УДК 004.42:801.3(045/046)

С. Ж. КАРАБАЕВА
КГУСТА ИМ. Н. ИСАНОВА,
БИШКЕК, КЫРГЫЗСКАЯ РЕСПУБЛИКА
E-MAIL:SONUN2008@MAIL.RU
S. ZH. KARABAEVA
KSUCTA N.A. N. ISANOV,
BISHKEK, KYRGYZ REPUBLIC

А. М. МАНСУРОВ
КГУСТА ИМ. Н. ИСАНОВА,
БИШКЕК, КЫРГЫЗСКАЯ РЕСПУБЛИКА
E-MAIL:LOGINIMMENIN@MAIL.RU
A. M. MANSUROV
KSUCTA N.A. N. ISANOV,
BISHKEK, KYRGYZ REPUBLIC

E.mail. ksucta@elcat.kg

МЕТОДЫ КОМПЬЮТЕРИЗАЦИИ ЛЕКСИКОГРАФИЧЕСКИХ РАБОТ И МАШИННЫХ СЛОВАРЕЙ

METHODS OF COMPUTERIZING LEXICOGRAPHIC WORKS AND MACHINE DICTIONARIES

Бул макалада машиналык сөздүк жана компьютердик лексикографиянын изилдөө методу каралган, компьютердик лексикография жөн гана илимдин бир бөлүгү эмес, өзгөчө илим катары саналары сыпатталган.

Чечүүчү сөздөр: *Компьютердик лексикография, машиналык сөздүк, компьютердик анализ, табигый тил, жасалма интеллект, детерминант, блингв, контекстуалдык сөздүк, маалыматтык текст, математикалык модель, механикалаштырылган лексикография.*

В этой статье рассмотрены методы исследования компьютерной лексикографии и машинные словари, утверждено, что компьютерная лексикография представляет собой особую науку, а не является отдельной областью уже определившихся наук.

Ключевые слова: *Компьютерная лексикография, машинный словарь, компьютерный анализ, естественный язык, искусственный интеллект, детерминант, билингв, контекстологический словарь, информатический текст, математическая модель, механизированная лексикография.*

This article is devoted to the computer lexicography machine dictionaries, methods research and, it is confirmed that computer lexicography is a special field of science, and is not a separate field of existing sciences.

Keywords: *Computer lexicography, machine dictionary, computer analysis, natural language, artificial intelligence, determinant, bilingual, contextual dictionary, informative text, mathematical model, mechanized lexicography.*

В настоящее время машинный перевод и компьютерный анализ текстов начинают приобретать независимое значение в сфере гуманитарной информатики.

Обычная лексикографическая технология состоит из следующих процедур: отбора информатических текстов, анализа текстов, составления словников, а также словоуказателей, анализа данных, составления полных или частичных конкордансов, т.е. алгоритма текста

ника по выбранным словам или выписывания из текста иллюстративных примеров использования отобранных слов, составления словарных статей и компоновки словаря.

Словарь в системе машинного перевода играет ведущую роль. Если грамматический анализ для машинного перевода уникален, то в части словаря накоплен большой опыт автоматической обработки текстов, который позволяет сформулировать основные черты компьютерной лексикографии.

Компьютерная лингвистика занимается прежде всего методами автоматического анализа и синтеза, морфологией, синтаксисом, семантикой и особенно словарями, так как тот или иной словарь необходим для работы алгоритма. В настоящее время вопросы компьютерной лексикографии часто ставились в научной литературе. Между тем многие зарубежные ученые неоднократно обращали внимание на тот факт, что машинный словарь в системе автоматической обработки текстов моделирует многие важные функции человеческого интеллекта и поэтому изучение словарей имеет самостоятельное значение для развития искусственного интеллекта.

Задачей компьютерной лексикографии является изучение способов построения и использования машинных словарей естественных языков, т.е. использование компьютерной техники для автоматической обработки текстов. Можно выделить такие направления как лингвистическое обеспечение информационных систем разных типов; машинный перевод; разработка систем, понимающих естественный язык (лингвистические задачи в системах искусственного интеллекта); разработка систем использования информации, содержащейся в звуковом речевом сигнале и др.

