Разработки 2015 г. по теме
«Новая усовершенствованная версия Базы знаний БД ИЯз РАН "Языки мира" как уникальный фактографический и программный аппарат для типологических, компаративных, квантитативных, квалитативных и прикладных лингвистических исследований»

исполнители А.К.Зотова, В.Дьячков, Г.А.Черкасова.

Терминосистема лингвотипологической БД ИЯз РАН "Языки мира": Проект синонимического терминологического русско-английского указателя

Работа ведется одновременно в практическом и теоретическом направлениях. Так в частности, для разработки оптимальной модели синонимического терминологического русско-английского указателя оказалось необходимым рассмотреть значительное количество потенциальных вариантов, используемых в отечественных и зарубежных источниках в диапазоне от учебных, отраслевых и других указателей в традиционном печатном виде до новейших электронных ресурсов.

На предварительном этапе была проведена тщательная экспертиза, как формата представления, так и терминологических фрагментов содержания БД. Были проанализированы источники неточностей и ошибок, связанные с *адаптацией* описания явлений и признаков, имеющихся в Энциклопедии, к требованиям реферата и модели реферата в его окончательном виде. Очевидно что, с одной стороны, модель реферата включает в себя терминологию различных лингвистических школ в области типологии и компаративистики. С другой стороны, требования унифицированного описания приводят к заданной в модели реферата формальной классификации и искусственному отнесению признаков к той или иной позиции, что в ряде случаев становится причиной неадекватного представление информации. В этом смысле, уточнение терминологии с помощью ряда возможных синонимов и английских эквивалентов может стать одним из инструментов решения проблемы возникновения *артефактов.*

В 2015 г. анализ терминосистемы БД «Языки мира» включал в себя продолжение исследования терминологии оригинальных статей, написанных авторами для Энциклопедии “Языки мира”, справочного аппарата к отдельным томам энциклопедии и терминологии соответствующих разделов рефератов и общей модели реферата БД.

Были выявлены значительные расхождения в объеме и структуре терминологических указателей в изданиях разных лет, причем основная тенденция состоит в расширении и диверсификации справочного аппарата.

Так, в более ранних публикациях материал представлялся преимущественно в виде списка сокращений терминов, относящихся ко всем языкам данного тома, и структурно сохранял или частично воспроизводил

элементы модели реферата БД с незначительными изменениями (использование аббревиатур, алфавитного порядка), например:

Тюркские языки - терминологический указатель в виде списка сокращений с сохранением структуры модели реферата БД

…
вр[емя]
  буд[ущее]
  давнопрош[едшее]
  наст[оящее]
  прош[едшее]
…
накл[онение]
  усл[овное]
…
п[адеж]
  вин[ительный]
  дат[ельный]
  дат.-вин. – дательно-винительный
  дат.-местн. – дательно-местный
  дат.-напр. – дательно-направительный
  дир[ективный]
  им[енительный]
  инстр[ументальный]
  исх[одный]
  лишит[ельный]
  местн.-врем. – местно-временно́й
  местн[ый]

варианты синонимии:

орудный=орудийный падеж

В 2015 году материал был значительно расширен и дополнен за счет анализа новых публикаций Энциклопедии, в частности томов “Семитские и эфиосемитские языки”, “Реликтовые индоевропейские языки Передней и Центральной Азии” и др.

Было обнаружено, что помимо традиционных указателей в этих изданиях использованы качественно отличные терминологические описания языковых явлений. В частности, наряду с указателями в виде сокращенных вариантов грамматических терминов, в них включены т.н. *пометы при языковых примерах* и *глоссы*. В совокупности они представляют собой фрагмент глоссария лингвистических терминов, но обладают рядом особенностей. Обнаруживаются значительные количественные и качественные расхождения с терминосистемой компьютерной версии БД по таким параметрам как:

- комбинированное использование шрифтов и алфавитов;

- использование терминологических маркеров в виде буквенных символов на базе латинского и русского алфавитов (N – любой носовой согласный, R – любой сонорный согласный, КО – косвенная основа);

- введение обобщающих квази-терминов в виде аббревиатур на базе латинского и русского алфавитов (POST – послелог/ПОСЛ - послелог);

- употребление синонимов (медиальный=медий и др).

