Электронные грамматические словари в интегрированной лексикографической системе

Electronic grammar dictionaries in the integrated lexicographical system

Любченко Т.П.

Украинский языково-информационный фонд НАНУ

Рассматриваются вопросы разработки структуры электронных грамматических словарей (для русского, немецкого, испанского, английского языков) и их функциональные возможности, а также некоторые аспекты построения пользовательского интерфейса. Словари предназначаются для использования в информационно-справочной системе, а также для применения в контурах систем автоматической обработки текстовой информации (в алгоритмах морфологического анализа и синтеза текста). При разработке структуры словаря учтены требования, предъявляемые к программным продуктам подобного типа.

Электронные грамматические словари (ЭГС) разрабатываются как часть интегрированной лексикографической базы Украинского языково-информационного фонда Национальной Академии наук Украины (УЯИФ НАНУ).

В основу разработки положена базовая концепция информационной теории лексикографических систем, разработанная д.т.н. Широковым В.А. Основные положения информационной теории лексикографических систем изложены в работах [1, 2, 3]. Примером успешного применения информационной теории лексикографических систем является интегрированная лексикографическая система (ИЛС), объединяющая функции словоизменения, орфоэпии, фразеологии, синонимии и антонимии украинского языка. Как готовый лексикографический продукт ИЛС украинского языка реализована в виде лазерного диска [4].

Нами выполняется работа по созданию многоязычной ИЛС. Одной из компонент этой системы являются грамматические словари. ЭГС разрабатываются прежде всего для языков, включенных в систему МП: украинского[1], русского, английского, немецкого, испанского, а в перспективе французского, турецкого языков.

Словари ориентированы на письменные варианты языков.

Электронные грамматические словари предназначены главным образом для использования их в качестве инструмента автоматического морфологического анализа в системе МП (на этапах морфологической разметки текста, лемматизации и синтеза). Помимо этого, словарь должен быть доступен пользователю как справочное средство (поиск слов, предоставление информации относительно словоизменения конкретных реестовых единиц словаря). Такие особенности назначения словаря выдвигают и определенные требования к его структуре (лингвистическая информация, представленная в словаре должна быть достаточной для выполнения всех требуемых функций; должно быть обеспечено широкое разнообразие способов доступа к этой информации).

При МП текстов на разных его этапах возникает необходимость получать исходную форму слова от текстовой (при анализе входного текста) и синтезировать требуемую словоформу от исходной, которой слово

[1] Разработка ЭГС украинского языка выполнена ст.н.с. УЯИФ НАНУ И.В. Шевченко.

задается в автоматическом переводном словаре (при синтезе выходного текста). Решение упомянутых задач связано с необходимостью построения парадигматической классификации лексики, разработки структуры и создания базы данных, репрезентирующей данную систему словоизменения, разработки алгоритмов формирования как развернутой парадигмы слова, так и конкретной его формы.

В качестве основного инструмента для решения этих задач предлагается использовать электронный грамматический словарь (ЭГС).

Принципы построения парадигматической классификации лексики

В ЭГС аналогом словоизменительного типа является парадигматический класс (ПК).

Под парадигматическим классом мы понимаем группу лексем, парадигма которых характеризуется одинаковым количеством грамматических форм и внутри которой словоизменение осуществляется в соответствии с единым правилом формирования парадигматических форм. Для слов одного парадигматического класса в языках флективного типа это предполагает одинаковость флексий в соответствующих грамматических значениях и совпадение характера чередования в основе. Для языков аналитико-синтетических и синтетико-аналитических к этим требованиям добавляется требование одинаковости моделей образования аналитических форм.

В соответствии с этим определением слово 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. Грамматические значения в структуре данных ЭГС представлены двухсимвольными кодами: первый символ обозначает часть речи, второй – грамматические значения словоизменительной формы.

Технология создания компьютерной базы русского грамматического словаря

Источником лингвистической информации по русскому словоизменению является Грамматический словарь русского языка А.А.Зализняка (в дальнейшем – ГСЗ) [5], достаточно полно моделирующий словоизменительную систему русского языка

Технология создания компьютерной базы русского грамматического словаря включала в себя следующие этапы:

- перевод бумажного ГС в электронную форму (сканирование);
- корректура отсканированного текста;
- разработка структуры лексикографической системы ГС, языка ее разметки и установление идентификаторов элементов структуры;
- автоматическая конверсия электронного текста ГС в лексикографическую базу данных (ЛБД) в соответствии с разработанной структурой;
- разработка алгоритмов парадигматической классификации ГС и их программная реализация;

- формирование парадигматической ЛБД (автоматическое индексирование лексем кодами грамматических классов и номерами парадигматических классов).