Методы компьютерной лексикографии подразделяются на механизированные и машинные. В механизированной машинный словарь представляет собой словарь обычный. К механизированной лексикографии можно отнести все виды использования механических и автоматических средств помощи в обычной лексикографической работе.

В машинной лексикографии машинный словарь используется как орудие автоматической обработки текстов с определенными целями, т.е. часть системы обработки естественных языков.

Механизированная лексикография в использовании словарей не имеет принципиальных отличий от лексикографии обычной. Машинная лексикография действует в пределах науки информатики и подчиняется закономерностям лексикографии.

В словарях хранятся знания, касающиеся слов. Очевидно, что они должны быть всеобъемлющими и достаточно полными, чтобы показать смысл слов (анализ), их форму или другую информацию, относящуюся к языковым производителям. Тем не менее, качество словаря зависит не только от охвата, но и от доступности информации. Стратегии доступа варьируются в зависимости от задачи (понимание текста и информативность текста) и знания. В отличие от читателей, которые ищут значения, авторы начинают с поиска соответствующие слова. В то время как бумажные словари статичны, позволяя использовать только ограниченные стратегии для доступа к информации, их электронные сопоставления обещают динамический, активный поиск по нескольким критериям и через разнообразные маршруты доступа к словам. В электронном словаре навигация происходит в огромном концептуальном лексическом пространстве, и результаты отображаются во множестве форм (например, в виде деревьев, в виде списков, в виде графиков или в алфавитном порядке по темам по частоте).

Главной смысловой особенностью информатического текста словаря является его справочный характер. Информатический текст не должен содержать ошибок в калькуляции сведений или в точности извлечения сведений из исходного документа. Но информатический текст не может рассматривать содержание исходного документа с точки зрения его отношения к действительности, оценивать содержание исходного документа. Поэтому в информатический текст не входят все фигуры стиля документа, его риторическая установка.

Построением информатических текстов в указанном смысле занимаются информационно-поисковые и информационно-логические системы. В той мере, в какой естественный текст, подвергнутый машинному переводу, должен удовлетворять требованиям информационного, соответствующие задачи должны решаться и алгоритмом машинного

вода. Это делает машинный перевод задачей гораздо более сложной, чем прочие лингвистические задачи прикладного характера.

Создание адекватной для машинного перевода грамматики и словаря, безусловно равноценно созданию цельной теории языка или его модели, в случае, если это создание охватывает не фрагмент языка, а достаточно широкую языковую область и позволяет осуществить обработку большого массива текстов.

В компьютерной лексикографии можно отметить общие методы, объединяющие ее с лексикологией и лексикографией, а также с другими науками лингвистического цикла, и частные методы, обусловленные спецификой лингвистических вычислений. Из общих методов наибольшее значение имеет использование языка-эталона для описания значений слов. Практическое применение этого метода имеет место в теории детерминант.

Детерминант – свободная словоформа, находящаяся обычно в начале предложения и осуществляющая грамматическую связь со всей предикативной единицей и является ее распространителем. Среди частных методов важно использование билингв. Кроме того, последнее время больше внимания стало уделяться вопросам математического моделирования словаря.

Метод билингв имеет широкое применение в виде использования параллельных текстов для составления словарей и словников. В качестве примера можно рассмотреть составление семантических частотных словарей. Состав словника при этом во многом зависит от выбранной методики выделения единиц перевода в исходном и переводящем тексте. При составлении англо-русского семантического частотного словаря по информатике единица перевода была определена как минимальный отрезок исходного английского сегмента (слово или словосочетание), для которого в соответствующем русском сегменте нет лексических единиц, передающих лексическое значении составных частей выделенной единицы перевода, если она является словосочетанием. Это рабочее определение исходит из понятия сегмента.

Математическая модель словаря исходит из идеализированных представлений о словаре как некоторой структуре, задающей отношения порядка на множестве слов. Сравнительно с другими объектами математической и вычислительной лингвистики словарь лишь недавно стал объектом математического моделирования. Природа множества, на которых словарь задает отношения порядка, может быть различной. Словарь может рассматриваться и как структура, упорядочивающая буквенные комбинации, и как структура, строящая отношения между элементами смысла, выраженными словами и словосочетаниями.