- несовпадения терминообразовательных моделей (не-каузатив или т.каузатив= только каузальная форма глагола).

В целом можно отметить тенденцию к унификации терминообразовательных схем:

- использование родовых категорий вместо видовых, свойственное европейским языкам (энклитическое местоимение = энклитика);

- калькирование иноязычных терминообразующих элементов и структур (не-каузатив);

- использование латинизированных терминов и др.

В практическом направлении продолжались исследования и разработка лексикографической модели терминологического указателя в виде словаря синонимов (полных и/или частичных) и общего русско-английского/англо-русского терминологического указателя. Обобщался и классифицировался фактический материал, анализировались возможные варианты его представления:

- русско-английский терминологический указатель с учетом формата модели реферата - указатель тезаурусного типа, например:

Абруптив[ные] <ларингальные признаки< шумные< согласные<фонемный состав – **Abruptives;**

- алфавитный русско-английский синонимический указатель лингвистических терминов с формированием словника путем фильтрации терминов из всего корпуса описания в формате модели реферата, например:

Абессив – абессив**,** лишительный падеж, изъявительный падеж. Abessive.
Аблатив – творительный падеж, относительный падеж, отделительный падеж, удалительный падеж, внешне-местный падеж – Ablative;

- алфавитный русско-английский синонимический указатель лингвистических терминов с элементами толкового и переводного глоссария, например:

Аблатив 1 – Аблатив 8 - формы относительного падежа – Ablative 1-Ablative 8
Абсолютив – падеж субъекта непереходных и объекта переходных глаголов -.Absolutive case
…
Агглютинативные языки – языки, в которых грамматические отношения передаются с помощью суффиксов - Agglutinative languages

В целом, результаты теоретического анализа терминологических расхождений между моделями описания явлений в семьях и группах языков могут быть использованы для решения целого ряда других практических задач:

- верификации сведений о языках и выявление адекватности их описания в БД «Языки мира»;

- повышения качества работы по пополнению и редактированию БД «Языки мира» на этапе перехода от текстового описания к формальному представлению данных о языках;

- оптимизации применения БД «Языки мира» в дидактических целях при интерпретации формальных описаний с использованием терминов и классификаций, не представленных в типовой модели реферата.

Систематизация контента БД ИЯз РАН "Языки мира"

В рамках работы над БД «Языки мира» продолжается работа над теоретически значимыми аспектами базы данных. В результате выполненной в 2014 году тотальной проверки признакового пространства базы данных стало возможным постепенное видоизменение ресурса и повышение его ценности как справочной базы данных.

Работы над БД в 2015 году велись в следующих направлениях:

1) откорректированы инструкции по исправлению ошибок и неточностей с учетом возможностей архитектурной реализации новой версии базы данных; 2) составлены рекомендации по реализации отдельных компонентов базы данных в новой версии (список языков, глоссарий).

Поскольку концепция дальнейшего развития базы данных предполагает ее реализацию средствами программного обеспечения, написанными на языке Java (в отличие от предыдущих версий, использующих другие языки программирования), необходима первичная подготовка данных для того, чтобы их можно было использовать в новой версии БД.

Подготовка данных на текущий момент ведется независимо от реализации новой версии БД средствами программного обеспечения.

В рамках коррекции инструкций по исправлению ошибок и неточностей были проведены следующие предварительные работы:

а) выявлен список языковых признаков, которые в базе данных имеют частоту 1 (т. е. выявлены только в одном языке); эти признаки в силу низкой частотности проанализированы отдельно, не имеющие большого значения признаки удалены;

б) выявлен список языков, которые в базе данных отражены на основании недостоверных данных, список диалектов, которые нельзя отразить в базе данных как отдельные значимые единицы языкового континуума, а также – отдельно – список мертвых языков.

