Т. А Грязнухина., Н. М. Заика, Т. П. Любченко, В. А. Широков

ИНТЕГРИРОВАННАЯ ЛЕКСИКОГРАФИЧЕСКАЯ СИСТЕМА КАК ИНСТРУМЕНТ АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ НОВЫХ ЭЛЕКТРОННЫХ СЛОВАРЕЙ И КАК БАЗА ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ В ОБЛАСТИ ЛЕКСИЧЕСКОЙ СЕМАНТИКИ

Одной из основных задач интегрированной лексикографической системы (ИЛС) украинского языка, разрабатываемой в Украинском информационно-языковом фонде НАН Украины, является создание надежных информационно-языковых технологий, обеспечивающих исследования в области автоматической переработки языковой информации.

Предполагаемое включение в систему электронных версий существующих словарей современного украинского языка[1], программное интегрирование их между собой, а также создание программных средств экстракции из этих словарей содержащейся в них семантической информации дает право рассматривать ИЛС в качестве надежного базиса для построения лингвистических процессоров автоматического распознавания смысла обрабатываемой информации, а задачу формирования интегрированных лексикографических систем включить в круг задач, связанных с проблемой создания искусственного интеллекта.

Кроме того, программно-технологический комплекс, которым снабжена ИЛС, обеспечивает выполнение ею функции автоматизированного рабочего места (АРМ) для создания новых автоматизированных словарей, в том числе автоматических.

В основе ИЛС лежит теория лексикографических систем В. А. Широкова, в частности ее стержневое положение о наличии определенного лексикографического эффекта в любой информационной системе [1].

Техническим обеспечением ИЛС являются разработанные в Фонде программно-лингвистические технологии автоматизации интеграционных процессов [2].

В качестве базовой лексикографической системы в ИЛС выбрана электронная версия 20-томного Толкового словаря украинского языка (ЭТС), поскольку этот словарь является наиболее репрезентативным источником информации о семантических свойствах лексических единиц украинского языка и представляет собой обобщение опыта работы лексикографов многих поколений в области исследований лексической семантики.

Форма представления данных в ЛБД словаря, строгая формализация описания его макро- и микроструктуры, обеспечивающие возможность обработки данных компьютерными программами, обусловили многофункциональность электронного толкового словаря. С одной стороны, ЭТС является информационно-поисковой системой, принимающей участие в коммуника-

[1] В настоящее время в систему входят электронные версии: Украинского толкового словаря в 20 томах, грамматического, орфоэпического, этимологического, нескольких переводных и терминологических словарей.

ционном процессе человек – компьютер. В этом случае словарь может быть записан на диске, и тогда обращение к нему происходит по требованию пользователя, или он является рабочим компьютерным словарем, и обращение к нему осуществляется через интернет. С другой стороны, ЭТС выполняет функцию резидентного словаря, представляющего собой часть автоматизированного рабочего места (АРМ) лингвиста в процессе формирования ЛБД новых электронных словарей, описывающих различные лексико-грамматические, семантико-синтаксические и семантические контексты лексических единиц украинского языка. То есть речь идет об автоматических электронных словарях – грамматическом (ЭГС) и синтаксическом глагольного управления (ЭГУ), а также автоматизированных электронных семантических словарях синонимов (ЭСС) и паронимов (ЭСП).[1]

Предполагается, что в дальнейшем после интеграции в ИЛС вновь созданных электронных словарей и сам ЭТС будет параметризован семантическими характеристиками, экстрагируемыми из этих словарей. Это в значительной мере повысит информационный потенциал электронного толкового словаря в плане расширения и специализации последовательного представления в нем семантических свойств лексической подсистемы украинского языка.

