ПОСТРОЕНИЕ ЧАСТОТНОГО СЛОВАРЯ СЛОВОФОРМ НЕСКОЛЬКИХ ТЕКСТОВ

Клименкова Е.Г.
*Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60, e-mail: klim789@bk.ru*

В современном мире, как научная сфера, так и повседневная жизнь людей невообразима без автоматизированных информационных технологий. На протяжении последних десятилетий их значение стремительно растет. Появление ПК и быстрое развитие кибернетических идей укрепило надежды исследователей-лингвистов в том, что современные точные науки (и прежде всего математика) помогут лингвистике обрести недостающую ей точность. Появилась возможность автоматизировать многие трудоемкие процессы, например, статистическую обработку текстов, ведение разнообразных словарных и лексических картотек. Но с появлением компьютеров почти сразу возникла проблема общения с ними неподготовленных пользователей. Наилучшей формой для таких пользователей мог быть привычный естественный язык. Но для организации такого взаимодействия надо прежде понять законы и особенности использования естественного языка в процессе общения людей между собой [1].

В настоящее время поиск решения проблем автоматической обработки текстовой информации на естественном языке представляет особый интерес. И это объясняется не только тем, что естественный язык является инструментом мышления и общения между людьми, но и тем, что естественный язык — это универсальное средство накопления, хранения, обработки и передачи информации.

Словари играют большую роль в современной культуре, в них отражаются знания, накопленные обществом на протяжении веков. Они служат целям описания и нормализации языка, содействуют повышению правильности и выразительности речи его носителей. Словарь – справочная книга, содержащая собрание слов (или морфем, словосочетаний, идиом и т. д.), расположенных по определенному принципу, и дающая сведения об их значениях, употреблении, происхождении, переводе на др. язык и т. п. (лингвистические словари) или информацию о понятиях и предметах, ими обозначаемых, о деятелях в каких-либо областях науки, культуры и др [2].

В большинстве словарей описывается семантическая структура слов, т. е. словам сопоставляются объяснения их значений и употребления. Термин "словарь" используют для обозначения всей совокупности слов того или иного языка (другими словами, его лексику) и противопоставляют термину "грамматика", который обозначает совокупность правил построения из слов осмысленных речевых отрезков. Лингвистическая наука, которая занимается разработкой методов составления словарей и их изучением, называется лексикографией. В последние десятилетия в рамках лексикографии складывается новое направление – лексикографическая статистика. Лексикографическая статистика занима-

ется созданием частотных словарей и решает связанные с этой задачей вопросы теории и методики создания такого словаря [2].

Частотный словарь — вид словаря, в котором лексические единицы характеризуются с точки зрения степени их употребительности в совокупности текстов, представительных либо для языка в целом, либо для отдельного функционального стиля, либо для одного автора. Словарь может быть отсортирован по частоте, по алфавиту (тогда для каждого слова будет указана его частота), по группам слов и т. д. Частотные словари могут строиться на основе словоформ, лемм (нормальных форм слова) или словосочетаний [3].

Обычно частотные словари строятся не для одного текста, а для корпусов текстов. То есть, берется набор текстов из определенной предметной области или представительный для языка в целом, для конкретного функционального стиля речи, для творчества конкретного автора, и из него извлекаются словоформы, части речи, словосочетания или основы слов. По данным частотных словарей выделяются слова с высокой частотностью и низкочастотные слова. Это позволяет выявить ядро и периферию лексики, разграничить активный и пассивный запас, определить стилистическую принадлежность и жанровую приуроченность лексики, ее социально-возрастное расслоение. Данные о частотности употребления необходимы, например, при установлении авторства текста. Важны данные о наиболее частотных словах и при разработке компьютерных программ проверки орфографии. Частотный словарь также используются для создания эффективных методик обучения языку [3].

Проблемы при создании частотных словарей следующие: воспроизводимость (будут ли результаты идентичны на другом аналогичном корпусе); всплеск частоты отдельных слов (частота слова в одном тексте может повлиять на его позицию в частотном списке); сложность определения позиции менее частотных слов, что не дает возможности ранжировать их рационально; естественно, что практически в любом наборе текстов на первых местах по встречаемости будут служебные слова — союзы, предлоги и т. д.

При разработке прототипа системы, предназначенной для создания частотного словаря словоформ нескольких текстов, использовался метод подсчета слов. На его основе подсчитывается количество использования каждого слова и составляется словарь. Такую программу может использовать преподаватель иностранного языка для составления списка слов, которые необходимо знать его ученикам для изучения какой-либо темы. Также, такая программа может помочь человеку, который самостоятельно изучает иностранный язык. Можно составить словарь до начала чтения книги на другом языке, затем изучить наиболее встречаемые слова, что поможет лучшему пониманию текста.

Список литературы:

1. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. / Г.Г.Белоногов, Ю.П. Калинин, А.А. Хорошилов. - М.: Русский мир, 2004. -248 с.

2. Баранов А. Н. Введение в прикладную лингвистику - М.: УРСС Эдиториал. - 2001. -360 с.

3. Марчук Ю.Н. Компьютерная лингвистика. -АСТ.: Восток-Запад**.** -2007. -320с.