СОВРЕМЕННЫЕ ЛИНГВИСТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ ДЛЯ ИССЛЕДОВАНИЯ ТЕРМИНОЛОГИИ

Гатиатуллин А.Р., кандидат технических наук
Кириллович А.В.,
Хакимов Б.Э., кандидат филологических наук

MODERN LINGUISTIC INFORMATION RESOURCES FOR THE STUDY OF TERMINOLOGY

Gatiatullin A.R., Kirillovich A.V., Khakimov B.E.

Задача разработки моделей связывания терминосистем в языках носит фундаментальный характер и ставит главной целью исследование набора и структуры фундаментальных признаков, позволяющих соотносить терминологию в различных языках.

С появлением и развитием новых областей и форм деятельности, с расширением международной интеграции в различных сферах активно обновляется лексикон национальных языков народов РФ, в том числе и терминология. Новая лексика строится на базе национальных языков с их национально-специфическими особенностями. При этом факторы взаимодействия языков оказывают существенное влияние на развитие терминологии в конкретных сферах.

Несмотря на то, что терминология в татарском языке достаточно хорошо разработана для отдельных областей знаний и сфер деятельности, до недавнего времени отсутствовали информационные ресурсы исследовательского типа, представляющие для пользователя татарские термины с их вариантами использования, и не было создано комплексной интегративной модели терминосистемы татарского языка.

В рамках проекта Российского научного фонда № 16-18-02074 «Разработка моделей связывания терминологии в разных языках (на материале русского и татарского языков)» были созданы электронные ресурсы, аккумулирующие актуальную татарскую терминологию в общественно-политической и IT-областях и дающие новые возможности для ее изучения.

В ходе реализации проекта проекта были разработаны новые лингвистические ресурсы, размещенные в открытом доступе:

– русско-татарский тезаурус общественно-политической (10000 концептов) и IT-лексики (3000 концептов) (http://tattez.antat.ru/);

– татарско-русский словарь сочетаемости общественно-полити-

ческой лексики (4000 коллокаций) (http://spdict.turklang.tatar/);

– лингвистический ресурс типа WordNet для именной лексики татарского языка (5000 существительных).

– тематические подкорпусы – размеченные коллекции документов на татарском языке в общественно-политической и IT-областях (24 млн. словоформ) (http://tugantel.tatar/corpus/op/, http://tugantel.tatar/corpus/it/);

– параллельный русско-татарский корпус общественно-политической лексики и языка информационных технологий (4,5 млн. словоформ) (http://tugantel.tatar/corpus/it, http://tugantel.tatar/parallel/).

Проект базируется на гипотезе о том, что языковые выражения и их значения позволяют выделить основные, существенные для современной жизни людей смыслы и реалии. Это достигается с помощью тезаурусного моделирования – исследования терминов и построения моделей, отражающих сложные иерархические и внутрикомпонентные связи терминов.

Таким образом, основное внимание коллектива исполнителей было сосредоточено на разработке компьютерного тезауруса как центрального элемента группы взаимосвязанных лингвистических ресурсов.

Русско-татарский тезаурус общественно-политической и IT-лексики базируется на тезаурусе русского языка РуТез (http://www.labinform.ru/pub/ruthes/) и включает понятия-концепты из различных областей.

При проектировании татарской части тезауруса в целом сохраняется структура концептуальных связей РуТез, при этом важной задачей является отображение специфики лексико-семантической системы татарского языка. Это достигается тремя основными способами.

1). Поиск татарских переводных эквивалентов (соответствий) для концептов тезауруса РуТез;

2). Добавление новых концептов, отображающих не представленные или представленные неполно в тезаурусе РуТез тематические блоки (например, понятия, связанные с исламом; социальная иерархия традиционного общества; татарские этнокультурные реалии и т.п.);

3). Достраивание концептов промежуточных уровней, не выраженных в русском языковом мышлении, но выявляемых при сопоставлении с татарским языком (например, в русском языковом сознании имеются концепты «пасынок» и «падчерица», но не представлен обобщающий концепт, обозначающий ребенка одного из супругов вне зависимости от пола – такой концепт – «үги бала» – имеется в татарском языке).

