Н.К. Рубашко, Г.П. Невмержицкая (Минск, БГУ)

К ВОПРОСУ РАЗРАБОТКИ СЛОВАРЕЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ «МАШИННЫЙ ФОНД БЕЛОРУССКОГО ЯЗЫКА»

В научно-исследовательской лаборатории интеллектуальных информационных систем БГУ в рамках Государственной программы «Электронная Беларусь» была создана информационная система «Машинный фонд белорусского языка» (ИС МФБЯ). Данная система представляет собой полнофункциональный информационно-программный комплекс, создающий основу для решения большинства практических задач обработки текстов: корректировки орфографии, машинного перевода, реферирования, семантического поиска, автоматизации обучения языку и других.

Главной составляющей ИС МФБЯ является лингвистическая база знаний (ЛБЗ), включающая базу данных (классификатор свойств белорусского языка, корпусы текстов, словари) и базу данных распознающих лингвистических моделей белорусского языка, формализующих языковую компетенцию в целях автоматического анализа текста на всех уровнях его глубины.

Одним из основных компонентов любой ЛБЗ является машинный словарь (МС) естественного языка.

Существуют различные классификации типов МС, из которых можно выделить классификацию по характеру лексических единиц, включенных в словарь, и по способу организации словника. По характеру лексических единиц МС чаще всего подразделяются на словари основ и словари словоформ. По способу организации словников машинные словари подразделяются на алфавитные, обратные, частотные, тезаурусы, конкордансы и др.

Среди словарей особо выделяются двуязычные машинные словари. В таком словаре отдельной словоформе одного языка ставится в соответствие некоторое конечное множество альтернативных вариантов другого языка.

Технология разработки машинных словарей белорусского языка ИС МФБЯ включала следующие этапы:

– сбор статистического материала из различных источников белорусскоязычных документов и составление картотеки всех словоупотреблений, которые встречаются в этих источниках;

– создание на основе полученной картотеки словаря белорусского языка;

– развертывание для каждой лексической единицы всех ее грамматических форм;

– перевод каждой лексической единицы с белорусского языка на русский (для двуязычных словарей);

– перенос словаря на машинные носители информации и последующая его корректировка;

– кодирование машинных словарей с учетом разработанного классификатора лексико-грамматических свойств белорусского языка;

– автоматическое пополнение словарей за счет новых электронных текстов с развертыванием для каждой лексической единицы всех ее грамматических форм и кодирование этих форм.

Первоначальным источником для составления картотеки служили печатные тексты белорусского языка, представленные сплошными выборками по 5–10 тысяч словоупотреблений из СМИ и литературы по литературоведению, языкознанию, биологии, химии, медицине, технике, физике, праву, истории и др.

В дальнейшем словари пополнялись за счет различных словарных источников, учебных пособий и текстов из сети Интернет.

В состав ИС МФБЯ входят следующие словари белорусского языка:

– базовый аннотированный словарь (содержит 157 487 парадигм, что составляет 2 293 117 словоформ);

– словарь имен собственных (15 926 парадигм);

– словарь сокращений и аббревиатур (374 слов);

– словарь синонимов (5 927 синонимичных рядов);

– словарь омонимов (50 562 группы);

– словарь антонимов (188 групп);

– частотный словарь (16 480 слов);

– обратный словарь (142 028 слов);

– словарь ударений (112 227 парадигм);

и двуязычные словари:

– белорусско-русский словарь (237 435 парадигм);

– русско-белорусский словарь (237 702 парадигмы);

– словарь идиом (5075 белорусских идиом);

– терминологические словари (общее число терминов — 32 795).

Разработанный базовый словарь белорусского языка содержит слова, принадлежащие всем существующим в языке частям речи, включая причастия, а также вводные слова и предикативы, и является словарем словоформ, сгруппированных в парадигмы. Под парадигмой понимается совокупность всех грамматических форм некоторого слова, представленных вместе с соответствующими им лексико-грамматическими кодами (ЛГК). Каноническая форма слова представлена первой в парадигме.

При загрузке каждой парадигмы в память компьютера в прикладных программах парадигматический класс представляется в виде его неизменяемой части (квазиосновы) и квазифлексии, в состав которой включаются флексия и часть основы с чередованием. Квазифлексии с соответствующими им грамматическими значениями представляют уровень парадигм для языка флективного типа.

Разработанный словарь имен собственных содержит слова, которые повседневно употребительны, но, будучи именами собственными, традиционно в словари общей лексики не включаются. В состав словаря имен собственных входят словники: личных имен, фамилий и отчеств; наименований физико-географических объектов и территориальных единиц Беларуси; наименований мировых физико-географических объектов и

территориальных единиц; названия государственных и общественных организаций, религиозных праздников, литературных памятников, языков программирования.

Были разработаны словообразовательные цепочки «существительное–прилагательное», содержащие географические наименования и образованные от них прилагательные (около 1 000 цепочек). Эти цепочки являются комплексным справочником в границах тематически связанной группы слов, поскольку образование прилагательных от географических наименований — один из самых сложных вопросов словообразования. Разработанные цепочки могут использоваться при автоматическом анализе текста и его трансформации.

