УДК 81’374

АВТОМАТИЧЕСКАЯ КОМПИЛЯЦИЯ БАЗЫ ДАННЫХ КОМПЛЕКСНОГО ЭЛЕКТРОННОГО СЛОВАРЯ

Л. Н. Беляева, А. Н. Ефремова

Российский государственный педагогический университет имени А. И. Герцена

Поступила в редакцию 31 января 2015 г.

Аннотация: *в статье рассматривается способ автоматической компиляции базы данных комплексного электронного словаря из различных бумажных словарей, основанный на распознавании структуры исходных словарей и словарных статей и на представлении извлеченной лексикографической информации с помощью фреймовой модели.*
Ключевые слова: *комплексный электронный словарь, автоматическая компиляция словаря, распознавание структуры словаря, фрейм.*

Abstract: *the paper describes the method for automatic compilation of complex electronic dictionary database from various paper dictionaries. The method is based on the recognition of the structure of the source dictionaries and their lexical entries. The extracted information is then represented as a frame model.*
Key words: *complex electronic dictionary, automatic dictionary compilation, dictionary structure recognition, frame.*

Одной из основных задач современной компьютерной лексикографии является создание и ведение электронных словарей и терминологических баз данных, предназначенных для переводчиков, тратящих до 30 % времени на собственно терминологическую работу [1]. Именно для переводчиков необходимы комплексные электронные словари (далее – КЭС), которые включают в свою структуру различные словари, как общие, так и отраслевые, как переводные, так и энциклопедические и толковые, а также специализированные словари переводчика [2].

Для автоматизации отдельных этапов терминологической работы сегодня создано довольно много инструментальных средств, однако нет универсального решения для основных задач извлечения терминологии и ведения комплексного переводческого ресурса [3]. Задача извлечения терминологии решается на основе сопоставительного анализа параллельных или сопоставимых текстов и основана на применении статистических метрик и лингвистических фильтров. Напротив, создание основы КЭС предполагает извлечение информации из различных лексикографических источников, бумажных в том числе, и объединение ее в единый ресурс. Рассмотрение особенностей его создания, выбора макро- и микроструктуры, а также способа структурирования лексикографической информации представляет особый интерес и в теоретическом, и в практическом плане.

В составе электронного словаря принято различать собственно базу данных (коллекцию словарных статей) и систему программ, осуществляющую работу с этой базой [4]. При разработке такого комплекса создается исходный лингвистический ресурс, предназначенный для формирования КЭС, который представляет собой множество статей, извлеченных из различных словарей, и оформляется в виде базы данных. Создание такой базы данных является чрезвычайно трудоемким процессом, значительно упростить и ускорить который может автоматическая компиляция базы из уже существующих бумажных словарей, преобразованных в электронную форму путем сканирования с помощью современных средств распознавания текста [4]. Словари, на основе которых строится база данных КЭС, в дальнейшем будем называть исходными, а их электронное представление – текстом.

Основную сложность для процесса автоматической компиляции КЭС представляют такие задачи, как:

1) разработка алгоритмов распознавания макро- и микроструктуры исходных словарей;

2) выбор рационального способа представления полученных данных.

Рассмотрим возможности решения этих задач.

1. Распознавание макро- и микроструктуры исходных словарей

Автоматизация процесса создания КЭС требует разработки универсального метода распознавания макро- и микроструктуры исходных словарей. Одним из способов решения задачи распознавания является разработка и программная реализация процедуры распознавания на основе выделенных признаков распознаваемых элементов.

Распознавание макроструктуры словаря основано на выделении границ словарных статей в исходном тексте. Решение этой задачи требует определения признаков словарной статьи и зависит от способа представления статей в словаре – алфавитного, гнездового, алфавитно-гнездового, тематического и т. д.

Распознавание микроструктуры словаря заключается в выделении границ зон внутри словарных статей и их классификации, что может устанавливаться на основе признаков каждой зоны. К признакам границ зон и их типов относятся, в частности, различные специальные символы, отделяющие одну зону от другой, позиционные характеристики расположения конкретной зоны относительно других зон словаря, изменение языка, используемого для описания информации в зоне (для переводных словарей), регистр букв и т.д. На этом этапе важно учитывать, что в разных словарях признаки зон, содержащих информацию одного типа, могут отличаться кардинально.