Параметризация однозначных единиц и конкретных значений многозначных единиц словника ЭТС с помощью грамматических словарей (ЭГС и ЭСУ) обеспечивает системное описание лексико-грамматических характеристик слов украинского языка с точки зрения

- частеречной принадлежности слова;
- зависимости между конкретным значением многозначного слова и полнотой/ неполнотой его словоизменительной парадигмы;
- принадлежности к семантико-грамматическим категориям вида и переходности/ непереходности глаголов (для многозначного глагола его конкретных значений);
- принадлежности к семантико-морфологической категории рода и к категории одушевленности/неодушевленности существительных (конкретных значений многозначного существительного);
- описания моделей глагольного и именного управления.

Параметризация ЭТС с помощью семантических словарей ЭСС и ЭПС вводит в описание конкретных значений многозначного слова семантические контексты (синонимический и паронимический) в качестве их дифференциального признака.

При определении первоочередности ввода указанных электронных словарей в ИЛС учитывалась степень важности параметров, представленных в этих словарях, для разработки лингвистического обеспечения систем АОТ. С этой точки зрения приоритет электронного грамматического словаря, обеспечивающего лексико-грамматический и семантико-грамматический

[1] Описание понятий *электронная версия бумажного словаря*, *автоматизированный словарь* и *автоматический словарь* см. в работе Л. Н. Беляевой [3].

лексикографические эффекты, очевиден. Лексико-грамматический параметр является таковым, без которого не может обойтись ни один из модулей лингвистического процессора, поскольку информация о принадлежности слова к лексико-грамматическому классу является исходной в синтаксическом, лексическом и в семантическом анализаторах. Представленная в параметре «парадигматический класс» информация о типе словоизменения слов внутри одного грамматического класса делает этот параметр определяющим в программе лемматизации (при переходе от текстовой словоформы к канонической, фиксируемой в словаре) и в программе синтеза (при построении конкретной словоформы по заданной исходной).

В ИЛС грамматический словарь представлен в двух ипостасях – как информационно-справочная система, содержащая данные о словоизменительной подсистеме украинского языка, и как автоматический словарь, ориентированный на выполнение резидентной функции в программном обеспечении систем АОТ.

Целесообразность интегрирования в ИЛС синтаксического словаря управления обусловлена тем, что выбор в качестве единицы словника ЭСУ не лексемы, а слова в его конкретном значении делает возможным представление в эксплицитном виде существующих корреляций между конкретными значениями многозначного слова и их синтаксическими контекстами (схемами управления). Осуществление параметризации электронного толкового словаря с помощью ЭСУ превращает ЭТС в репрезентативный массив языкового материала для исследований еще одного аспекта синтаксической семантики.

Автоматические словари управления, прежде всего, ориентированы на использование их в автоматических редакторах для выявления в тексте ситуаций неправильного употребления схем управления («дякувати [кого]»: «дякувати волонтерів» вместо правильного в украинском языке «дякувати [кому]»: «дякувати волонтерам»). Интеграция ЭСУ с электронным толковым словарем позволяет вводить в правила автоматического редактора подсказки о необходимости изменения формы зависимого слова. Подключение в этом случае ЭГС, в свою очередь, позволит автоматически синтезировать по словарю управления соответствующую форму.

В ходе автоматического морфологического анализа текста обращение к ЭСУ оказывается очень эффективным на этапе снятия грамматической омонимии именных форм, находящихся в тексте в постпозиции к глаголам, зафиксированным в ЭСУ. Составление с помощью операции логического умножения валентных свойств глагольной формы, заданных в словаре управления, и морфологических характеристик именной формы, установленных в результате автоматического морфологического анализа, приводит к полному снятию омонимии или уменьшению длины цепочки омонимичного кода:

замовити – {кого, що, кому} U кулі {род., дав., місц. одн., наз., знах. множ., іменник жін. роду} → кулі { знах. мн., іменник жін. роду, наз., знах. мн., іменник чол. роду}.

В системах МП с украинским компонентом (в частности, украинско-русского перевода) установление в тексте для переводимого глагола реализуемой им схемы управления может сокращать количество переводных эквивалентов (ПЭ). Например, глагол «вистрибнути» имеет два русских ПЭ – «выпрыгнуть» и «запрыгнуть». Если в тексте алгоритмически устанавливается реализация схемы управления «з/ зі/ із + род.пад.», то будет выбран первый ПЭ – «выпрыгнуть», если – схема управления «на + вин. пад.» – второй ПЭ «запрыгнуть».