Для предполагаемой версии БД «Языки мира» на языке Java требуется подготовить текстовые файлы с откорректированным пространством признаков для каждого языка. С этой целью были привлечены результаты тотальной проверки признакового пространства, произведенной в 2013 – 2014 гг. Результаты этой проверки были проанализированы повторно и

формализованы с целью предоставить однозначную инструкцию для исполнителей работ по корректировке пространства признаков.

На данный момент полностью реализована подготовка инструкции по разделу «Синтаксис». В течение года проводилась также подготовка по разделам «Фонология» и «Морфология», однако полная реализация инструкций по этим разделам базы данных запланирована на 2016 год.

Пример реальной инструкции по заполнению базы данных приведен ниже.

признак 4 уровня	признак 5 уровня	признак 6 уровня	действие над признаком	комментарий
предикат определяет форму актанта	объектное согласование	по модальности	удалить	
	субъектное согласование	по модальности	удалить	
		по роду	удалить для: ишкашимский язык	
		по числу	удалить для: ишкашимский, белуджский язык	для удинского языка нет статьи в базе
	субъектно-объектное согласование	по времени	удалить	

Составление инструкции потребовало консультации с исходным для базы текстом энциклопедии «Языки мира», а также внешними источниками с целью верификации спорных данных. Эти работы также не проводились во время тотальной проверки признакового пространства.

Полученные результаты были скоординированы внутри рабочей группы, обсуждены спорные моменты теоретического подхода к удалению признаков.

В рамках работы над базой данных были предложены также решения проблем, связанных с несовершенством глоссария. Работы над глоссарием были сосредоточены в основном в направлении общетеоретических вопросов, связанных с универсальностью используемых в базе данных лингвистических терминов.

В ходе теоретической работы было проанализировано представление отдельных признаков в признаковом пространстве (вид глагола и слоговые структуры). Были приняты решения о реструктуризации этих признаков: так, для слоговых структур было принято решение об ином способе представления признака, для вида глагола – о полной смене наименований признаков для всех языков. При этом были выделены компетентные

источники, которые приняты за отправную точку работ по смене наименований.

В рамках работы над списком языков были выделены языки, для которых выделены параметры, затрудняющие работу с языковыми данными. К числу таких параметров относится, например, упоминание в базе данных диалектных континуумов, не соотнесенных с конкретными языками, наличие непроверяемых данных о языках, которые не охвачены серией энциклопедий «Языки мира». Кроме того, для отдельных языков указаны субареалы их распространения, которые не влияют существенным образом на представления об ареальном распространении (например, для некоторых языков выборочно были указаны страны, где проживают диаспоры, говорящие на этом языке).

Наименования государств в разделе базы данных, посвященном ареалам распространения языков, были также тщательно проанализированы. К видоизменению характера представления в базе данных были рекомендованы названия государств, соответствующие ареалам распространения вымерших языков.

Методы визуализации материалов БД ИЯз РАН "Языки мира" и результатов проводимых исследований

В новой усовершенствованной версии Базы знаний "Языки Мира" данные, представленные в БД, связываются с фрагментами книжного энциклопедического издания записанными в формате PDF. Это, пожалуй, самый наглядный способ визуализации, когда текст из полиграфического издания соответствующего тома представляется постранично. Графическое изображение привязано к элементами базы данных по конкретному языку, что позволяет интерпретировать структурные элементы БД по тексту томов.

Для сравнительного и сопоставительного анализа конкретных параметров языков в БД исследовалось применение способов графической визуализации данных, хранимых в БД ИЯз РАН «Языки мира». Для наглядного отображения статистических характеристик предлагается для сопоставительных исследований двух языков использовать визуализацию их такими способами как вертикальные столбиковые диаграммы и графики по одной или двух осям абсцисс (см. пример ниже). Для визуализации характеристик множеств языков и распространенности конкретных параметров использовать круговые диаграммы, при условии когда их количество не превосходит двух десятков.

