ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Н. В. Зыгмантович

АВТОМАТИЧЕСКОЕ СОЗДАНИЕ ОБРАТНОГО СЛОВАРЯ ТЕКСТА

В статье рассматривается история проблемы создания обратного словаря текста. Отмечается, что формирование подобного рода словарей вручную является достаточно трудоемким процессом, требующим значительных временных затрат. С развитием методов обработки лингвистических единиц с помощью компьютера появилась возможность создавать обратные словари в автоматическом режиме. Подчеркивается важная роль таких словарей при анализе словоформ флективных языков лингвистами, а также различными системами автоматической обработки текста. В публикации приводится алгоритм создания обратного словаря русскоязычного текста и демонстрируется принцип его действия с помощью компьютерной программы Cygwin.

Современным средством поддержки работы лингвистов являются информационные технологии и созданные на их основе специализированные лингвистические ресурсы и системы автоматической обработки текста, обеспечивающие решение задач информационного поиска, смысловой компрессии содержания текстовых документов, поиска и извлечения данных и знаний, осуществления перевода, проведения лингвистической экспертизы и т.д. Долгие годы перечисленные выше и целый ряд других систем переработки текста, написанного на естественном языке, разрабатывались в рамках так называемого инженерного подхода, основанного на машинных словарях и определенных правилах. Необходимость создания электронных специализированных словарей, совершенствование компьютерных технологий привели к становлению и развитию новой области – компьютерной или электронной лексикографии.

Компьютерная лексикография является разделом прикладной лингвистики (точнее компьютерной лингвистики) и занимается созданием в автоматизированном режиме электронных словарей (лингвистических баз данных), компьютерных карточек, программ обработки текста, которые позволяют в автоматическом режиме формировать словарные статьи, хранить словарную информацию и определенным образом ее обрабатывать [1, с. 88; 2, с. 274]. К основным задачам компьютерной лексикографии относится определение структуры словаря и зон словарной статьи, а также разработка принципов составления разных видов словарей. Компьютерная лексикография является важным направлением компьютерной лингвистики, поскольку создаваемые ею продукты обладают мультимедийностью, легкостью в использовании и обновлении, а также предоставляют пользователю широкий доступ к необходимой информации.

Электронные словари могут различаться по типу пользователя: одни предназначены для непосредственного использования человеком, другие – для обработки текста определенной компьютерной системой [2, с. 275].

Основные преимущества электронных словарей по сравнению с их бумажными аналогами сводятся к следующему. В отличие от электронного словаря традиционный словарь требует больших расходов полиграфических материалов и с течением времени физически изнашивается. Электронный словарь занимает меньше места и более удобен в работе, чем бумажное издание. Электронный словарь может включать максимально полную информацию и различные справочные материалы, обеспечивает легкий и быстрый поиск информации (может обрабатывать навигационные, общие запросы), а бумажный словарь всегда имеет ограниченный объем и поддерживает традиционный поиск данных (как правило, по алфавиту). В состав электронного словаря можно включать заметки, комментарии, а также электронные ссылки на другие словари или справочные издания. Электронный словарь позволяет визуализировать разные типы данных и обладает вариативным цветовым решением, что не характерно для бумажного словаря. Очень важной характеристикой электронного словаря является наличие автоматизированной или автоматической системы обновления информации. В то же время переиздание бумажных словарей требует больших затрат в области материальных и временных ресурсов [3, с. 182].

С помощью компьютера создаются разные типы словарей: терминологические, частотные, обратные, словоуказатели, конкордансы, тезаурусы и т.д. В контексте данной статьи рассмотрим подробнее историю возникновения, развития, основные характеристики, применение и, самое главное, автоматическое формирование обратного словаря. Необходимо отметить, что отличие обратного словаря от любого другого словарного издания состоит в следующем. В обычных словарях (двуязычных, толковых и др.) слова располагаются в алфавитном порядке, а в обратном – используется так называемый алфавитный порядок по концу слова или обратный (инверсионный) алфавитный порядок, что не означает перевернутого алфавита от буквы *я* до буквы *а*. Из любых двух слов первым в такой словарь помещается то, у которого конечная буква ближе к началу алфавита. Например, в обратном словаре русского языка на первом месте стоят слова, оканчивающиеся на *-а*, потом на *-б*, *-в*, *-г* и т.д.