Универсальная процедура распознавания структуры исходных словарей должна быть основана на полном наборе признаков, характерных для различных словарей. Установление такого набора требует специального исследования макро- и микроструктур большого количества словарей различных типов, объединение которых в КЭС целесообразно с точки зрения использования в работе переводчика.

В качестве пилотного проекта было проведено исследование двенадцати бумажных словарей различного типа: переводных словарей общей лексики [5–8], переводных словарей отраслевой лексики [9–12] и толковых словарей [13–16]. В результате были выделены общие черты организации словарных статей и используемые в них пометы и признаки, на основе чего разработаны базовые алгоритмы распознавания их макро- и микроструктур. Для каждого из исследованных словарей на основе разработанных алгоритмов была создана отдельная программа распознавания его макро- и микроструктуры. Средний процент правильного распознавания структуры соответствующего словаря оказался достаточно высоким – 95,8 %, что доказало адекватность выделенных признаков распознаваемых элементов и эффективность разработки процедуры на основе этих признаков.

Следует иметь в виду, что при построении универсальной процедуры распознавания невозможно полностью учесть все особенности любого привлекаемого словаря, поэтому необходима либо ее постоянная доработка, либо дальнейшая «ручная» проверка распознанных данных. Ошибки в процессе сканирования и опечатки в исходном словаре позволяет устранить «ручная» проверка отсканированного текста [4], проведенная до запуска процесса автоматического распознавания словарной микроструктуры.

2. Выбор рационального способа представления полученных данных

Данные, извлекаемые с помощью процедур автоматического распознавания макро- и микроструктуры исходного словаря, необходимо представить в виде модели организации знаний.

Одной из форм представления знаний является фреймовая модель, основанная на концепции Марвина Минского [17], в которой фрейм понимается как иерархически организованная структура данных, репрезентирующая знания о некоторой стереотипной ситуации или классе ситуаций. Верхние уровни структуры содержат данные, всегда справедливые для анализируемой ситуации, а нижние – пустые узлы, заполняемые конкретными данными соответствующей ситуации [5]. Идея М. Минского получила дальнейшее развитие в работах Р. Шенка [18], Е. Черняка [19], Й. Уилкса [20] и их последователей.

Фреймы, ориентированные не на конкретные ситуации или объекты, а на описывающие их тексты, называют лингвистическими [21]. Лингвистический фрейм представляет собой упорядоченные множества признаков текста и их значений. В качестве признаков могут выступать сведения двух типов: о самом тексте (автор, название и т.п.) и об объектах и ситуациях, описанных в нем [21].

Основным преимуществом фреймов как модели представления знаний является их способность отражать наши представления о концептуальной организации памяти человека, а также их гибкость. Фреймовая модель позволяет достаточно наглядно представлять информацию, что облегчает как ее восприятие человеком, так и дальнейшую компьютерную обработку.

Рассмотрим возможность представления словарной статьи в базе данных КЭС на основе фреймовой модели. Каждая словарная статья при таком подходе организуется как система фреймов. Для определения структуры данной системы необходимо было выбрать общий (универсальный) формат статьи.

В результате проведенного предварительного исследования нескольких словарей [5–16] был разработан формат статьи, включающий в себя все возможные зоны, выделяемые в словарных статьях так, чтобы структура любой статьи каждого рассмотренного словаря вписывалась в этот формат. Структура системы фреймов, предлагаемых для моделирования извлекаемой лексикографической информации, основана на этом формате. Соответственно, система включает набор фреймов разного уровня описания (например, основной фрейм статьи, фрейм зоны заголовка, фрейм зоны перевода и т.д.). Имена слотов фреймов соответствуют наименованиям зон или компонентов зон (например, в зоне «Пример» можно

выделить подзоны «Заголовок примера» и «Перевод примера»). Значение слота – контент соответствующей зоны. Слот либо является указателем на другой фрейм в системе (или набор фреймов такого типа), либо имеет текстовое значение. Общий вид фрейма представлен на рис. 1.