На основе данной технологии был создан ЭГС русского языка [6], который является частью интегрированной лексикографической базы и может быть использован в качестве информационно-справочной системы пользователями-филологами, работающими с русскими текстами относительно словоизмененения современного русского языка в его письменной форме.

В качестве источников лингвистической информации для составления грамматических словарей немецкого, английского и испанского языков использованы грамматики и словари этих языков [7-8].

Принципы представления лексического материала в ЛБД (Структура данных)

На внутреннем уровне архитектуры лексикографической системы ЭГС структура лингвистических данных русского языка представляется реляционной моделью, отношения которой представлены следующими таблицами:

- таблицей **nom** реестровых единиц *Reestr* с указанием кода лексико-грамматического класса *part* и номера парадигматического класса (поле *type*);
- таблицей квазифлексий **flex**, в которой для каждой грамматической формы (поле *NumbOfGrForm*) каждого парадигматического класса (поле *type*) заданы квазифлексии *flex*;
- таблицей **indent**, задающая параметры и характеристики, являющиеся одинаковыми для каждого из парадигматических классов;
- таблицей **Parts** лексико-грамматических классов и их кодов;
- таблицей **gr** словоизменительных типов.

Связь между таблицами **nom**, **flex**, **indent** осуществляется по номеру парадигматического класса (поле *type*); а между таблицами **nom**, **Parts**, **gr** – по полю *part*, что отражено на рис. 1.

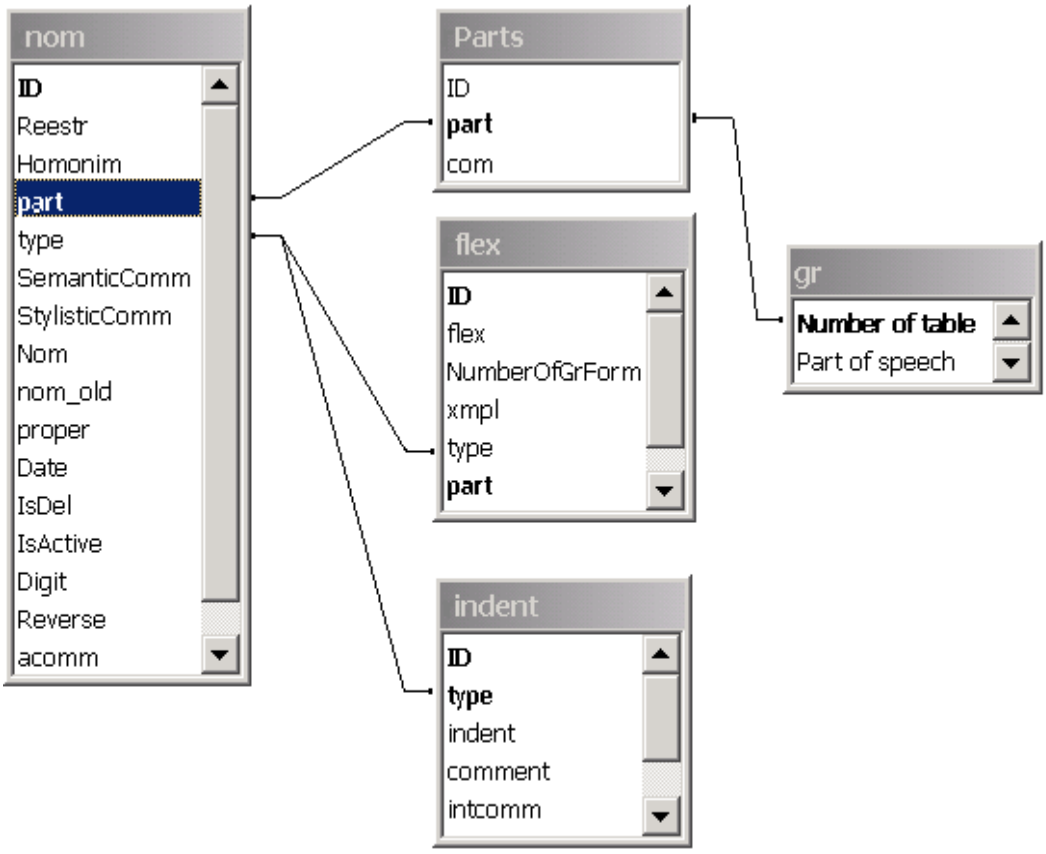

Рис. 1. Схема связей между таблицами данных.