Родоначальниками обратных словарей считаются средневековые арабские классические словари XIII–XIV вв. В Европе обратный алфавитный порядок слов XVIII в. использовался еще и при составлении словарей рифм (так называемых рифмовников). В конце XIX – начале XX в. появились лингвистические обратные словари древних индоевропейских языков: латинского, древнегреческого, санскрита, тохарского, древнеперсидского и старославянского. Первые обратные словари русского языка были созданы за рубежом: 1958 г. – в Берлине, 1958–1959 гг. – в Висбадене. В середине 50-х гг. XX в. Л. Успенский высказал мысль о том, что лингвистам был бы очень полезен словарь, в котором слова располагались по алфавиту не начальных, а конечных букв. Ученый писал: «Представьте себе, что я захочу узнать что-либо, связанное не с началами, а с окончаниями слов. Ну, положим, какое

значение имеет в русском языке суффикс ‘-л-’ в словах среднего рода, вроде ‘зерка-л-о’? Или каких суффиксов ‘-чик’ в нем больше: тех ли, которые образуют слова, означающие профессию, род занятий (вроде ‘лет-чик’, ‘рез-чик’, ‘пулемет-чик’), или образующих уменьшительные имена (‘маль-чик’, ‘паль-чик’ и пр.). Мне может понадобиться и сведение, какой суффикс более употребителен: ‘-чик’ или ‘-ник’ (а может быть, ‘-тель’) в тех же словах, означающих род занятий (‘гранат-о-мет-чик’ или ‘подрыв-ник’?)» [4, с. 66–67], имея в виду обратный словарь русского языка.

Русскоязычные обратные словари начали издаваться в 70-х гг. XX в. Первый «Обратный словарь русского языка» был создан в 1974 г. и содержал около 125 тыс. слов. В данном словаре при каждом слове указаны словари-источники, имеются грамматические пометы, приводятся статистические сведения о количестве слов, оканчивающихся на определенную букву или сочетание букв, о распределении слов по грамматическим классам [5, с. 416]. Компьютерная обработка словаря проводилась в вычислительном центре Академии наук СССР.

В настоящее время создано большое количество обратных словарей для разных, как правило, флективных языков. В таких словарях представлены грамматическая и словообразовательная структуры слов. Расположение слов в алфавитном порядке, начиная с конца слова, объединяет их с одинаковой конечной частью: сначала идут все слова, оканчивающиеся на первую букву алфавита, потом слова, оканчивающиеся на вторую букву алфавита и т.д., как если бы они читались справа налево [6, с. 112–116]. При совпадении последних букв учитываются предпоследние буквы, далее – третьи с конца и т.д. В обычном словаре рядом оказываются слова с одинаковыми приставками и корнями, например, *переброс*, *перевод*, *перегон*. В обратном словаре рядом находятся слова с одинаковыми окончаниями и суффиксами, например, *банщик*, *барабанщик*, *обманщик*, *денщик*, *каменщик*. При этом, естественно, объединяются слова, относящиеся к единому словообразовательному или словоизменительному типу, а также сложные слова с одинаковой последней составляющей.

Выделить из текста все слова и расположить их в алфавитном порядке, начиная от конца слова, вручную является не простой задачей. Поэтому разработаны различные компьютерные программы, позволяющие создавать обратные словари автоматически. Наиболее удобной из них является Cygwin – среда для запуска Linux-приложений из-под Windows [7]. Определенный набор команд данной среды позволяет формировать разные типы словарей: частотные, обратные, конкордансы.

Рассмотрим алгоритм построения обратного словаря русскоязычного текста программой Cygwin на конкретном примере. Предположим, что в память компьютера поступил следующий текст:

Недавно купил этот автомобиль. Авто отличное! Двигатель 2,5 литра, турбодизель. Прежний хозяин сказал при продаже, что масло не жрет, солярку тоже, летит как угорелая! Так оно и есть. 140 км/ч нормальная крей-

серская скорость. Вообще немцы умеют делать автомобили. Дорогу держит отлично, так как достаточно широкая машина. Тормоза все дисковые. Главное передний привод, по сравнению с другими немецкими автомобилями. Такими как мерседес и БМВ этого же класса. Места в автомобиле очень много. Не тесно даже, когда сидят пять взрослых человек. Багажное отделение тоже огромно. Влезла стиральная машина. По соотношению цена-качество, отличный автомобиль. Всем рекомендую Ауди.

В соответствии с алгоритмом обработки текста компьютер должен выполнить следующую последовательность действий.