Имя фрейма		
Слот_1	...	Слот_N
значение слота_1	...	значение слота_N

Рис. 1. Общий вид фрейма: N – номер последнего слота фрейма

Фреймовая модель словарной статьи разработанного формата представлена на рис. 2, где связь между слотом и фреймом показана стрелками.

Часто бумажный словарь не просто преобразуется в его электронный аналог, а является основой для создания нового электронного словаря. В таких случаях фреймовая модель статьи в словарной базе данных предоставляет возможности для совершенствования словаря путем его пополнения, так как может содержать процедурные знания наряду с декларативными. Например, при недостатке или полном отсутствии в статье устойчивых сочетаний, примеров или энциклопедической информации значения соответствующих слотов могут быть организованы как процедуры, извлекающие необходимые данные из некоторого корпуса текстов. Вид фрейма «Значение» представленной модели при включении в него присоединенных процедур в качестве значений слотов «Примеры» и «Устойчивые сочетания» приведен на рис. 3.

Словарная статья

Заголовок	Тело
текст	набор фреймов «Тело статьи»

Тело статьи

ФИ	ГСИ	ЭИ	ГВ	УС	СТ	ПС	ЭС	А
текст	текст	текст	набор фреймов «Грамматический вариант»	набор фреймов «Устойчивое сочетание»	набор фреймов «Составной термин»	текст	текст	текст

Грамматический вариант

ГСИ	З
текст	набор фреймов «Значение»

Устойчивое сочетание

Сочетание	П	Т
текст	текст	текст

Составной термин

Термин	П	Т
текст	текст	текст

Значение

ГСИ	П	Т	ЭИ	Пр	СО	СУС	ОУС	УС	ПС
текст	текст	текст	текст	набор фреймов «Пример»	текст	текст	текст	набор фреймов «Устойчивое сочетание»	текст

Пример

Пример	П	Т
текст	текст	текст

Устойчивое сочетание

Сочетание	П	Т
текст	текст	текст

Рис. 2. Фреймовая модель словарной статьи: ФИ – фонетическая информация, ГСИ – грамматическая и стилистическая информация, ЭИ – этимологическая информация, ГВ – грамматический вариант, З – значение, П – перевод, Т – толкование, Пр – пример, СО – смысловые оттенки, СУС – символическое употребление слова, ОУС – особенности в употреблении слова, УС – устойчивые сочетания, ПС – производные слова, СТ – составные термины, ЭС – энциклопедическая справка, А – аналоги

Значение

ГСИ	П	Т	ЭИ	Пр	СО	СУС	ОУС	УС	ПС
текст	текст	текст	текст	процедура «Извлечение примеров из корпуса текстов»	текст	текст	текст	процедура «Извлечение устойчивых сочетаний из корпуса текстов»	текст

Рис. 3. Фрейм «Значение» при включении присоединенных процедур в качестве значений слотов «Примеры» и «Устойчивые сочетания»

Программная реализация процедуры распознавания структуры бумажного словаря показала эффективность распознавания словарных элементов на основе выделенных признаков и выявила основные направления исследования для построения универсальной процедуры распознавания структуры различных бумажных словарей.

ЛИТЕРАТУРА

1. *Gornostay T.* Terminology management in real use / T. Gornostay // Proceedings of the 5th International Conference «Applied Linguistics in Science and Education». – SPb., 2010. – P. 25–26.

2. *Климзо Б. Н.* Ремесло технического переводчика : об английском языке, переводе и переводчиках научно-технической литературы / Б. Н. Климзо. – 2-е изд., перераб. и доп. – М. : Р. Валент, 2006. – 508 с.

3. *Vasiljevs A.* Service model for semi-automatic generation of multilingual terminology resources / A. Vasiljevs, M. Pinnis, T. Gornostay // Terminology and Knowledge Engineering 2014 : Proceedings of the Conference, 19–21 Jun 2014. – Berlin, 2014. – Mode of access: http://tke2014.sciencesconf.org/conference/tke2014/eda_en.pdf

4. *Беляева Л. Н.* Автоматизированная лексикография : гуманитарные технологии / Л. Н. Беляева. – СПб. : Изд-во РГПУ им. А. И. Герцена, 2011. – 75 с.