При разработке структуры данных для других языков (английского, испанского, немецкого) с целью унификации их представления в ЛБД, был использован подход, аналогичный вышеописанному. Т.е. основной единицей словаря выбрана квазиоснова. Отличие заключается лишь в том, что в ЛБД этих языков в описание парадигмы помимо флексий вводится еще и тип процедуры, используемый при построении конкретных форм (аналитических).

В английском, испанском языках квазиоснова правильных глаголов, большинства существительных и прилагательных совпадает с исходной формой слова. Представление словарной единицы в виде квазиосновы касается по существу только неправильных глаголов и незначительного количества существительных. Немецкий язык, который относят к языкам синтетико-аналитического типа, характеризуется тем, что простые (синтетические) формы образуются в нем флективным способом, а сложные (аналитические) – по определенным схемам (процедурно). Характерно также большое количество чередований в основе слова и явление отделяемости префиксов у определенной группы глаголов. Все эти особенности словоизменительных процессов немецкого языка потребовали, во-первых, учета их при разбиении множества словоизменительных единиц языка на парадигматические классы, и, во-вторых, введения в структуру данных некоторых дополнений. Помимо таблиц, представленных на рис. 1., в структуру данных включается таблица, задающая типы отделяемых префиксов, а также таблица, задающая типы процедур построения аналитических форм. В таблицу **indent** вводятся дополнительные поля, задающие тип чередований, а также признак отделяемости префикса.

Программные средства для подготовки и редактированния грамматических ЛБД

Грамматические ЛБД функционируют под СУБД Microsoft SQL Server 7.0. Клиентская программа, служащая инструментом для редактирования и пополнения грамматических ЛБД, разработана и создана в среде Microsoft Visual Studio 6.0. Программа работает под управлением операционной системы Microsoft Windows 2000 или Microsoft Windows XP.

Программа реализует следующие функции:

- просмотр реестра;
- получение полной словоизменительной парадигмы выбранного из реестра слова и его основных грамматических характеристик;
- вывод и просмотр части реестра (по заданной части речи, по номеру парадигматического класса, по произвольному запросу (на языке SQL); выдача всех грамматических омонимов, имен собственных и т.п.);
- выдача количественных характеристик относительно наполняемости парадигматических классов, частей речи, омонимов и т.п.
- поиск слов в реестре;
- построение прямого или инверсного словаря (установка прямой или обратной сортировки в реестре);
- ввод новых и редактирование уже введенных реестровых слов, удаление слов из реестра;
- ввод, редактирование, удаление парадигматических классов (задание их дифференцирующих характеристик; ввод и редактирование квазифлексий – для флективных языков, типов процедур образования аналитических форм для языков аналитических);
- запись в файл/ вывод на печать выделенных фрагментов (например вывод полной парадигмы конкретного слова; запись в файл части реестра и т.п.);
- построение словаря квазиоснов (для языков флективного типа; словарь квазиоснов используется программами МА и СА).

Структура ЛБД, описанная выше, удобна для ее формирования и редактирования. Для программ, требующих быстрой обработки парадигмы слов (реализация функций лемматизации и синтеза, например, в программе морфологического анализа), более оптимальным с точки зрения быстродействия и использования дискового пространства вариантом хранения информации является вариант словаря в виде квазиоснов и квазифлексий. Поэтому была разработана оптимизированная ЛБД, которую реализует файл типа gram.dic а также библиотека функций для работы с файлом такой структуры. Библиотека функций может динамически подсоединяться к внешней программе (например, программе МА и СА).

Рабочее окно программы изображено на рис. 2.

