1 | 0,01 | -0,02 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 94 | 1437 | 2.3.4. | 200 | ..прочие актантные падежи | К | 0,67 | 0,58 | 0,63 | 1 | -0,04 | -0,10 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| 95 | 1492 | 2.3.4. | 165 | .падежное оформление атрибутивного отношения | К | 0,60 | 0,50 | 0,52 | 1 | -0,07 | -0,10 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 96 | 1518 | 2.3.4. | 284 | .падежное оформление посессивного отношения | К | 0,90 | 0,92 | 0,90 | 1 | -0,01 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 97 | 1559 | 2.3.4. | 253 | .падежное оформление локативных отношений | К | 0,85 | 0,85 | 0,80 | 1 | -0,05 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 98 | 1560 | 2.3.4. | 251 | ..способ выражения | К | 0,85 | 0,83 | 0,79 | 1 | -0,05 | -0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 99 | 1561 | 2.3.4. | 196 | ...аффиксы именные | К | 0,77 | 0,56 | 0,62 | 1 | -0,15 | -0,21 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 100 | 1562 | 2.3.4. | 178 |падеж | Ф | 0,71 | 0,58 | 0,56 | 1 | -0,15 | -0,13 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 101 | 1573 | 2.3.4. | 189 | ..локативные формы имени | К | 0,77 | 0,54 | 0,60 | 1 | -0,17 | -0,23 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 1757 | 2.3.5. | 315 | 2.3.5.ГЛАГОЛЬНЫЕ КАТЕГОРИИ | | | | | | | | | | | | | | | | | |
| 102 | 1759 | 2.3.5. | 253 | .залог | К | 0,77 | 0,81 | 0,80 | 1 | 0,03 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 103 | 1760 | 2.3.5. | 244 | ..категориальный состав кроме активного | К | 0,77 | 0,77 | 0,77 | 1 | 0,00 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 104 | 1768 | 2.3.5. | 160 | ...каузатив | Ф | 0,48 | 0,52 | 0,51 | 0 | 0,03 | 0,04 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 105 | 1777 | 2.3.5. | 183 | ...пассив | К | 0,54 | 0,69 | 0,58 | 1 | 0,04 | 0,15 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| 106 | 1792 | 2.3.5. | 243 | ..способ выражения | К | 0,71 | 0,77 | 0,77 | 1 | 0,06 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 107 | 1795 | 2.3.5. | 212 | ...морфологическая категория | Ф | 0,71 | 0,65 | 0,67 | 1 | -0,04 | -0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 108 | 1814 | 2.3.5. | 303 | .видо-временные категории | К | 0,92 | 0,96 | 0,96 | 1 | 0,04 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 109 | 1815 | 2.3.5. | 294 | ..основные временные противопоставления | К | 0,87 | 0,94 | 0,93 | 1 | 0,06 | 0,08 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 110 | 1822 | 2.3.5. | 201 | ...настоящее/прошедшее/будущее | Ф | 0,42 | 0,67 | 0,64 | 0 | 0,21 | 0,25 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| 111 | 1904 | 2.3.5. | 242 | ..виды прошедшего времени | К | 0,60 | 0,75 | 0,77 | 1 | 0,17 | 0,15 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| 112 | 2081 | 2.3.5. | 278 | ..способ выражения | К | 0,77 | 0,92 | 0,88 | 1 | 0,11 | 0,15 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 113 | 2084 | 2.3.5. | 268 | ...морфологическая категория | Ф | 0,77 | 0,88 | 0,85 | 1 | 0,08 | 0,12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 114 | 2085 | 2.3.5. | 150 | ...служебные слова | Ф | 0,31 | 0,52 | 0,47 | 0 | 0,17 | 0,21 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 115 | 2086 | 2.3.5. | 144 | .способ действия | К | 0,38 | 0,62 | 0,46 | 1 | 0,07 | 0,23 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 116 | 2087 | 2.3.5. | 139 | ..основные значения | Ф | 0,38 | 0,58 | 0,44 | 1 | 0,06 | 0,19 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 117 | 2214 | 2.3.5. | 139 | ..способ выражения | К | 0,37 | 0,62 | 0,44 | 1 | 0,07 | 0,25 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Класс/Факт | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРОПИСЬМЕННЫЙ МОНГОЛЬСКИЙ | АРАБСКИЙ КЛАССИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ |
|----------------------|------------|-----------------|---------|----------------------------------|------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|------------------|-----------------------------|-----------------------|-------|----------|------------|----------|----------------|
| 118 | 2217 | 2.3.5. | 291 | .наклонения кроме изъявительного | К | 0,85 | 0,92 | 0,92 | 0 | 0,07 | 0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 119 | 2218 | 2.3.5. | 291 | ..основные значения | К | 0,87 | 0,90 | 0,92 | 1 | 0,06 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 120 | 2219 | 2.3.5. | 281 | ...повелительное наклонение | К | 0,85 | 0,87 | 0,89 | 1 | 0,04 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 121 | 2220 | 2.3.5. | 211 |императив | Ф | 0,44 | 0,65 | 0,67 | 1 | 0,23 | 0,21 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 122 | 2246 | 2.3.5. | 266 | ...косвенные наклонения | К | 0,79 | 0,85 | 0,84 | 1 | 0,05 | 0,06 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 123 | 2337 | 2.3.5. | 277 | ..способ выражения | К | 0,87 | 0,88 | 0,88 | 1 | 0,01 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 124 | 2338 | 2.3.5. | 269 | ...морфологическая категория | Ф | 0,87 | 0,87 | 0,85 | 1 | -0,01 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 125 | 2449 | 2.3.5. | 206 | .переходность | К | 0,58 | 0,63 | 0,65 | 1 | 0,07 | 0,06 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 126 | 2450 | 2.3.5. | 199 | ..маркер переходности | К | 0,58 | 0,63 | 0,63 | 1 | 0,05 | 0,06 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |

Статистические универсалии. Разделы 2.3.6-2.4.0

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Класс/Факт | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРИТОНСКИЙ
НЬИ МОНГОЛЬ-
СКИЙ | АРАБСКИЙ КЛАС-
СИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ | |
|----------------------|------------|-----------------|---------|---|------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|------------------|--------------------------------------|----------------------------|-------|----------|------------|----------|----------------|---|
| | 2520 | 2.3.6. | 316 | 2.3.6.ДЕЙКТИЧЕСКИЕ КАТЕГОРИИ | | | | | | | | | | | | | | | | | | |
| 127 | 2522 | 2.3.6. | 299 | ..местоимения и местоименные слова | К | 0,90 | 0,92 | 0,95 | 1 | 0,04 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 128 | 2523 | 2.3.6. | 296 | ..разряды местоимений | К | 0,90 | 0,90 | 0,94 | 1 | 0,03 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 129 | 2542 | 2.3.6. | 201 | ...вопросительные | Ф | 0,54 | 0,58 | 0,64 | 1 | 0,10 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | |
| 130 | 2550 | 2.3.6. | 292 | ...личные | К | 0,87 | 0,90 | 0,92 | 1 | 0,06 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 131 | 2563 | 2.3.6. | 157 | ...неопределенные | Ф | 0,25 | 0,52 | 0,50 | 0 | 0,25 | 0,27 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | |
| 132 | 2595 | 2.3.6. | 276 | ...указательные | Ф | 0,81 | 0,81 | 0,87 | 1 | 0,07 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 133 | 2613 | 2.3.6. | 240 | ..дейктический статус | К | 0,69 | 0,77 | 0,76 | 1 | 0,07 | 0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | |
| 134 | 2614 | 2.3.6. | 233 | ..указательные местоимения | К | 0,69 | 0,73 | 0,74 | 1 | 0,05 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | |
| 135 | 2615 | 2.3.6. | 177 | ...ближний план/дальний план | Ф | 0,56 | 0,52 | 0,56 | 0 | 0,00 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | |
| 136 | 2692 | 2.3.6. | 277 | ..грамматическая категория лица | К | 0,83 | 0,92 | 0,88 | 1 | 0,05 | 0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 137 | 2694 | 2.3.6. | 168 | ..в имени | К | 0,58 | 0,67 | 0,53 | 0 | -0,05 | 0,10 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | |
| 138 | 2697 | 2.3.6. | 158 | ...притяжательные аффиксы | Ф | 0,54 | 0,65 | 0,50 | 0 | -0,04 | 0,12 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| 139 | 2699 | 2.3.6. | 269 | ..в глаголе | К | 0,83 | 0,92 | 0,85 | 1 | 0,02 | 0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 140 | 2702 | 2.3.6. | 225 | ...одноличное спряжение | К | 0,73 | 0,87 | 0,71 | 1 | -0,02 | 0,13 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | |
| 141 | 2716 | 2.3.6. | 261 | ..пространственная ориентация действия | К | 0,81 | 0,87 | 0,83 | 1 | 0,02 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 142 | 2717 | 2.3.6. | 260 | ..выражение | К | 0,81 | 0,87 | 0,82 | 1 | 0,02 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 143 | 2733 | 2.3.6. | 149 | ...поселологи | Ф | 0,44 | 0,58 | 0,47 | 1 | 0,03 | 0,13 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | |
| 144 | 2740 | 2.3.6. | 199 | ..определенность/неопределенность имени | К | 0,44 | 0,73 | 0,63 | 0 | 0,19 | 0,29 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | |
| 145 | 2756 | 2.3.6. | 125 | ..синтаксическое выражение | К | 0,27 | 0,56 | 0,40 | 0 | 0,13 | 0,29 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 146 | 2801 | 2.3.6. | 249 | ..отрицание | К | 0,79 | 0,77 | 0,79 | 1 | 0,00 | -0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 147 | 2803 | 2.3.6. | 135 | ..именное >< глагольному | Ф | 0,38 | 0,58 | 0,43 | 1 | 0,04 | 0,19 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | |

| | | | | | | | | | | | | | | | | | | | | | |
|-----|------|--------|-----|--------------------------------------|---|------|------|------|---|-------|-------|---|---|---|---|---|---|---|---|---|---|
| 148 | 2805 | 2.3.6. | 181 | ...лексическое | К | 0,67 | 0,58 | 0,57 | 1 | -0,10 | -0,10 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | |
| 149 | 2807 | 2.3.6. | 169 | ...морфологическое | К | 0,48 | 0,54 | 0,53 | 0 | 0,05 | 0,06 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| 150 | 2811 | 2.3.6. | 236 | ...позиция отрицательного показателя | К | 0,73 | 0,73 | 0,75 | 1 | 0,02 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| | 2828 | 2.3.7. | 316 | 2.3.7.ЧАСТИ РЕЧИ | | | | | | | | | | | | | | | | | |
| 151 | 2831 | 2.3.7. | 303 | ...знаменательные | К | 0,96 | 0,92 | 0,96 | 1 | 0,00 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 152 | 2834 | 2.3.7. | 308 | ...глагол | К | 0,96 | 0,96 | 0,97 | 1 | 0,01 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 153 | 2835 | 2.3.7. | 297 | ...финитные формы | Ф | 0,88 | 0,94 | 0,94 | 1 | 0,06 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 154 | 2836 | 2.3.7. | 267 | ...нефинитные формы | К | 0,71 | 0,83 | 0,84 | 1 | 0,13 | 0,12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 155 | 2854 | 2.3.7. | 253 |причастие | Ф | 0,69 | 0,71 | 0,80 | 1 | 0,11 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 156 | 2866 | 2.3.7. | 299 | ...местоимение | Ф | 0,94 | 0,92 | 0,95 | 1 | 0,00 | -0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 157 | 2867 | 2.3.7. | 261 | ...наречие | Ф | 0,81 | 0,90 | 0,83 | 1 | 0,02 | 0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 158 | 2872 | 2.3.7. | 272 | ...прилагательное | Ф | 0,83 | 0,88 | 0,86 | 1 | 0,03 | 0,06 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 159 | 2875 | 2.3.7. | 286 | ...существительное | Ф | 0,85 | 0,94 | 0,91 | 1 | 0,06 | 0,10 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 160 | 2876 | 2.3.7. | 278 | ...числительное | Ф | 0,85 | 0,90 | 0,88 | 1 | 0,03 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 161 | 2877 | 2.3.7. | 297 | ...служебные | К | 0,92 | 0,92 | 0,94 | 1 | 0,02 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 162 | 2885 | 2.3.7. | 201 | ...послелог | Ф | 0,62 | 0,62 | 0,64 | 1 | 0,02 | 0,00 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 163 | 2894 | 2.3.7. | 271 | ...союз | Ф | 0,92 | 0,85 | 0,86 | 1 | -0,07 | -0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 164 | 2897 | 2.3.7. | 234 | ...частица | К | 0,73 | 0,77 | 0,74 | 1 | 0,01 | 0,04 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 165 | 2899 | 2.3.7. | 215 | ...междометия | Ф | 0,46 | 0,77 | 0,68 | 0 | 0,22 | 0,31 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| | 2902 | 2.4.0. | 315 | 2.4.0.ПАРАДИГМЫ | | | | | | | | | | | | | | | | | |
| 166 | 2917 | 2.4.0. | 298 | ...глагол | К | 0,90 | 0,96 | 0,94 | 1 | 0,04 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 167 | 2918 | 2.4.0. | 297 | ...финитные формы | К | 0,90 | 0,96 | 0,94 | 1 | 0,04 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 168 | 2919 | 2.4.0. | 295 | ...словоизменение | К | 0,88 | 0,96 | 0,93 | 1 | 0,05 | 0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 169 | 2926 | 2.4.0. | 287 |время | Ф | 0,85 | 0,94 | 0,91 | 1 | 0,06 | 0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 170 | 2928 | 2.4.0. | 210 |залог | Ф | 0,65 | 0,75 | 0,66 | 1 | 0,01 | 0,10 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| 171 | 2932 | 2.4.0. | 253 |лицо | Ф | 0,83 | 0,85 | 0,80 | 1 | -0,03 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 172 | 2935 | 2.4.0. | 282 |наклонение | Ф | 0,85 | 0,90 | 0,89 | 1 | 0,05 | 0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 173 | 2952 | 2.4.0. | 273 |число | Ф | 0,81 | 0,88 | 0,86 | 1 | 0,06 | 0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 174 | 3113 | 2.4.0. | 179 | ...причастие | К | 0,37 | 0,52 | 0,57 | 1 | 0,20 | 0,15 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 175 | 3266 | 2.4.0. | 219 | ...прилагательное | К | 0,63 | 0,67 | 0,69 | 1 | 0,06 | 0,04 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 176 | 3267 | 2.4.0. | 216 | ...словоизменение | К | 0,63 | 0,65 | 0,68 | 1 | 0,05 | 0,02 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 177 | 3314 | 2.4.0. | 291 | ...существительное | К | 0,85 | 0,94 | 0,92 | 1 | 0,07 | 0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 178 | 3315 | 2.4.0. | 286 | ...словоизменение | К | 0,83 | 0,92 | 0,91 | 1 | 0,08 | 0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 179 | 3334 | 2.4.0. | 254 | ...падеж | Ф | 0,81 | 0,75 | 0,80 | 1 | 0,00 | -0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 180 | 3345 | 2.4.0. | 148 | ...принадлежность | К | 0,50 | 0,54 | 0,47 | 0 | -0,03 | 0,04 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 181 | 3367 | 2.4.0. | 282 | ...число | Ф | 0,83 | 0,90 | 0,89 | 1 | 0,07 | 0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Статистические универсалии. Разделы 2.5.1-2.5.4