Оценивая технологические возможности математической теории словаря, следует признать, что здесь еще не получено таких результатов, пользуясь которыми можно было бы оценивать проектируемые или имеющиеся словари по существенным для них содержательным лингвистическим параметрам.

Специфические положения, лежащие в основе теории компьютерной лексикографии:

- соотношение словаря и алгоритма автоматической обработки текстов;
- типология машинных словарей в машинной и механизированной лексикографии с противопоставлением общих для них и специальных для каждой типов словарей;
- представление о языке как много уровневой иерархической системе, основным назначением которой является коммуникация;
- чем больше охват и глубина описания слов для каждой лексической единицы, тем больше будет лексикографическая задача.

Эта система допускает постепенную расшифровку основных черт, релевантных коммуникации. Такое представление лежит в основе использования приближенных методов вычислений применительно к проблемам лексикографии и составления машинных словарей.

Обработка текстов с помощью ЭВМ происходит в тесном взаимодействии человека с машиной.

Машинные словари можно классифицировать по различным признакам. Самой общей классификацией является классификация по двум основаниям: - по характеру лексических единиц, включенных в словарь, и по принципу упорядочения в нем лексических единиц, т.е. по

собу организации словаря. По характеру лексических единиц мы делим словари на: 1) словари основ; 2) словари словоформ; 3) словари оборотов. По способу организации словаря машинные словари подразделяются на: 1) частотные; 2) алфавитные (прямые и обратные); 3) словари тезаурусы; 4) словари-конкордансы; 5) специальные словари, к которым можно отнести, например, автоматический контекстологический словарь для перевода многозначных слов.

Машинная и механизированная лексикография отличаются лишь некоторыми разновидностями из перечисленных выше словарей. Так, в рамках той или иной используются частотные словари, алфавитные, конкордансы и пр. Разница, однако, в принципиальной конечной цели, которой служит словарь. В механизированной лексикографии словарь служит конечной целью исследования. С помощью ЭВМ получаются некоторые данные о лексическом составе текстов или лексических характеристиках языка. В машинной лексикографии машинный словарь используется как орудие автоматической обработки текстов, с его помощью добываются некоторые новые данные.

Преимущества компьютеризации в лексикографии очевидны: компьютер может быстро предоставить доступ к массовому цитатному материалу и быстро дать информацию о многозначном слове, а также позволяют хранить и обрабатывать большие массивы словарной и текстовой информации, т.е. могут использоваться для создания одно- и многоязычных словарей, конкордансов, контекстологических и прочих современных компьютерных словарей.

В наши дни идея компьютеризации лексикографии выполняет важную методологическую роль не только в лингвистике, но начинает проникать в кибернетику, робототехнику и биоинженерию.

Мы надеемся, что интерес к компьютеризации лексикографии среди лингвистов и программистов занимающихся проблемами компьютерной лексикографией позволит получить выдающие результаты в кибернетике и в науке.

В этом отношении, размышляя о новых способах реализации задачи по созданию лексических ресурсов, важны для самих лексикографов, но также для всех, кто интересуется лексиконами как ментальными структурами.

Список литературы

1. PankovP.S., Karabaeva S.J. Mathematical and computer models of spatialconceptsin Kyrgyz language // Интернет-журнал ВАК КР, 2016, № 3. 7 с.
4. PankovP.S., Karabaeva S.J. Independent computer presentation of spatial notions in Turkic languages The Vth International Conference on Computer Processing of Turkic Languages “TurkLang 2017, Kazan. Труды конференции. Том-1., С- 68-79.
5. Karabaeva S. Peculiarities of spatial relations in Kyrgyz language // Abstracts of the Issyk-Kul International Mathematical Forum. - Bishkek: Kyrgyz Mathematical Society, 2015, p. 79.
6. Karabaeva S. Presentation of spatial-temporal relations in Kyrgyz language// Труды международной конференции по компьютерной и когнитивной линг-вистике TEL-2016. – Казань: Изд-во Казанского ун-та, 2016. – С. 274-277.