Иванка Атанасова, Преслав Наков, Светлин Наков (Болгария)

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ПОМОЩЬ ИССЛЕДОВАТЕЛЮ-ЛИНГВИСТУ

1. Стандартные и специализированные компьютерные технологии

Информационные технологии предлагают два основных типа средств в помощь исследователю – стандартные и специализированные. Стандартные средства – это общедоступные программные продукты и технологии, а специализированные средства – специально разработанные для конкретного исследования программные продукты.

Текстообрабатывающий софтвер очень широко использованное стандартное средство для создания и редактирования текстовых материалов и документов. Большинство современных текстообрабатывающих программ (как, например, Microsoft Word) позволяет форматировать и оформлять тексты, включать графики, таблицы, формулы, иллюстрации, примечания, аннотации и другие объекты в документы, распечатывать их на принтере. Реализованы также страницирование, поиски, замещение, устранение, копирование отрывков текста, корректирование орфографических и грамматических ошибок, перенос слов и еще много других функций.

Электронные таблицы (как, например, Microsoft Excel) тоже широко использованные стандартные средства для выполнения различных подсчетов и исчислений. Наряду с простыми подсчетами, как суммирование, определение среднего арифметического числа, исчисление процентов и другие, могут выполняться и более сложные совокупности вычислительных действий. Так, например, по предварительно заданной вычислительной схеме можно подсчитывать результаты многих научных экспериментов, а потом эти результаты визуализировать в виде таблиц, график и диаграмм.

Системы управления баз данных (как, например, Microsoft Access) являются стандартным средством хранения и обработки информации. В них данные сохраняются в структурированном виде, чаще всего в таблицах, между которыми могут быть определенные связи. Разрешается хранение, добавление и устранение данных, а также самые разнообразные поиски по различным (нередко довольно сложным) критериям. К сожалению, работа с помощью этого софтвера сложная, требующая специальных познаний и умений, что делает его неудобным для массового потребления.

Компьютерные словари (как, например Babylon Translator) являются стандартным средством перевода слов и выражений с одного языка на другой. Разумеется, есть и более новые технологии – автоматические электронные переводчики, выполняющие автоматический машинный межъязыковой перевод текстов. Известны два вида таких переводчиков – персональные (как, например, Socrates и Magic Goody), представляющие собой софтуер для межъязыковых переводов, установленные локально, и on-line – предлагаемые в качестве услуг в Интернете (как, например, Alta Vista Translation Services). Для всех автоматических переводчиков характерно, что полученный машин-

ный перевод не очень удачный, и поэтому специалисту по соответствующим языкам необходимо дополнительно корригировать переведенные тексты. Несмотря на этот недостаток, они экономят труд и время.

Существует большое разнообразие общедоступных средств, помогающих исследователю-лингвисту в его работе, которых, однако, вовсе недостаточно. Нередко исследовательская работа требует узко специфической и нестандартной с точки зрения общедоступного софтвера обработки данных. В таком случае даже профессиональные умения для работы со стандартными средствами мало помогают. Возникает необходимость в разработке специализированного софтвера для выполнения соответствующей специфической обработки данных, который создается высококвалифицированным компьютерным программистом или коллективом программистов. Разумеется, далеко не каждое исследование может быть выполнено при помощи компьютера, однако в определенных случаях можно добиться исключительно хороших результатов в короткий срок. Это зависит не только от цели исследования, но также от использованных исследовательских способов и методов формализации проблем, от которых зависят компьютерные алгоритмы, при помощи которых они решаются. От точности и эффективности этих алгоритмов зависят конечные результаты. Ключ к удачному компьютерному исследованию – формализировать проблемы, над которыми работают, и потом создать софтверные продукты, которые решают их или помогают в их решении.

Приведем несколько примеров для специфического софтвера, разработанного нами специально для нужд конкретного исследования в области русской и болгарской терминологии изобразительного искусства в плане сопоставления, а именно: компьютерный словарь терминов изобразительного искусства и автоматическое извлечение гипонимических рядов из компьютерных терминологических словарей с двумя дополняющими друг друга техниками – формальной и семантической. В настоящем исследовании представлена только формальная техника.

2. Компьютерный словарь терминов изобразительного искусства

Компьютерный словарь терминов изобразительного искусства (КСТИИ) – это специально разработанный софтверный продукт для построения и поддержки компьютерных словарей. Он представляет собой небольшую информационную систему, основанную на реляционных базах данных и обеспечивающую поддержку двух словарей – болгарского и русского. Каждая словарная статья состоит из слова, т. е. *однословного термина (ОТ) или терминологического словосочетания (ТС)* и краткого толкования соответствующего термина, включающего все его значения.

Составленные нами два КСТИИ для русских и болгарских терминов изобразительного искусства (ИИ) включают соответственно 2633 русских и 2894 болгарских лексических единиц (ОТ и ТС).

Главное окно КСТИИ разделено по горизонтали на две половины, причем верхняя часть предназначена для болгарского компьютерного словаря, а

нижняя – для русского. Таким образом, при сравнительном лингвистическом исследовании можно наблюдать одновременно более короткие эквивалентные словарные статьи в обоих языках. Более длинные словарные статьи можно наблюдать последовательно в соответствующих диалоговых окнах.