1. Заменить каждый пробел в тексте на символ *Enter*, то есть преобразовать текст в словарь словоформ. Фрагмент такого словаря приведен ниже:

недавно
купил
этот
автомобиль.
авто
отличное!
двигатель
литра,
турбодизель.
прежний
хозяин
сказал
при
продаже,
что
масло
не
жрет,
...

2. Удалить из состава некоторых словоформ присоединенные к ним знаки препинания [. , ! - , ? и т.д.]. Данная процедура необходима для правильного выделения в дальнейшем окончаний слов. Фрагментарно словарь очищенных от знаков препинания словоформ представлен ниже:

недавно
купил
этот
автомобиль
авто
отличное
двигатель
литра
турбодизель
прежний
хозяин
сказал
при
продаже
что
масло
не
жрет
солярку
тоже
...

3. Поменять в каждой словоформе словаря буквы справа–налево, например, *недавно – онваден*. Фрагментарно результаты выполнения такой процедуры приведены ниже:

онваден
липук
тотэ
ьлибомотва
отва
еончилто
ьлетагивд
артил
ьлезидобрут
йинжерп
ниязох

лазакс
ирп
ежадорп
отч
олсам
ен
терж
укрялос
ежот
…

4. Отсортировать преобразованный таким образом словарь по алфавиту. Фрагмент полученного словаря представлен ниже:

онваден
липук
тотэ
ьлибомотва
отва
еончилто
ьлетагивд
артил
ьлезидобрут
йинжерп
ниязох

лазакс
ирп
ежадорп
отч
олсам
ен
терж
укрялос
ежот
…

5. Поменять в каждой словоформе словаря буквы слева–направо, то есть вернуть ее в исходную форму, например, *онваден – недавно*. Фрагментарно результаты выполнения такой процедуры приведены ниже:

когда
тормоза
влезла
цена
машина
машина
литра
класса
места
бмв
привод

же
даже
продаже
тоже
тоже
отделение
автомобиле
не
главное
…

В полном виде обратный словарь рассмотренного в качестве примера текста выглядит следующим образом:

когда
тормоза
влезла
цена
машина
литра
класса
места
бмв
привод
же
даже
продаже
тоже
отделение
автомобиле
главное
багажное
отличное
все
вообще
дисковые
ауди
передний
прежний
автомобили
другими
такими
немецкими
автомобилями
при
отличный
как
так
человек
сказал
купил
всем
хозяин
качество
много
этого
масло
недавно
огромно
оно
тесно
отлично
достаточно
авто
что
мерседес
жрет
держит
летит
этот
умеют
сидят
дорогу
солярку
взрослых
км/ч
немцы
турбодизель
двигатель
автомобиль
очень
делать
есть
скорость
пять
сравнению
соотношению
рекомендую
широкая
крейсерская
угорелая
нормальная
стиральная

В отличие от обратных словарей языков синтетического (флективного) типа обратные словари языков аналитического типа имеют несколько иной вид. Приведем фрагмент обратного словаря испанского языка [8]:

Всего слов, оканчивающихся на -a: 14682

Написание в прямой транскрипции	Написание в обратной транскрипции
aba	*aba*
abab	*baba*
ababa	*ababa*
ababer	*rebaba*
abadla	*aldaba*
abaeb	*beaba*
abagla	*algaba*
abah	*haba*
abahcac	*cachaba*
abaj	*jaba*
abajla	*aljaba*
abajnoj	*jonjaba*
...	...

Из вышеизложенного очевидно, что обратный словарь наглядно представляет морфологические характеристики конкретного языка. Если грамматические описания часто содержат утверждения о том, что слова с такими-то окончаниями обладают определенным свойством, то обратный словарь, в котором содержатся группы одинаково оканчивающихся слов, позволяет установить все слова, которые обладают тем или иным свойством, а также те, которые этим свойством не обладают. Обратный словарь необходим для выявления полного инвентаря имеющихся в языке типов словоизменения (склонения и спряжения), а также количественных соотношений между различными типами словоизменений и словообразований [9, с. 20].

Обратный алфавитный порядок удобен для построения грамматического словаря определенного языка. В таком словаре оказываются рядом слова со сходными грамматическими характеристиками, потому что одинаковый или сходный тип словоизменения в пределах одной части речи имеют слова со сходным концом. Например, в русском языке почти все прилагательные в форме м. р., ед. ч., им. п. оканчиваются на *-ый*, *-ий*, *-ой*, поэтому они сгруппированы в одной зоне словаря – в составе слов на *-й*. Можно также выявить все слова, имеющие одинаковое строение концов, но разные грамматические характеристики, и получить данные о соотношении окончания слова и его принадлежности к определенному словоизменительному типу.