5. Большой англо-русский словарь : ок. 100 000 слов / авт.-сост. Н. В. Адамчик. – Минск : Литература, 1998. – 1168 с.

6. *Борш А.* Русско-молдавский словарь : ок. 30 000 слов / А. Борш, И. Запорожан. – Кишинев : Главная редакция молдавской советской энциклопедии, 1990. – 504 с.

7. *Ганшина К. А.* Французско-русский словарь : ок. 51 000 слов / К. А. Ганшина. – 7-е изд., стер. – М. : Русский язык, 1977. – 912 с.

8. Русско-латышский словарь : ок. 40 000 слов / А. Гутманис [и др.]. 2-е изд., испр. и доп. – Рига : Аввотс, 1988. – 603 с.

ИСТОЧНИКИ

9. Англо-русский словарь по полиграфии и издательскому делу : ок. 30 000 терминов. – М. : Русский язык, РУССО, 1993. – 582 с.

10. *Газизов М. Б.* Англо-русский химический словарь : ок. 45 000 терминов / М. Б. Газизов [и др.]. – М. : Альфа-М, 2010. – 624 с.

11. Русско-английский медицинский словарь : ок. 50 000 терминов. – М. : Русский язык, 1975. – 648 с.

12. *Хютер П.* Русско-немецкий политехнический словарь : ок. 85 000 терминов / П. Хютер. – 3-е изд., стер. – Берлин : Техника ; М. : Сов. энциклопедия, 1969. – 1271 с.

13. Большой толковый словарь русского языка / сост. и гл. ред. С. А. Кузнецов. – СПб. : Норинт, 2000. – 1536 с.

14. *Крысин Л. П.* Толковый словарь иноязычных слов / Л. П. Крысин. – М. : Эксмо, 2010. – 944 с.

15. *Ожегов С. И.* Толковый словарь русского языка : 80 000 слов и фразеологических выражений / С. И. Ожегов, Н. Ю. Шведова / Российская академия наук ; Ин-т русского языка им. В. В. Виноградова. – 4-е изд., доп. – М. : А ТЕМП, 2006. – 944 с.

16. Словарь русского языка : в 4-х т. / АН СССР, Ин-т рус. яз. ; под ред. А. П. Евгеньевой. – 3-е изд., стер. – М. : Русский язык, 1985–1988.

17. *Минский М.* Фреймы для представления знаний : пер. с англ. / М. Минский. – М. : Энергия, 1979. – 152 с.

18. *Шенк Р.* Обработка концептуальной информации : пер. с англ. / Р. Шенк. – М. : Энергия, 1980. – 360 с.

19. *Charniak E.* Jack and Janet in Search of a Theory of Knowledge / E. Charniak. – Mode of access: http://ijcai.org/Past%20Proceedings/IJCAI-73/PDF/035.pdf

20. *Wilks Y.* Seven Theses on Artificial Intelligence and Natural Language / Y. Wilks. – Mode of access: http://www.issco.unige.ch/working-papers/Wilks-1975.pdf

21. *Гончаренко В. В.* Фреймы для распознавания смысла текста / В. В. Гончаренко, Е. А. Шингарева. – Кишинев : Штиинца, 1984. – 198 с.

Российский государственный педагогический университет имени А. И. Герцена

Беляева Л. Н., доктор филологических наук, профессор, заслуженный деятель науки РФ, профессор кафедры образовательных технологий в филологии

E-mail: lauranbel@gmail.com

Тел.: 8-921-905-71-62

Ефремова А. Н., аспирант кафедры образовательных технологий в филологии

E-mail: e_alena_n@mail.ru

Тел.: 8-911-230-08-33

Russian State Pedagogical University named after A. I. Herzen

Beliaeva L. N., Doctor of Philology, Professor, Honoured Scientist Worker of Russian Federation, Professor of the Educational Technologies in Philology Department

E-mail: lauranbel@gmail.com

Tel.: 8-921-905-71-62

Efremova A. N., Post-graduate Student of the Educational Technologies in Philology Department

E-mail: e_alena_n@mail.ru

Tel.: 8-911-230-08-33