Перевод лексических вариантов концептов – текстовых входов предполагает по возможности максимальный охват существующих вариантов – параллельных наименований одного и того же явления (реалии), функционирующих в татарском языке. Для этого

анализируются словарные статьи специальных словарей и словарей общего назначения, ведется поиск лексического материала в реальных текстах общественно-политической тематики на татарском языке. Поскольку языковая ситуация в Республике Татарстан находится в состоянии развития, для одних и тех же реалий могут использоваться параллельные названия, которые фиксируются в тезаурусе при условии ненарушения законов грамматической и лексической систем татарского языка. Так, на рис. 1 мы можем увидеть варианты перевода сочетания «Конституционный суд» и частотности их употребления в корпусе.

Рис.1. Варианты перевода термина «конституционный суд» на татарский язык.

Система электронного тезауруса реализована в формате веб-приложения (http://tattez.antat.ru/) и содержит открытую (поиск и навигация) и закрытую (функции добавления, правки и удаления единиц) части.

Частотные двухкомпонентные термины и их варианты (синонимы) фиксируются также в «Татарско-русском словаре коллокаций» (устойчивых сочетаний слов), который основывается на корпусных данных. В текущей версии словаря представлено свыше 4000 сочетаний. Материалы словаря дают возможность получить актуальные татарские переводы слов и словосочетаний, используемых в современном общественно-политическом дискурсе, отслеживать актуальные тенденции в функционировании та-

тарского языка, достаточно полно и детально отслеживать новые слова и сочетания, а также многочисленные синонимические номинации.

Разработанные ресурсы позволяют заключить, что характерной чертой современной татарской общественно-политической лексики является наличие большого числа синонимов. Локализация татарской культуры на пересечении Запада и Востока является основной причиной сосуществования слов тюрко-татарского, арабо-персидского, русского или западноевропейского происхождения, обозначающих одно и то же понятие. Широкие возможности для исследования такой синонимии в татарском языке как раз и предоставляют корпусные технологии. Анализ контекстов употребления синонимов в речи позволяет определять степень близости синонимов, возможность их взаимозамены, приоритеты использования носителями языка и т.д.

В рамках проекта были разработаны методы связывания терминологии на лингвистическом уровне. Выявлены основные словообразовательные и синтаксические модели, используемые при переводе русских терминов, описаны основные случаи изменения структуры термина при переводе, получены количественные данные о распределении грамматических моделей. Так, русские сложные слова, как правило, в татарском языке имеют двухкомпонентные переводные соответствия: «крестоносец» – «тәре йөртүче», «законопроект» – «закон проекты», «налогообложение» – «салым салу».

На рисунках 2–3 приведена статистика по терминам IT-области, которая фиксирует использование заимствований, а также собственных слов в созданном IT-подкорпусе татарского языка на выборке из 1583 терминов.

Рис. 2. Происхождение татарских IT-терминов.

На рис. 3 мы видим соотношение полных и частично освоенных заимствований (т.е. имеющих в своем составе словообразовательные элементы татарского языка), а также собственно татарских слов среди терминов IT-сферы с учетом их частотности в корпусе. Очевидно, что незаимствованные слова составляют большую часть, т.е. активно идет процесс создания новых терминов на национальном языке. Но при учете частоты употребления в текстах соответствующей тематики преобладают уже заимствованные слова. Это значит, во-первых, что заимствованная посредством русского языка интернациональная лексика составляет активное ядро; во-вторых, это свидетельствует о высокой доле терминов-синонимов и вариативности в татарском языке.