Программы интеграции электронных словарей в ИЛС и организация АРМ лингвиста, работающего над составлением словаря, обеспечивают автоматическое заполнение полей ЛБД ЭСУ информацией, необходимой для:

- формирования словника словаря, основной единицей которого является слово в его конкретном значении;
- индексации единиц словника номерами значений/ оттенков значения, совпадающими с соответствующими им номерами в ЭТС;
- заполнения полей «переходность» и «вид»;
- заполнения полей «схема предложного управления», «схема беспредложного управления», экстрагированными из поля синтаксических признаков в ЭТС, представленными там относительными местоимениями «кого», «що», «кому» и т.п. или предложными сочетаниями типа «з ким», «для чого»;
- заполнения поля «сфера употребления» в виде стилистических помет, экстрагированных из соответствующего поля ЭТС;
- заполнения рабочего поля «иллюстрации»;
- формирования поля толкования значений из ЭТС. Речь идет не только о простом перенесении соответствующей информации из поля ЭТС в поле ЭСУ, но и о восстановлении в эксплицитном виде толкований, заданных в ЭТС схемами «доконаний до» (ПРОДИБАТИ – док. до **дибати**), «однократний до» (АГАКНУТИ – однокр. до **ага́кати**), «пасивний до» (АБОНУВАТИСЯ – пас. до **абонува́ти**), «безособовий» или формулами отсылок к толкованиям других слов: «те саме, що» (ПЕРЕГАТА, и, *ж.* Те саме, що **зага́та** 1. ЗАГАТА, и, *ж.* 1. Споруда для затримання руху води в річці, потоці і т. ін.), «дивитися» (ПЕРЕГАТИТИ див. **перега́чувати**).

Поле лексико-грамматических признаков заполняется путем параметризации словника с помощью электронного грамматического словаря.

Сервисные программы АРМ обеспечивают возможность составителям словаря в автоматизированном режиме осуществлять обращение к конкретным полям, редактирование информации (удаление, внесение новой, исправление).

В будущем в рамках АРМ предусматривается создание в автоматизированном режиме текстовой корпусной поддержки путем снабжения единиц словника, подчиненных одной лексеме, списком коллокаций, извлеченных из корпуса [4].

Для перехода от автоматизированного словаря управления к автоматическому, выполняющему резидентную функцию внутри программного обеспечения автоматических систем обработки языковой информации

разработаны специальные программы перекодирования содержащейся в них информации, предусмотрено внесение определенных модификаций во внутреннюю структуру словаря.

Интегрирование электронного словаря синонимов и электронного словаря паронимов[1] в ИЛС обусловлено тем, что синонимия и паронимия принадлежат к широко распространенным в языке явлениям, которые из-за своей причастности к плану содержания языковых знаков требуют особого внимания к себе в системах автоматической обработки текстовой информации, в частности это касается систем машинного перевода, автоматического редактирования, систем информационного поиска. Если синонимы играют положительную роль в системах МП (они могут быть использованы как инструмент для уточнения и сокращения количества вариантов перевода), а также в системах информационного поиска (проиндексированные синонимы обеспечивают повышение показателя полноты поиска), то паронимия негативно влияет на качество работы этих систем. Неправильное употребление компонента паронимической пары в переводимом тексте может приводить к полному искажению его смысла и поэтому крайне негативно влиять на качество перевода. При информационном поиске выбор неправильного паронима в качестве ключевого слова увеличивает показатели шума (находится лишнее) или молчания (ничего не находится) информационно-поисковой системы.

Паронимическая параметризация ИЛС может стать полезной при составлении семантических правил модуля автоматического редактирования исходных текстов в системах машинного перевода, а также в информационно-поисковых системах. Стало очевидным, что степень эффективности этих систем находится в прямой зависимости от их способности распознавать и исправлять в обрабатываемом тексте ситуации с неправильным употреблением одного из компонентов паронимической пары.

