УДК 81’374

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ СТРУКТУРЫ БУМАЖНЫХ ЛЕКСИКОГРАФИЧЕСКИХ ИСТОЧНИКОВ

А. Н. Ефремова

Российский государственный педагогический университет имени А. И. Герцена

Поступила в редакцию 30.05.2016 г.

Аннотация. В статье предложен способ автоматического распознавания структуры бумажных лексикографических источников, который может быть использован в процессе создания основы комплексного электронного словаря для переводчиков путем объединения существующих бумажных словарей различных типов.
Ключевые слова: комплексный электронный словарь, распознавание структуры словаря, макроструктура словаря, микроструктура словаря, словарная статья.
Annotation. The paper presents the method for automatic recognition of printed dictionary structure which can be used for creation of complex electronic dictionary database via incorporating of existing printed dictionaries of different type.
Keywords: complex electronic dictionary, dictionary structure recognition, dictionary macrostructure, dictionary microstructure, dictionary entry.

Одной из задач современной компьютерной лексикографии является создание электронных словарей, среди которых особое место занимают комплексные электронные словари (КЭС). Такие словари объединяют в своей структуре множество словарей различных типов; подобный ресурс, на наш взгляд, может значительно помочь в работе переводчику, тратящему в процессе своей деятельности более 30 % времени на собственно терминологическую работу [1]. Сегодня создано довольно много инструментальных средств для автоматизации отдельных этапов терминологической работы, но нет универсального решения для основных задач извлечения терминологии и ведения комплексного переводческого ресурса [2]. Именно поэтому рассмотрение особенностей создания КЭС представляет особый интерес и в теоретическом, и в практическом плане.

Создание подобного словаря осуществляется в несколько этапов, количество которых зависит от числа объединяемых словарей и необходимости модификации базовой словарной статьи. На первом этапе должна быть создана основа КЭС путем извлечения информации из уже существующих различных лексикографических источников, бумажных в том числе, и объединения ее в форме единого ресурса. Автоматизация этого процесса позволит значительно ускорить создание КЭС. Бумажные словари при этом необходимо предварительно преобразовать в электронную форму путем сканирования с помощью современных средств распознавания текста [3]. Создание КЭС также предполагает решение задач выбора его макро- и микроструктуры [4; 5], при этом выбор микроструктуры словаря для электронного формата представляет собой более сложную проблему, поскольку требует анализа и выявления всего комплекса потенциально возможных и представленных в словарях разных типов видов информации.

Автоматическое извлечение информации из бумажных лексикографических источников требует распознавания структуры каждого исходного словаря, что предполагает последовательное распознавание его макро- и микроструктуры. Распознавание макроструктуры словаря основано на выделении границ словарных статей в исходном тексте, распознавание микроструктуры словаря заключается в выделении границ зон внутри словарных статей и их классификации.

Для определения набора всех возможных распознаваемых зон словарных статей необходимо предварительно разработать такой формат статьи, который включал бы в себя все возможные зоны статей исходных словарей. Подобный формат в дальнейшем может быть взят за основу формата статьи КЭС, так как в статье комплексного словаря должна быть отражена вся информация о лексической единице, полученная из исходных лексикографических ресурсов. Разработка базового варианта словарной статьи КЭС требует проведения предварительного исследования микроструктуры традиционных бумажных словарей. С точки зрения типологии лингвистических словарей [5; 6; 7; 8] принято выделять толковые, двуязычные (переводные) и др. словари; двуязычные словари, в свою очередь, делятся на общие и специальные [4].

В соответствии с такой типологией в качестве экспериментального проекта было проведено исследование двенадцати бумажных словарей различного типа: переводных словарей общей лексики, переводных словарей отраслевой лексики и толковых словарей. В результате исследования был разработан формат статьи, включающий в себя все возможные зоны статей рассмотренных словарей, согласно которому заголовочная единица словарной статьи может быть описана с помощью следующей информации:

- фонетическая информация;
- грамматическая и стилистическая информация;
- этимологическая информация;
- перевод или толкование, пояснения;
- иллюстративные примеры;
- смысловые оттенки;
- символическое употребление слова;
- особенности в употреблении слова;
- словосочетания;
- составные термины;
- производные слова;
- энциклопедическая справка;
- аналоги.