| Номер на графике Ж.1 | Номер в БД | Раздел реферата | Частота | Наименование признака | Класс/Факт | Мертвые языки (МЯ) | Тестовая группа (ТГ) | Вся БД | АВЕСТЫ | Сдвиг_1 (МЯ-Вся БД) | Сдвиг_2 (МЯ - ТГ) | ЛАТИНСКИЙ | ДРЕВНЕЕВРЕЙСКИЙ | ДРЕВНЕПЕРСИДСКИЙ | СТАРОПИСЬМЕННЫЙ МОНГОЛЬСКИЙ | АРАБСКИЙ КЛАССИЧЕСКИЙ | ТЮРКИ | УРАТСКИЙ | ПОЛОВЕЦКИЙ | ХЕТТСКИЙ | ДРЕВНЕОГУЗСКИЙ |
|----------------------|------------|-----------------|---------|--|------------|--------------------|----------------------|--------|--------|---------------------|-------------------|-----------|-----------------|------------------|-----------------------------|-----------------------|-------|----------|------------|----------|----------------|
| | 3409 | 2.5.1. | 314 | 2.5.1.СТРУКТУРА СЛОВОФОРМЫ | | | | | | | | | | | | | | | | | |
| 182 | 3413 | 2.5.1. | 281 | ..типы морфем, присоединяемых к корневым | К | 0,98 | 0,88 | 0,89 | 1 | -0,09 | -0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 183 | 3414 | 2.5.1. | 268 | ..словообразовательные аффиксы | Ф | 0,94 | 0,87 | 0,85 | 1 | -0,09 | -0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 184 | 3415 | 2.5.1. | 273 | ..словоизменяемые аффиксы | Ф | 0,90 | 0,87 | 0,86 | 1 | -0,04 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 185 | 3416 | 2.5.1. | 273 | ..средства словоизменения | К | 0,94 | 0,88 | 0,86 | 1 | -0,08 | -0,06 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 186 | 3429 | 2.5.1. | 271 | ..суффиксы | К | 0,90 | 0,88 | 0,86 | 1 | -0,05 | -0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 187 | 3439 | 2.5.1. | 220 | ..модели словоформы | К | 0,75 | 0,75 | 0,70 | 0 | -0,05 | 0,00 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 188 | 3443 | 2.5.1. | 216 | ..преимущественно суффиксальная | К | 0,75 | 0,75 | 0,68 | 0 | -0,07 | 0,00 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 3593 | 2.5.2. | 313 | 2.5.2.СЛОВООБРАЗОВАНИЕ | | | | | | | | | | | | | | | | | |
| 189 | 3595 | 2.5.2. | 290 | ..деривация | К | 0,96 | 0,94 | 0,92 | 1 | -0,04 | -0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 190 | 3596 | 2.5.2. | 290 | ..аффиксация | К | 0,96 | 0,94 | 0,92 | 1 | -0,04 | -0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 191 | 3604 | 2.5.2. | 283 | ...суффиксы | Ф | 0,94 | 0,90 | 0,90 | 1 | -0,05 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 192 | 3622 | 2.5.2. | 219 | ..сложение основ | К | 0,62 | 0,73 | 0,69 | 1 | 0,08 | 0,12 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| | 3633 | 2.5.3. | 314 | 2.5.3.ПРОСТОЕ ПРЕДЛОЖЕНИЕ | | | | | | | | | | | | | | | | | |
| 193 | 3635 | 2.5.3. | 271 | ..строй | К | 0,77 | 0,73 | 0,86 | 1 | 0,09 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 194 | 3639 | 2.5.3. | 198 | ..номинативный | К | 0,63 | 0,65 | 0,63 | 1 | -0,01 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| 195 | 3652 | 2.5.3. | 282 | ..линейный порядок членов предложения | К | 0,92 | 0,88 | 0,89 | 1 | -0,03 | -0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 196 | 3655 | 2.5.3. | 263 | ..основной | К | 0,85 | 0,83 | 0,83 | 0 | -0,01 | -0,02 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 197 | 3657 | 2.5.3. | 184 | ...SOV | Ф | 0,69 | 0,58 | 0,58 | 0 | -0,11 | -0,12 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 198 | 3662 | 2.5.3. | 231 | ..линейный порядок в группах | К | 0,71 | 0,69 | 0,73 | 0 | 0,02 | -0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |

| | | | | | | | | | | | | | | | | | | | | |
|-----|------|--------|-----|---------------------------------------|---|------|------|------|---|-------|-------|---|---|---|---|---|---|---|---|---|
| 199 | 3663 | 2.5.3. | 218 | ..в именной группе | К | 0,71 | 0,63 | 0,69 | 0 | -0,02 | -0,08 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 200 | 3711 | 2.5.3. | 208 | ..согласование | К | 0,54 | 0,58 | 0,66 | 1 | 0,12 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 201 | 3712 | 2.5.3. | 193 | ..актант определяет форму предиката | К | 0,52 | 0,56 | 0,61 | 1 | 0,09 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 202 | 3713 | 2.5.3. | 184 | ...субъектное согласование | К | 0,50 | 0,54 | 0,58 | 1 | 0,08 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 203 | 3715 | 2.5.3. | 158 |по лицу | Ф | 0,46 | 0,50 | 0,50 | 1 | 0,04 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 204 | 3720 | 2.5.3. | 173 |по числу | Ф | 0,50 | 0,52 | 0,55 | 1 | 0,05 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 205 | 3752 | 2.5.3. | 183 | ..коммуникативно значимые средства | К | 0,44 | 0,62 | 0,58 | 1 | 0,14 | 0,17 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| | 3773 | 2.5.4. | 312 | 2.5.4.СЛОЖНОЕ ПРЕДЛОЖЕНИЕ | | | | | | | | | | | | | | | | |
| 206 | 3807 | 2.5.4. | 259 | ..тип построения | К | 0,81 | 0,85 | 0,82 | 1 | 0,01 | 0,04 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 207 | 3809 | 2.5.4. | 248 | ..сочинение/подчинение | Ф | 0,71 | 0,81 | 0,78 | 1 | 0,07 | 0,10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 208 | 3812 | 2.5.4. | 264 | ..тип связи | К | 0,81 | 0,83 | 0,84 | 1 | 0,03 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 209 | 3813 | 2.5.4. | 208 | ..союзная/бессоюзная | Ф | 0,50 | 0,63 | 0,66 | 1 | 0,16 | 0,13 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| 210 | 3816 | 2.5.4. | 223 | ..союзы | К | 0,65 | 0,65 | 0,71 | 1 | 0,05 | 0,00 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 211 | 3820 | 2.5.4. | 216 | ...самостоятельные служебные элементы | Ф | 0,62 | 0,63 | 0,68 | 1 | 0,07 | 0,02 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |



Рис. 3.1 Распределение используемых в одном из расчетов 11 мертвых языков по оси времени

Условные обозначения

БД – база данных
ИЕ – индоевропейские
ИЯ РАН (ИЯз РАН) – Институт языкознания Российской Академии Наук
КГУ – Казанский государственный университет
Класс. – классифицирующий
МБ – мегабайт
МГЛУ – Московский государственный лингвистический университет
МГУ – Московский государственный университет
МР – модель реферата
МЯ – мертвые языки
НОЦ – Научно-образовательный центр
Н.э – наша эра
ОТиПЛ – Отделение теоретической и прикладной лингвистики
Прим. – примечание
РГНФ – Российский гуманитарный научный фонд
Рис. – рисунок
Совр. – современный
СУБД – система управления базой данных
Табл. – таблица
ТГ – тестовая группа
Факт. – фактографический
ФГУП НТЦ – федеральное государственное унитарное предприятие «Научно-технический центр»
ЮУрГУ – Южно-Уральский государственный университет
ЯП – язык-признак
BDE – Borland DataBase Engine
DOS (MS DOS) – Disk Operational System фирмы Microsoft
LENCA – Languages of Europe, North, and Central Asia
URL – Universal Resource Locator
WALS – The World Atlas of Language Structures
Win – Windows

Предметный указатель

- Data Mining. см. Интеллектуальный анализ данных
- LENCA, 12
- WALS – The World Atlas of Language Structures, 7
- База данных Языки мира
- Web-версия, 41
 - Windows-версия, 35
 - история создания, 22
 - контент, 22
 - модель реферата, 26
 - назначение, 43
 - принципы организации, 26
 - программное обеспечение, 35
 - языковые семьи и языки, 23
- Генетические гипотезы
- методика проверки, 90
 - обзор, 11
- Интеллектуальный анализ данных, 64
- Квантитативные портреты по языковым семьям и ветвям, 87
- Классификация признаков по частотным основаниям, 88
- Классификация языков, 9
- ареальный подход, 10
 - генеалогический подход, 9
 - типологический подход, 11
- Кластеризация языков
- общие сведения, 57
 - описание алгоритмов, 59
- Компаративистика, 7
- Концепция зон
- spread и residual, 13
- Маркеры генетического родства, 88
- Меры близости
- методика оценки качества мер, 50
 - общие подходы, 45
 - результаты расчетов, 51
 - сводная таблица, 49
 - учет структуры признакового пространства, 46
- Модель структурной эволюции языков, 71
- Сравнительно-исторический метод, 9
- Типологические (структурные) свойства языков
- по Г. Сильницкому, 18
 - по Дж. Николс, 13
- Типология, 7
- Феномен типологического сдвига на евразийском пространстве, 106
- ЯП-диаграммы, 75