При параметризации электронного толкового словаря с помощью ЭСС индексируются конкретные значения слов соответствующими этим значениям синонимическими группами. Внутри группы синонимы проиндексированы номерами значений, взятыми из ЭТС. При параметризации текстовой информации индексируемым словоформам ставятся в соответствие все синонимические группы, сформированные по конкретным значениям лексемы данной словоформы с указанием соответствующего значения. Аналогичным образом происходит и параметризация ЭТС и по электронному словарю паронимов.

В рамках корпусной лингвистики электронный грамматический словарь и автоматический словарь управления выполняют функцию инструмента соответственно морфологической и синтаксической разметок Украинского национального лингвистического корпуса. Электронные словари синонимов и паронимов предлагается использовать в качестве основного инструмента семантической разметки.

---

[1] Принятое определение понятий синонимов и паронимов, внешние и внутренние структуры словарей ЭСС и ЭСП, принципы формирования описаны в статьях [5–9].

ЛИТЕРАТУРА

1. *Широков, В. А.* Інформаційна теорія лексикографічних систем / В. А. Широков. – Київ : Довіра, 1998. – 331 с.
2. *Широков, В. А.* Лінгвістичні та технологічні основи тлумачної лексикографії / В. А. Широков, В. М. Білоноженко, О. В. Бугаков [та ін.]. – Київ : Довіра, 2010. – 293 с.
3. *Беляева, Л. Н.* Формирование переводного словаря средствами информационных технологий / Л. Н. Беляева // Прикладна лінгвістика та лінгвістичні технології : MegaLing-2012 : зб. наук. пр.; НАН України, Укр. мовно-інформ. фонд; редкол. : В. А. Широков [та ін.]. – Київ : УМІФ, 2013. – С. 22–36.
4. *Широков, В. А.* Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна [та ін.]; під ред. В. А. Широкова. – Київ : Довіра, 2005. – 471 с.
5. *Грязнухина, Т. А.* Индексирование паронимов в лексикографических системах / Т. А. Грязнухина, Т. П. Любченко, В. А. Широков // Прикладна лінгвістика та лінгвістичні технології : MegaLing'2009 : зб. наук. пр.; за ред. В. А. Широкова. – Київ : Довіра, 2009. – С. 171–179.
6. *Грязнухина, Т. А.* Технология и программно-инструментальные средства формирования електронного словаря паронимов украинского языка / Т. А. Грязнухина, Т. П. Любченко, В. А. Широков, К. Н. Якименко // Прикладна лінгвістика та лінгвістичні технології : MegaLing'2010 : зб. наук. пр.; за ред. В. А. Широкова. – Київ : Довіра, 2010. – С. 45–54.
7. *Грязнухіна, Т. О.* Словник української мови у 20-ти томах як інструмент для створення Електронного словника паронімів / Т. О. Грязнухіна, Т. П. Любченко // Лексикографічний бюлетень : зб. наук. пр.; відпов. ред. к. філол. н. І. С. Гнатюк. – Київ : Видавничий дім Дмитра Бураго, 2011. – Вип. 20. – С. 28–34.
8. *Устимець, О. В.* Формування електронного словника синонімів української мови / О. В. Устимець // Учён. зап. Таврического нац. ун-та им. В. И. Вернадского. Серия «Филология» : в 59 т. – Симферополь, 2007. – № 4. – Т. 20. – С. 57 – 61.
9. *Грязнухіна, Т.* Операційне визначення критерію семантичної подібності синонімів / Т. Грязнухіна, О. Устимець, В. Широков // Прикладна лінгвістика та лінгвістичні технології : MegaLing-2008 : зб. наук. пр. / НАН України, Укр. мовно-інформ. фонд; редкол. : Ю. Д. Апресян [та ін.]. – Київ : Довіра, 2009. – С. 42–57.