Словарь сокращений и аббревиатур представляет собой попытку собрать воедино и систематизировать наиболее употребительные аббревиатуры и сокращения современного белорусского языка. Особенностью данного словаря, в отличие от других словарей сокращений, является то, что в него включены только неизменяемые сокращения (например, *см, га, т.п.*) и несклоняемые буквенные аббревиатуры, для которых указана категория рода, определяемая по стержневому, центральному слову расшифровки. Сложносокращенные слова (например, *калгас*) и аббревиатуры (например, *МАЗ — Мінскі аўтамабільны завод*), изменяющиеся по падежам, включены в базовый словарь белорусского языка.

Разработанный словарь синонимов белорусского языка содержит не только синонимы в их классическом понимании, но и варианты слова (например, *дзіця — дзіцё*), необходимые для их полного отождествления при информационном поиске и синтезе текста. Данный словарь представляет собой список канонических форм слов, разделенных одним или несколькими пробелами, с приписанным каждому слову ЛГК и с возможностью получения полной парадигмы базового словаря для каждой канонической формы. Словарь синонимов был получен автоматически с использованием двуязычных словарей: белорусско-русского и русско-белорусского с последующей корректировкой и пополнением по словарям синонимов белорусского языка.

Словарь омонимов белорусского языка ИС МФБЯ представляет собой словарь омоформ (грамматических омонимов), поскольку лексические, или простые, омонимы содержатся в базовом словаре белорусского языка. В словаре омонимов не учитывается также внутрипарадигматическая омонимия, если омонимичные формы содержатся в одной парадигме (например, именительного и винительного падежей). Словарь грамматических омонимов был получен автоматически с использованием базового словаря белорусского языка.

Словарь антонимов белорусского языка ИС МФБЯ явился первой попыткой создания электронного словаря такого типа. Данный словарь представляет собой не просто список противоположных по значению слов, он содержит синонимические ряды, которые между собой являются антонимичными. Словарь антонимов белорусского языка был получен путем

перевода словаря антонимов русского языка и пополнен с помощью синонимичных рядов словаря синонимов белорусского языка ИС МФБЯ.

Источником частотного словаря белорусского языка явился аннотированный корпус текстов, созданный в рамках ИС МФБЯ, поэтому вся статистика дается на основе данного корпуса. В разработанном частотном словаре за элемент словника принимается лексема. Все лексемы представлены в исходных (канонических) формах с ЛГК и указанием абсолютной частоты встречаемости по всем текстам аннотированного корпуса. Слова с нулевой частотой встречаемости в итоговый словарь не включались.

В обратном словаре белорусского языка дается не просто список слов (точнее, их исходных словоформ), упорядоченных по концам, а каждое слово списка снабжено статистическими данными по количеству слов, оканчивающихся на любое четырехбуквенное, трехбуквенное, двухбуквенное сочетание и на одну букву. Обратный словарь автоматически строился для всего списка исходных (словарных) словоформ базового словаря белорусского языка с возможностью получения полной парадигмы для каждого слова.

Словарь ударений белорусского языка разрабатывался на основе базового словаря белорусского языка с указанием образования грамматических форм и особенностей расстановки ударения. Расстановка ударений осуществлялась автоматически по формальным признакам ударения с последующей корректировкой вручную.

Поскольку белорусский и русский языки являются языками с богатой формой флексий, то для двуязычных словарей также использовалась парадигматическая структура, описанная выше. Следует отметить, что в разработанном русско-белорусском словаре даются причастия, которые, в зависимости от типа (действительное или страдательное), могут переводиться на белорусский язык как одним словом, так и словосочетанием, имеющим в своем составе местоимение и глагол в соответствующем времени.

Словарь идиом представляет собой список двуязычных выражений. В словарь идиом включены наиболее частотные идиомы. Словарь содержит 1360 оригинальных русских идиом, каждой из которых соответствует несколько синонимичных белорусских, таким образом, количество оригинальных белорусских идиом составляет 5075.

Разработанные для ИС МФБЯ терминологические словари представляют восемь областей знания: биологию, военную науку, математику, физику, кибернетику, юриспруденцию, литературоведение, лингвистику. Особенностью разработанных тематических словарей является их двуязычность: словари представлены на белорусском и русском языках. Данные словари могут служить основой для создания тезаурусов по предметным областям.

Для доступа к разработанным словарям был разработан комплекс программных средств. Словари доступны пользователю как справочное средство (поиск слов, предоставление информации относительно словоизменения конкретных единиц словаря). Также была разработана справочно-информационная

система МФБЯ, дающая возможность пользователю просмотреть как грамматические характеристики заданного слова, так и его словоизменительную парадигму, список синонимов, антонимов, правильную расстановку ударений, частотные характеристики, собранные по корпусу текстов, а также варианты перевода на русский язык. При подключении корпуса текстов пользователю предоставляется возможность просмотреть указанное слово в определенном контексте.