Именной указатель

Арапов А.Е. 92, 133
Ахтямов Р.Б. 27, 133
Балалыкина Э.А. 12, 133
Бодуэн де Куртенэ И.А. 1, 7, 12, 27, 74, 133
Бугаева А. 12, 133
Бурлак С.А. 11, 12, 24, 28, 30, 30, 31, 85, 91, 96, 133, 186
Виноградов В.А. 9, 27, 133, 186
Глас Р. 85, 133
Гончаров Е. 27
Двоеносова М.В. 70
Дресслер В. 13, 133
Жамбю 70, 133
Журинская М. А. 27, 133
Завьялова М.В. 133
Звягинцева В.А. 133
Зотова А.К. 27
Кибрик А.А. 7, 9, 133, 186
Кибрик А.Е. 9
Климов Г.А. 13, 133
Комри Б. 7, 14, 24, 133
Коряков Ю.Б. 103
Логунов В.В. 27, 133, 186
Молдован А. М. 133, 186
Мэттьюс С. 133
Нестерова Н.Н. 27
Никуличева Д.Б. 39
Новиков А.И. 27, 133, 186
Полински М. 133
Пенроуз Р. 96, 133
Поляков В.Н. 2, 27, 39, 54, 62, 70, 103, 133, 186
Раушенбах Г.В. 55, 133
Рогова Н.В. 27, 133
Романова О.И. 27, 133
Рябцева Н.К. 27
Савельев А.А. 27, 77, 78, 133
Сводеш М. 8, 11, 85, 133
Сильницкий Г. Г. 21-23, 28-31, 133, 186
Скокан Ю.П. 27
Скорвид С.С. 133, 186
Соловьев В.Д. 2, 27, 54, 62, 70, 103, 133, 186
Старостин С.А. 11, 12, 24, 28, 30, 30, 31, 85, 91, 96, 133, 186
Тестелец Я.И. 27
Топоров В.Н. 133
Феллер В. 95, 133
Херц М.М. 92, 133
Хомутова Т.Н. 9
Чемоданова Н.С. 133
Шахматов А.А. 12, 133
Щербинин Т. 27
Якобсон Я. 13, 133

Ярославцева Е.И. 7, 9, 27, 32, 103, 133, 186
Ярцева В.Н. 7, 27, 133
Bishop C.M. 80, 133
Croft W. 11, 92, 96, 133
Dahl Ö. 102, 133
Janhunen Ju. 94, 133
Givón T. 11, 133
Keenan E. 13, 133
Kruskal J.B. 77, 133
Kohonen T. 80, 133
Laudan L. 96, 133
Legendre L. 78, 133
Legendre P. 78, 133
Mithun M. 13, 86, 133
Nichols J. 15-21, 84-87, 133
Rosch E. 94, 133
Schleicher A. 11, 133
Shopen T. 133
Svensen M. 133
Williams C.K.I. 133

Computer models and methods in Typology and Comparative Linguistics

Vladimir Polyakov and Valery Solovyev

The book by V.N. Polyakov and V.D. Solovyev «Computer models and methods in Typology and Comparative Linguistics» is devoted to the detailed description of the database «Languages of the World» and the ways of its usage. The database is one of the two greatest typological databases in the World. It was created on the basis of the cognominal series of monographs in the Institute of Linguistics of the Russian Academy of Science and contains the representation of 315 languages, mostly languages of Eurasia according to more than 3800 features.

This monograph introduces the complex of mathematical methods of analyses of multidimensional space of the language data. The ultimate aim is the work out of methods of analyses which would help to decide a point of the origin of languages, to identify the proximity of languages as genetic or as taking its genesis from borrowings, to find out the general complex of origin and diffusion of language structures.

The main advantage of the amplified approach is the usage of the vast range of linguistic data and application of certain mathematical methods for its analyses. Usually languages are classified on the basis of a few parameters. The regularities of evolution mostly of the phonetic level of language lie at the root of comparative researches. The database «Languages of the World» presents the opportunity to use all the existing array of the data about the structure of languages that gives the new prospect in classification of languages.

Different measures of proximity of languages had been decomposed in detail that is the main instrument for assorted researches. The reckoning were made owing to expounded methods, the first elicited new statistic regularity of language evolution. The phenomenon of typological shift on the Eurasian space was revealed. The core of it is in growth of frequency of frequent features and elimination of rare ones in the process of evolution (within the last 2-3 thousand of years).

The systematization of description of languages in the form of the database gave an option to introduce a new definition Language-Feature diagram (graphic chart which reflects the law of number of features from the degree of their diffusion). This definition is a strong instrument of analyses of mechanisms of borrowings.

The given facts and regularities present the opportunity to consider the phenomenon of the language diversity as anew and to form a mathematical model of structural evolution of languages. The methods of verification of genetic hypothesis were introduced and they underlie in the usage of rare features. Thanks to these methods new data were obtained that can be used for clarification of origin of isolate languages.