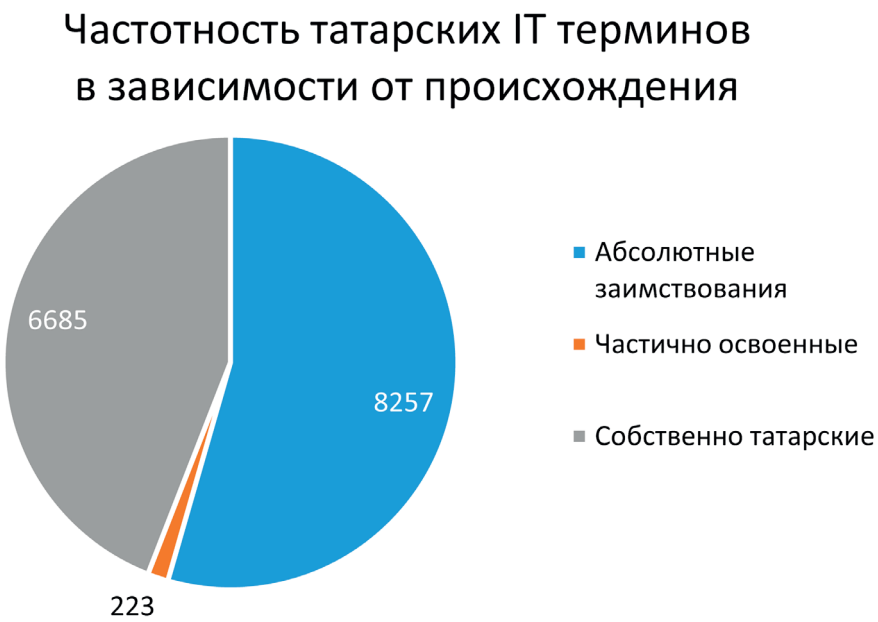

Рис. 3. Частотность татарских IT терминов в зависимости от происхождения.

Одним из возможных подходов к интеграции знаний о языке и мире является публикация лингвистических ресурсов на основе технологий семантической паутины в облаке лингвистических открытых связанных данных (LLOD). Лингвистические ресурсы для многих языков мира уже опубликованы в облаке LLOD, однако ресурсы для русского языка и языков народов России представлены в нем лишь фрагментарно. Чтобы восполнить этот пробел, в проекте разработано новое облако RuThes Cloud, которое является многоуровневым ресурсом лингвистических открытых связанных данных для русского языка и языков народов России. В настоящее время этот ресурс включает следующие компоненты:

1) сеть понятий (концептов), связанных между собой отношениями «род–вид», «часть–целое», ассоциации;

2) лексические единицы (отдельные слова или многословные выражения) на русском, английском и татарском языках, обозначающие концепты;

3) грамматические формы лексических единиц;

4) синтаксические деревья для многословных лексических единиц;

5) семантические и синтаксические фреймы (семантический фрейм содержит описание семантических ролей, связанных с некоторой ситуацией; синтаксический фрейм содержит описание модели управления лексической единицы и ее связь с семантическими ролями);

6) связи между лексическими единицами, такие как антонимия, деривация, коллокация и др.;

7) связи с внешними ресурсами из облака открытых связанных данных.