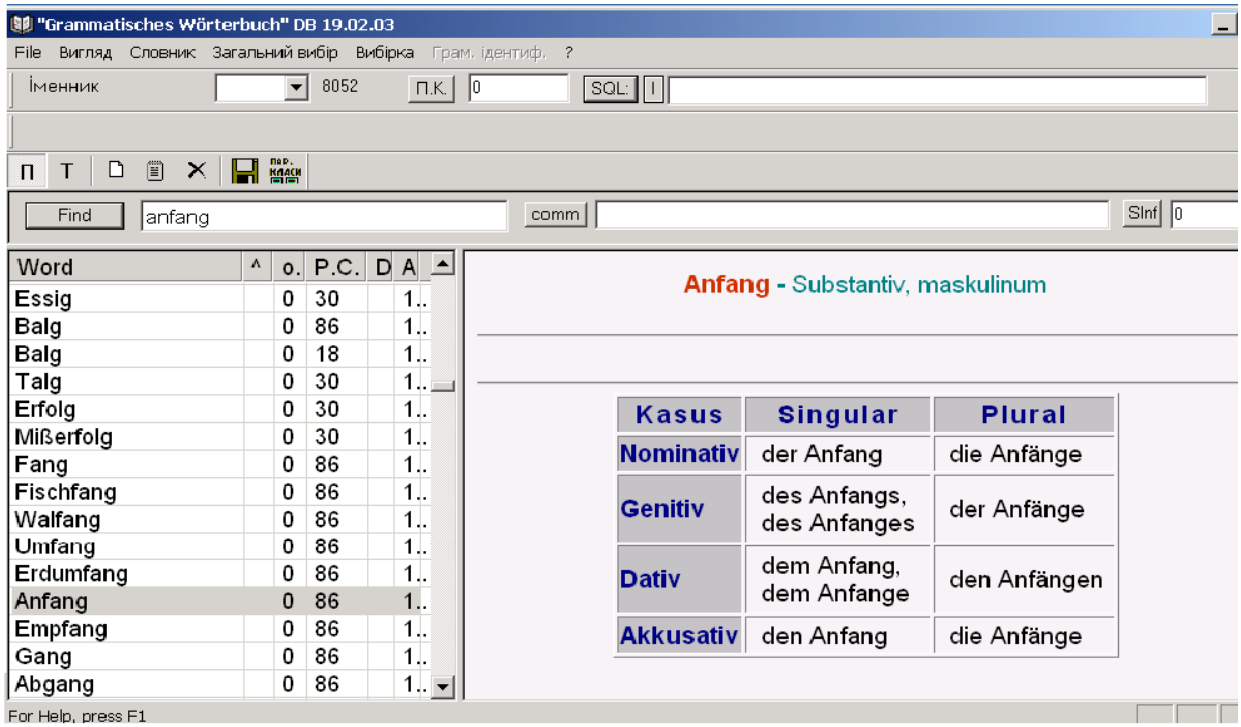

Рис. 2. Рабочее окно программы редактирования немецкого грамматического словаря.

Для русского (и украинского) языка разработана и создана программа морфологического анализа текста. Программа выполняет морфологическую разметку текста двухсимвольными кодами («часть речи»-«грамматическое значение») и осуществляет лемматизацию и синтез словофом. В работе программы используются словари квазиоснов и квазифлексий, записанные в специальном оптимизированном формате. Для других языков программы морфологического (морфолого-синтаксического) анализа еще предстоит разработать.

Отметим, что созданные грамматические словари для русского и украинского языков а также программный инструмент прошли аппробацию на значительных лексических массивах (русский язык – ок.170 тыс.лексексических единиц; украинский – ок. 200 тыс. лексических единиц).

Литература

1) Широков В.А. Інформаційна теорія лексикографічних систем. Київ, Довіра, 1998. – 331с.

2) Широков В. А. Інформаційно-лінгвістичні основи сучасної тлумачної лексикографії // Мовознавство.— 2002.— № 6.— С. 7-48.

3) Широков В. А., Рабулець О. Г., Костишин О. М., Шевченко І. В., Якименко К. М. Технологічні основи сучасної тлумачної лексикографії // Там же.— С. 49–86.

4) Інтегрована лексикографічна система «Словники України» // Широков В.А., Шевченко І.В., Рабулець О.Г., Пещак М.М., Костишин О.М. – Київ, 2001 (електронне видання).

5) Зализняк А.А. Грамматический словарь русского языка: Словоизменение. – М.: Русский язык, 1978. – 878 с. (4-е изд., испр. и доп.: 2003).

6) Грязнухина Т.А., Любченко Т.П., Рабулец А.Г. Электронная версия грамматического словаря русского языка (А.А.Зализняк) как инструмент автоматического морфологического анализа русского текста. //

Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных», Санкт-Петербург, март 2002. - С. 63-70.

7) G. Wahrig. Deutsches Wörterbuch. Wissen Media Verlag GmbH, Gütersloch/ München 2002 (vormals Bertelsmann Lexikon Verlag GmbH). – 1451s.

8) А.В. Садиков, Б.П. Нарумов. Испанско-русский словарь современного употребления. М.: Русский язык, 2001. – 752 с.

Electronic grammar dictionaries in the integrated lexicographical system

Lyubchenko T.P.

Problems of development of the grammar Russian, German, Spanish and English dictionaries structure, their functionalities and some aspects of constructing the user interface are examined. Dictionaries are intended for using in an information and reference system, as well as for application in the language processing systems (morphological analysis and text synthesis). When developing the dictionary structure the requirements to software of similar type were taken into account.