Для разработки алгоритма распознавания макро- и микроструктуры бумажных словарей обратимся к теории распознавания образов, одним из методов которой является распознавание элементов на основе их признаков. Согласно этому методу на основе признаков распознаваемых элементов строится распознающая процедура. Решение задачи распознавания макроструктуры исходного словаря при этом требует определения признаков границ словарной статьи и зависит от способа представления статей в словаре (алфавитного, гнездового, алфавитно-гнездового, тематического и т. д.); решение задачи распознавания микроструктуры словаря требует определения признаков зон и их типов. К таким признакам относятся различные специальные символы, отделяющие одну зону от другой, позиционные характеристики конкретной зоны, изменение языка, используемого для описания информации в зоне (для переводных словарей), регистр букв (верхний/нижний) и т. д. На этом этапе важно учитывать, что в разных словарях признаки зон, содержащих информацию одного типа, могут различаться.

Автоматизация процесса распознавания структуры бумажных словарей требует создания универсальной процедуры, позволяющей автоматически распознавать структуру различных словарей. Такая процедура должна быть основана на признаках не одного конкретного словаря, а на признаках нескольких словарей. Построение такой процедуры требует исследования макро- и микроструктур большого количества различных словарей.

Проведенное нами исследование словарей различных типов позволило выделить некоторые общие характеристики организации статей в них, основные из которых приведены ниже.

Макроструктура словаря

- С точки зрения объемно-графического представления словарная статья обычно составляет абзац, иногда – несколько абзацев.

Микроструктура словаря

- **Заголовок статьи** обычно выделяется жирным шрифтом и/или представлен в верхнем регистре. Может состоять как из одного слова, так и из нескольких. Словарная статья может содержать несколько *альтернативных заголовков* (перечисляются через запятую, точку с запятой, иногда разделяются соеди-

нительным союзом, либо каждый из альтернативных заголовков заключен в круглые/квадратные скобки). В некоторых словарях в заголовке статьи выделяется *неизменяемая часть* с помощью специального символа (часто |, ||, //).

• Омонимы выделяются в отдельные статьи. Информация о **номере омонима** чаще всего представлена сразу после заголовка, иногда – перед ним.

• После заголовка могут располагаться **фонетическая информация** (транскрипция в квадратных скобках), **грамматическая информация**, представляющая собой информацию о части речи (пометы), о возможных формах слова и т. п., а также **стилистическая информация**, представленная в виде помет.

• Статьи толковых словарей обычно содержат **этимологическую информацию** о заголовке, часто представленную в квадратных скобках.

• Различные **грамматические варианты перевода** или толкования чаще всего отделяются друг от друга арабской цифрой с точкой, иногда – римской цифрой с точкой.

• Различные **значения перевода** чаще всего разделяются арабской цифрой со скобкой, иногда – арабской цифрой с точкой.

• Обычно в **переводах** близкие значения разделяются запятой, более далекие – точкой с запятой.

• **Перевод** и **толкование** часто содержат комментарии, данные в круглых скобках курсивом.

• **Примеры** употребления заголовка располагаются в одной строке с переводом и разделяются символом ‘;’. Каждый пример представляет собой заголовок и перевод (толкование). Заголовок примера обычно содержит знак ‘~’, при формировании статьи в КЭС вместо него необходимо подставить либо заголовок статьи, либо неизменяемую часть заголовка. Вместо знака ‘~’ иногда используется сокращение заголовочного слова.

• Словарная статья может содержать примеры **словосочетаний** с заголовком, обычно они расположены после специального символа (часто: ♦, ✧, ◊, ✦).

• Статьи толковых словарей часто содержат дополнительную информацию о заголовке, такую как **смысловые оттенки** (может располагаться после ||), **символическое употребление слова** (может располагаться после |), **особенности в употреблении слова** (может располагаться после |, □).

• Статья в словарях, использующих алфавитно-гнездовую систему, может содержать **составные термины**, каждый из которых располагается с новой строки и имеет вид отдельной словарной статьи (включая перевод, грамматическую информацию и т. п.). Заголовок такой подстатьи содержит символ ‘~’ на месте заголовка основной статьи. Если после символа ‘~’ стоит запятая, то в некоторых словарях после подстановки основного заголовка вместо тильды изменяется порядок слов в заголовке составного термина.

• Статьи толковых словарей часто содержат **производные слова**, расположенные после знаков <, ||.