В левой половине словарей ОТ и ТС, включая дублеты и варианты, расположены по вертикали в алфавитном порядке, причем в скобках отмечаются происхождение заимствованных слов и форма единственного числа в тех случаях, когда их основная форма представлена во множественном числе. Дублеты и варианты располагаются и по горизонтали рядом с основным ОТ или ТС, которые разделяются запятыми.

В правой половине словарей приводятся толкования терминов, включая все значения полисемантичных терминов.

С помощью диалогового окна для редактирования словарной статьи можно вносить изменения в орфографию, пунктуацию, объем и содержание термина и его толкования (его значений).

При введении новых терминов компьютерная программа автоматически располагает все лексические единицы в алфавитном порядке, не позволяя два раза вводить один и тот же термин. При попытке добавить термин, включенный в словарь, к его толкованию добавляется только новое дополнительное толкование. Ненужные лексические единицы легко устраняются, причем во избежание случайных ошибок обязательно требуется подтверждение.

При помощи стандартных клавиш навигации можно последовательно рассматривать термины в алфавитном порядке. Компьютерная программа дает возможность быстро отыскать нужный термин или установить его отсутствие в соответствующем компьютерном словаре. Чтобы добиться этого результата необходимо написать термин или его начальные буквы в соответствующее окно и активировать кнопку поисков. Программа найдет термин, а в случае его отсутствия в словаре укажет на ближайший в алфавитном порядке.

КСТИИ служат не только для составления соответствующих словарей и хранения введенных данных. Программа предоставляет лингвисту дополнительную ценную информацию для сопоставительных научных исследований терминов, включенных в КСТИИ, возможность вносить изменения в их объем и содержание, легко и быстро переключаться из словаря в словарь, немедленно подавать информацию о точном количестве терминов в обоих словарях.

Софтверный продукт КСТИИ создан в среде для быстрой разработки приложений Borland Delphi. Delphi является визуальной средой для программирования, основанной на принципах объектно-ориентированного программирования и компонентной модели разработки приложений. В целях хранения и обработки информации использована встроенная в Delphi поддержка реляционных баз данных с помощью Borland Database Engine и базы данных PARADOX, а поиски в словарях реализуются языком структурированных запросов SQL.

3. Гипонимия

В лингвистической литературе принято слова, обозначающие родовые понятия, называть гиперонимами, а слова, обозначающие видовые понятия – гипонимами [Новиков, 1982, с. 241]. Гипероним обозначает общее родовое понятие или совокупность, целое по отношению к составляющим его элементам, частям. Гипоним обозначает видовое понятие или название элемента, части какого-нибудь множества, целого.

Гиперонимы и гипонимы образуют гипонимические ряды, в которых гипонимы занимают подчиненное положение по отношению к гиперонимам. В гипонимический ряд входят один гипероним, занимающий ведущее место и обозначающий общее понятие, и минимум два гипонима, занимающие подчиненное положение по отношению к нему. Гипонимы в гипонимическом ряду находятся в равноправных отношениях, т. е. в отношениях соподчиненности и называются “согипонимами” [Новиков, 1982, с. 241]. Со своей стороны, гипонимы тоже могут стать гиперонимами, образуя новые гипонимические ряды. Родо-видовые отношения в гипонимических рядах выражаются семантически или формально-семантически [Сперанская, 1984, с. 10].

4. Автоматическое извлечение гипонимических рядов

Более ранние попытки автоматического извлечения синонимов, гипонимов и гиперонимов (для английского языка) указывают на три основных метода: шаблонный, синтаксический и семантический. Шаблонный метод использован Хеарст [Hearst, 1992, p. 539], которая извлекает эксплицитно заданные в тексте гипонимы с помощью заранее заданных шаблонов, как “such that”, “or other” и др. Дас-Гупта [Das-Gupta, 1987, p. 245] делает попытки обнаружить гиперонимы в лексиконе на основе синтаксического анализа с целью идентификации термина, содержащего основные характеристики заданного гипонима-цели, Шайкевич [Shaikevich, 1985, p. 76] предлагает метод автоматического открытия синонимов на основе их дефиниций в лексиконе, исходя из идеи, что близкие по значению слова имеют сходные дефиниции.

Автоматическое извлечение гипонимических рядов из терминологических словарей – это специфический софтверный продукт, включающий две дополняющие друг друга техники: *формальная и семантическая.*

5. Формальная техника

Формальная техника используется для извлечения гипонимов (ОТ и ТС), содержащих общий терминоэлемент. Терминоэлемент – это широкое понятие, включающее производящую основу, словообразующую морфему (аффиксы) и слово (лексему) как компоненты в составе сложных терминов и ТС. [Даниленко, 1977, с. 37]. Терминоэлементами могут быть и ТС. Формальная техника применяется в тех случаях, когда родо-видовые отношения выражаются формально-семантически, т. е. семантически и одновременно синтаксически или морфологически. Благодаря специфической структуре компью-

терных словарей, посредством этой техники одновременно извлекаются дублеты и варианты соответствующих гипонимов, которые содержат или не содержат общий терминоэлемент. Как известно, с семантической точки зрения варианты считаются дублетами [Калинина, 1987, с. 11].