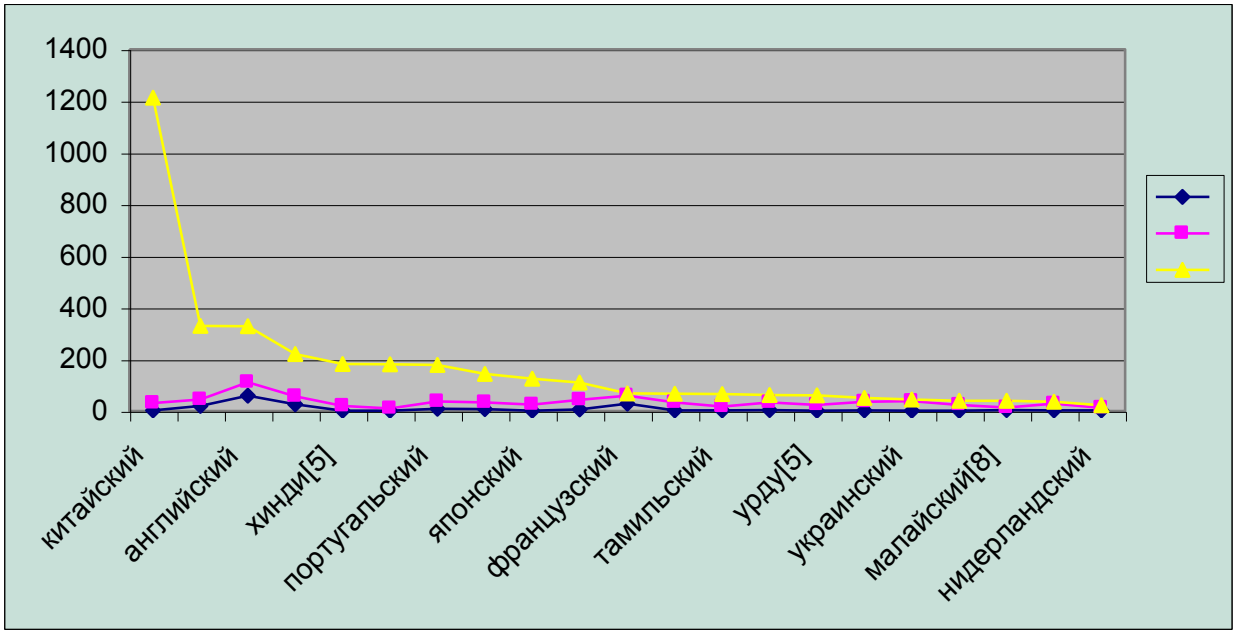
Так как языки привязаны к странам и территориям, то можно использовать статистические карты, которые представляют собой вид графических изображений на схематической географической карте, характеризующих уровень или степень распространения того или иного языка или диалекта на определенной территории. Однако, если территория распространения языка очень невелика, то на общей карте Земли, ее трудно зафиксировать, и этот способ удобен для основных языков мира.

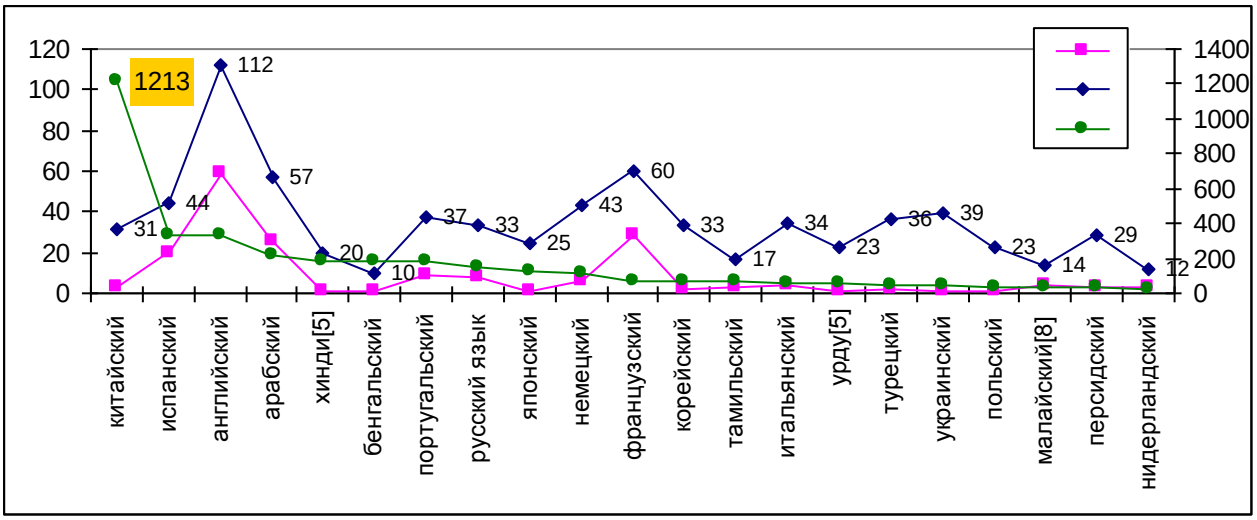
Средствами изображения территориального размещения являются штриховка, фоновая раскраска или геометрические фигуры. Используют картограммы и картодиаграммы. Первые штриховкой различной густоты, точками или окраской определенной степени насыщенности показываются величину исследуемого показателя в пределах ареала, нанесенного на карту. Вторая группа статистических карт – картодиаграммы, сочетающие диаграммы с географической картой. В качестве изобразительных знаков в картодиаграммах используются диаграммные фигуры (столбики, квадраты, круги, фигуры, полосы), которые размещаются на контуре географической карты.