• В конце статьи толкового словаря после специального символа (часто •, |) иногда представлена **энциклопедическая справка**.

• Статьи толковых словарей могут содержать **зону аналогов**, расположенную в конце словарной статьи после знака ||.

На основе выделенных характеристик бумажных словарей и признаков зон их статей была построена схема алгоритма распознавания макро- и микроструктуры словарей. В дальнейшем планируется совершенствование алгоритма с помощью исследования большего количества словарей и разработка универсальной процедуры распознавания. Следует иметь в виду, что при построении универсальной процедуры невозможно полностью учесть все особенности каждого привлекаемого словаря, поэтому необходима либо постоянная ее доработка, либо дальнейшая «ручная» проверка распознанных данных.

Программная реализация построенной схемы алгоритма распознавания макро- и микроструктуры бумажных словарей подтвердила принципиальную возможность построения единой универсальной процедуры распознавания, основанной на предложенном способе.

СПИСОК ЛИТЕРАТУРЫ

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

2. *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. – Режим доступа: http://tke2014.sciencesconf.org/conference/tke2014/eda_en.pdf

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

4. *Берков В. П.* Двуязычная лексикография: учебник / В. П. Берков. – 2-е изд., перераб. и доп. – М.: ООО «Издательство Астрель»: ООО «Издательство АСТ»: ООО «Транзиткнига», 2004. – 236, [4] с.

5. *Дубичинский В. В.* Лексикография русского языка: учеб.пособие / В. В. Дубичинский. – М. : Наука: Флинта, 2008. – 432с.

6. *Баранов А. Н.* Введение в прикладную лингвистику: учебное пособие / А. Н. Баранов. – М. : Эдиториал УРСС, 2001. – 360с.

7. *Попова Л. В.* Типологии и классификации словарей / Л. В. Попова // Вестник Челябинского государственного университета. – 2012. – № 20 (274). – С. 106–113.

8. *Табанакова В. Д., Сивакова Н. А.* Типология словарей сегодня / В. Д. Табанакова, Н. А. Сивакова. – Вестник Тюменского государственного университета. Социально-экономические и правовые исследования. – 2003. – № 4. – С. 114–119.

ЛЕКСИКОГРАФИЧЕСКИЕ РЕСУРСЫ

Переводные словари общей лексики

9. Большой англо-русский словарь:ок. 100000 слов / авт.-сост. Н. В. Адамчик. – Мн. : Литература, 1998. – 1168 с.

10. *Борш А.* Русско-молдавский словарь: ок. 30000 слов / А. Борш, И. Запорожан. – Кишинев : Главная редакция молдавской советской энциклопедии, 1990. – 504 с.

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

12. Русско-латышский словарь: ок. 40000 слов / А. Гутманис [и др.] / 2-е изд., исправленное и дополненное. – Рига : Аввотс, 1988. – 603 с.

Переводные словари отраслевой лексики

13. Англо-русский словарь по полиграфии и издательскому делу: ок. 30000 терминов. – М. : Рус. яз., РУССО, 1993. – 582 с.

14. Англо-русский химический словарь: ок. 45000 терминов / М. Б. Газизов [и др.] – М. : Альфа-М, 2010. – 624 с.

15. Русско-английский медицинский словарь: ок. 50000 терминов. – М. : Русский язык, 1975. – 648 с.

16. *Хютер П.* Русско-немецкий политехнический словарь: ок. 85000 терминов / П. Хютер. – 3-е изд., стереотип. – Берлин : Техника; М. : Сов. Энциклопедия, 1969. – 1271 с.

Толковые словари

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

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

19. *Ожегов С. И.* Толковый словарь русского языка: 80000 слов и фразеологических выражений / С. И.Ожегов, Н. Ю.Шведова / Российская академия наук. Институт русского языка им. В. В. Виноградова. – 4е изд., дополненное. – М. : ООО «А ТЕМП», 2006. – 944 с.

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

Ефремова А. Н. – аспирантка кафедры образовательных технологий в филологии, Российский государственный педагогический университет имени А. И. Герцена.
Тел.: 8-911-230-08-33
E-mail: e_alena_n@mail.ru

Efremova A. N. – post-graduate student of the Department for Educational Technologies in Philology, Herzen State Pedagogical University of Russia
Tel.: 8-911-230-08-33
E-mail: e_alena_n@mail.ru