Рассмотрим как работает формальная техника извлечения гипонимов, основывающаяся на общем терминоэлементе (корне или основе, аффиксе, слове в качестве компонента ТС или сложного слова, ТС), который содержится в гипонимах, но вовсе не обязательно в их гипероними. На программу КСТИИ подают гипероним, выраженный ОТ или ТС, и она автоматически приводит ряды ОТ и ТС, содержащих этот терминоэлемент. Во время поисков дополнительная информация в скобках, касающаяся происхождения заимствованных слов, не учитывается, так как она может внести ненужный термин в гипонимический ряд, если содержит заданный терминоэлемент, однако сам термин не является согипонимом, например: *р. вышивка – *фриз (фр. frise, по срвлат. frisium, phrygium 'вышивка'...), дерево – *цитрин (нем. zitrin, от лат. citrus 'лимонное дерево'); б. копие – *ланцетка (фр. lancette, по лат. lancea 'копие'), масло - *линолеум (анг. linoleum, по лат. linum 'ленено платно' + oleum 'масло').* Лишь в редких случаях дополнительная информация в скобках помогает включить в гипонимический ряд новый согипоним, не содержащий этот терминоэлемент, например: *р. статуя - колосс (нем. Kolob, фр. colosse, от лат. Colossus, по гр. kolossos 'большая статуя'); б. камък – пирит (нем. pyrit, фр. pyrite, от гр. pyrites 'огнен камък').*

Компьютерная программа извлекает список ОТ и ТС, содержащих заданный терминоэлемент (морфему, основу, слово, ТС). Оказывается, что не все ОТ и ТС в списке с общим терминоэлементом являются согипонимами, например: *р. портрет – *портретист, *портретировать, *портретируемый; б. акварел - *акварелист, *акварелистка.* С помощью одной формальной техники нельзя определить гипонимический ряд, так как родо-видовые отношения не могут выражаться только формально, а формально-семантически. Опираясь на свой опыт и знания, после дополнительных справок в компьютерных словарях, если это необходимо, специалист может точно определить состав гипонимического ряда и устранить ненужные термины из компьютерного списка.

Если общий терминоэлемент выражен словом, чья форма множественного числа не содержит какого-нибудь компонента (буквы) формы единственного числа, компьютерной программе подаются последовательно обе формы, например: *р. роспись – художественная роспись…и росписи – стенные росписи, церковные росписи; б. цвят – локален цвят… и цветове – ахроматични цветове, контрастни цветове….*

Формальная техника извлечения гипонимов с общим терминоэлементом, выраженным суффиксом, используется редко, так как она не очень эффективна. При использовании суффиксов, широко распространенных в терминологии ИИ, в списки обычно попадает большое количество терминов, не являющихся согипонимами, но содержащих этот суффикс, например: *б. худо-*

*жествени занаяти – ..., златарство, килимарство, леярство ..., *абстрактно изкуство, *декадентство, *майсторство ...(суфикс –ств-о).*

Недостатком формальной техники извлечения гипонимов с общим терминоэлементом является невозможность программы разграничивать морфемы или лексемы от случайно совпадающих с ними комплексов букв, например: *р. лак – *плакат, мел – *мельхиор, фон – *фонтанная скульптура; б. стил – *мастило, тон – *картон, туш – *картуш.*

6. Выводы

Настоящее исследование приводит нас к следующим выводам:

1. Компьютерные программы, разработанные специально для конкретных лингвистических исследований, очень нужный (порой незаменимый) и исключительно надежный помощник исследователя-лингвиста. С их помощью экономится много времени и сил, быстрее получаются более точные результаты.

2. Предложенные нами софтверные продукты могут быть использованы для сходных лингвистических исследований как конкретных, так и сопоставительных и в других областях познания. КСТИИ может быть использован также при составлении различных толковых и двуязычных словарей. Формальную технику извлечения гипонимических рядов можно удачно применять при конкретных и сопоставительных исследованиях не только в области гипонимии, но и лексикологии вообще, а также морфологии, словообразования и синтаксиса.

Литература

Даниленко В.П. Русская терминология. Опыт лингвистического описания. М., изд. “Наука”, 1977.

Калинина Р.П. Термины кузнечно-штамповочного производства в лексической системе русского языка (функционально-парадигматическое описание). АКД. Днепропетровск, 1987.

Новиков Л.А. Семантика русского языка. М., изд. “Высшая школа”, 1982.

Сперанская Н.Н. Лесохозяйственная терминология. АКД. Л., 1984.

Das-Gupta P. Boolean Interpretation of Conjunctions for Document Retrieval. Journal of the ASIS, 38(4), 1987.

Hearst M. Automatic Acquisition of Hyponyms from Large Text Corpora. Proc. of COLING 92, Nantes, 2, 1992.

Shaikevich A. Automatic Construction of a Thesaurus from Explanatory Dictionaries, Automatic Documenation and Mathematical Linguistics, 19(2), 1985.