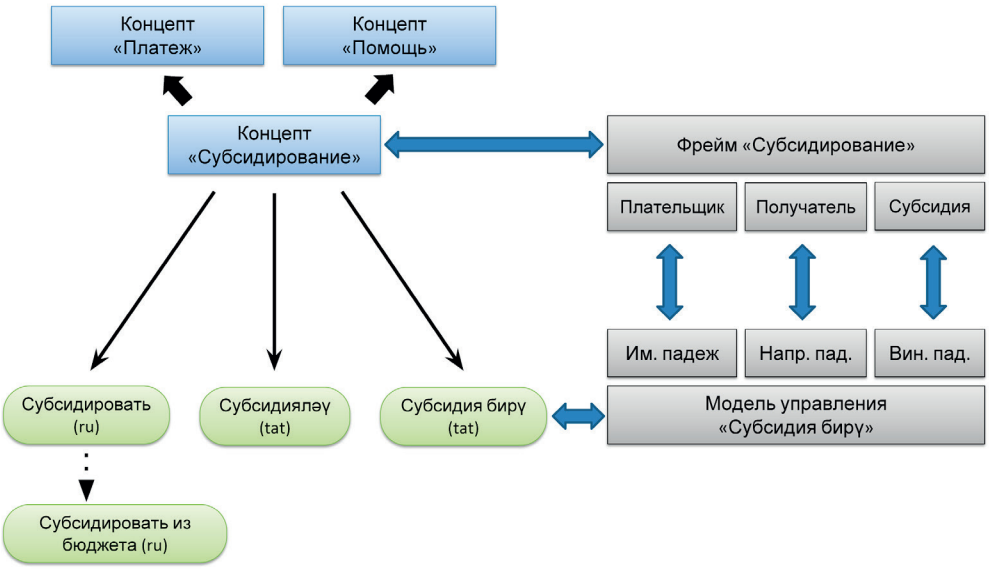

Рис. 4. Фрагмент ресурса RuThes Cloud.

На рис. 4 изображен фрагмент ресурса RuThes Cloud, содержащий концепт «Субсидирование» и некоторые связанные с ним элементы. Концепт «Субсидирование» связан с вышестоящими концептами «Платеж» и «Помощь». Кроме того, данный концепт связан с семантическим фреймом, который представляет ситуацию субсидирования с семантическими ролями «плательщик», «получатель» и «субсидия». Наконец, концепт «субсидирование» связан с лексическими единицами на русском и татарском языках, служащими для обозначения данного концепта, с их моделями управления.

Разработанная интегральная модель позволяет изучать взаимосвязи терминологических единиц в разных языках, их контекстное окружение и тем самым, обогащать наше знание о функционировании лексических систем разных языков.

Результаты проекта РНФ № 16-18-02074 «Разработка моделей связывания терминологии в разных языках (на материале русского и татарского языков)» имеют не только научную, но и общественную значимость. Появление общедоступных ресурсов, интегрирующих современную терминологию в общественно-политической и IT-областях, способствует сохранению и развитию языка и культуры, обеспечивает современный инструментарий для науки, образования и культуры.

Ресурсы, созданные в рамках проекта, вносят важный вклад в изучение фундаментальных процессов, связанных с закономерностями образования новой терминологии для национальных языков РФ, и позволяют определить направление, характер и степень влияния русского языка на процесс формирования актуального лексикона этих языков.

Исследование выполнено при поддержке Российского научного фонда (проект № 16-18-02074 «Разработка моделей связывания терминологии в разных языках (на материале русского и татарского языков)»).

Сведения об авторах: Гатиатуллин Айрат Рафизович, кандидат технических наук, Институт прикладной семиотики АН РТ, e-mail: agat1972@mail.ru; Кириллович Александр Витальевич, Казанский (Приволжский) федеральный университет, e-mail: alik.kirillovich@gmail.com; Хакимов Булат Эрнстович, кандидат филологических наук, Казанский федеральный университет, Институт прикладной семиотики АН РТ, e-mail: khakeem@yandex.ru.

Аннотация: В статье представлен обзор результатов, полученных в ходе реализации проекта Российского научного фонда «Разработка моделей связывания терминологии в разных языках (на материале русского и татарского языков)» в НИИ «Прикладная семиотика» Академии наук Республики Татарстан. Разработаны русско-татарский тезаурус общественно-политической и IT-лексики, размеченные тематические подкорпусы и другие лингвистические ресурсы. Данные ресурсы имеют большой потенциал для исследований современных процессов развития татарской терминологии.

Ключевые слова: тезаурус, терминология, татарский язык, общественно-политическая лексика, IT-терминология.

Abstract: The paper presents an overview of the results obtained during the project of the Russian Science Foundation named «The development of the terminology linking models in different languages (for Russian and Tatar languages)» at the Institute of Applied Semiotics of Tatarstan Academy of Sciences. The Russian-Tatar thesaurus of socio-political and IT-vocabulary, the annotated thematic subcorpora and other linguistic resources were developed. These resources have great potential for studies of modern processes in the sphere of Tatar terminology.

Key-words: thesaurus, terminology, the Tatar language, social and political vocabulary, IT-terminology.