Обратный словарь необходим также для решения многих вопросов из области фонетики и морфонологии. В случае необходимости получить количественные данные о фонемах исходный текст требуется представить в фонетической записи. Если словоформы включены в словарь в фонологической транскрипции, то на его основе устанавливаются все допустимые фонемы и конечные их сочетания (по две, три и т.д.). В обычном случае, когда единицы словаря представлены в своей орфографической форме, можно получить сведения о конечных буквах и их сочетаниях.

Обратный словарь предоставляет широкие возможности при проведении семантических исследований с целью изучения общего значения слов некоторой словообразовательной группы. Его можно использовать при корректировке поврежденных или неразборчиво написанных текстов, поскольку он позволяет осуществлять перебор всех слов, имеющих аналогичное окончание, но не поддающихся прочтению. Такой же перебор может понадобиться и при лингвистической дешифровке текста, если конец слова уже расшифрован [9, с. 20].

Обратные словари могут решать немалый круг задач не только в традиционной, но и в компьютерной лингвистике. В настоящее время они применяются в различных системах автоматической обработки текста для определения грамматических характеристик не найденных в специализированных встроенных словарях лексических единиц, например, в системах машинного перевода, системах извлечения информации и знаний, системах понимания и генерации текста на естественном языке и целом ряде других лингвистических текстовых процессоров.

ЛИТЕРАТУРА

1. *Палкова, А. В.* Основные понятия электронной лексикографии / А. В. Палкова // Вестн. ТвГУ. Сер. Филология. – 2015. – № 4. – С. 88–93.
2. *Чепик, Е. Ю.* Компьютерная лексикография как одно из направлений современной прикладной лингвистики / Е. Ю. Чепик // Учен. зап. Таврического нац. ун-та им. В. И. Вернадского. – 2006. – Т. 19. – № 3. – С. 274–279.
3. *Березовская, Е. А.* Современная лексикография: возможности электронных словарей / Е. А. Березовская, Е. В. Сухова // Русский язык: человек, культура, коммуникация: сб. материалов Междунар. науч. конф., 15–16 апреля 2014 г., г. Екатеринбург. – Екатеринбург : Изд-во Урал. ун-та, 2014. – С. 179–183.
4. *Успенский, Л.* Слово о словах / Л. Успенский // Звезда. – 1955. – № 1. – С. 27–32.
5. Обратный словарь русского языка: около 29 000 слов. – СПб. : Авалон; Азбука-классика, 2007. – 416 с.
6. *Еськов, Н.* Что такое «обратный словарь»? / Н. Еськов. – Наука и жизнь. – 1969. – № 12. – С. 112–116.
7. *Лейкинд, Ю.* Что такое CYGWIN? / Ю. Лейкинд [Электронный ресурс]. – Режим доступа : http://www.nestor.minsk.by/kg/2001/04/kg10410.html. – Дата доступа : 11.10.2018.
8. Испанско-русский словарь [Электронный ресурс]. – Режим доступа : https://diccionario.ru/diccionario-inverso-del-espanol/A#ixzz5f0q8LLtR. – Дата доступа : 15.10.2018.
9. *Козьмина, Е. Л.* Обратные словари. Принципы их создания и использования : автореф. дис. … канд. филол. наук : 10.02.19 / Е. Л. Козьмина. – М., 1988. – 20 с.

The article deals with the linguistic problems which may be solved with the help of reverse dictionaries. The paper provides the main aspects and stages of the algorithmic model developing for a Russian Text reverse dictionary based on the computer program Cygwin. The author analyses the results of the developed computer model.

Поступила в редакцию 22.11.2018

Е. М. Яркова

ПСИХОЛИНГВИСТИЧЕСКИЙ ЭКСПЕРИМЕНТ В ИССЛЕДОВАНИИ БИЛИНГВАЛЬНОГО ВОСПРИЯТИЯ СТИЛИСТИЧЕСКИ МАРКИРОВАННЫХ ВЫСКАЗЫВАНИЙ В МУЛЬТИМЕДИЙНОМ КОНТЕКСТЕ

В статье рассматриваются проблемы билингвального восприятия стилистически маркированных лексических единиц в текстах новостных выпусков и газетных статей. Известно, что при достаточно высоком уровне владения языком существенным фактором понимания являются фоновые знания. Особую роль в исследовании устной и письменной речи играет психолингвистический эксперимент.