В качестве примера графическое представления рассмотрим основные мировые языки по количеству носителей, количеству стран, в которых язык является официальным и где используется и др. Данные представлены в таблице.

ЯЗЫК	Является родным для (в млн.)	Число стран, где он официальный	Число стран, где используется	Процент сайтов	Процент мирового ВВП	Перечень стран, где язык является официальным
китайский	1213	3	31	4,5	12,5	Китай, Тайвань и Сингапур
испанский	329	20	44	4,8	6,5	Испания, Иберо-Америка (кроме Бразилии), Экваториальная Гвинея
английский	328	59	112	54,8	29,3	Великобритания, США, Индия, Австралия, Канада, Ирландия, Новая Зеландия, ЮАР и др.
арабский	221	26	57	1,2	2,5	страны Северной Африки и Ближнего Востока, см. подробнее Арабский мир
хинди	182	1	20	0,1	2,3	Индия, Фиджи
бенгальский	181	1	10	0,1	1	Бангладеш, индийские штаты и территории Западная Бенгалия, Трипура, Андаманские и Никобарские острова.
португальский	178	9	37	2,3	3,3	Португалия, Бразилия, Ангола, Мозамбик и др. страны Содружества
русский язык	144	8	33	6,1	2,6	Россия, Белоруссия, Казахстан, Киргизия
японский	125	1	25	4,2	7	Япония
немецкий	110	6	43	5,4	5,5	Германия, Австрия, Швейцария, Лихтенштейн, Бельгия, Люксембург
французский	67,8	29	60	4,4	4,6	Франция, Канада, Бельгия, Швейцария, Люксембург, Монако и др.
корейский	66,3	2	33	0,3	1,7	Северная Корея, Южная Корея
тамильский	65,7	3	17	0,1		Сингапур, Шри-Ланка
итальянский	61,7	4	34	1,5	3,2	Италия, Швейцария, Сан-Марино, Ватикан
урду	60,6	1	23	0,1		Пакистан
турецкий	50,8	2	36	1,4	0,9	Турция, Северный Кипр[7]
украинский	45	1	39	0.1	0.4	Украина
польский	40	1	23	1,8	0,9	Польша
малайский	39,1	4	14	0,3	1,4	Малайзия, Индонезия, Бруней, Сингапур
персидский	35,9	3	29	0,1		Иран, Афганистан, Таджикистан

На двух графиках показаны перечисленные выше показатели с использованием одной и двух осей Y. Если на вешнем рисунке с заданной максимальной величиной два других графика практически совпадают, то на нижнем графике, имеющим две оси (правая характеризует количество носителей данного языка, левая – два других параметра) картина абсолютно другая.





Представленные графики позволяют увидеть, что графическая визуализация требует специального исследования и доработки программного обеспечения для ее реализации на основе анализа количественных характеристик, записываемых в Базы данных.