8. *Кейзер, С. В*. Теоретико-методолоические основы анализа стилизованных текстов // Язык, речь, общение в контексте диалога языков и культу: Сб. науч. тр.; под ред О.И. Уланович. – Минск: БГУ, 2012. – С. 25-35.

АВТОМАТИЗАЦИЯ ПОСТРОЕНИЯ ЧАСТОТНОГО СЛОВАРЯ

А. А. Потёмкин

Частотный словарь представляет собой словарь языка, автора или отдельных произведений, в котором слова упорядочены по частоте их встречаемости в источнике. Для этого все словоформы приводятся к начальной форме по определённым правилам, после чего составляется таблица и делаются статистические выкладки.

Более подробно о методике построения частотного словаря и трудностях, с которыми сталкиваются их составители, можно прочитать во введении к частотному словарю русского языка под редакцией Л. Н. Засориной [4, с. 3].

Целью данной работы было применение лингвистических приложений и программ обработки текста для максимально возможной автоматизации процесса создания частотного словаря большого объёма. В работе использовалось следующее программное обеспечение: Microsoft Access 2013, Microsoft Excel 2013, Microsoft Word 2013, Mystem 3.0, Nooj v5, Notepad++ 6.6.7 и xMarkup 3.5. Объём словаря составил 2057 словоформ и 880 лемм.

Алгоритм построения частотного словаря состоит из нескольких частей лингвистического анализа, таких как блок преформатирования и блок морфологического анализа, а также этапа постредактирования с помощью таблиц. Дальнейшая часть работы содержит подробное описание по пунктам построения частотного словаря на примере произведения О. Генри «Комната на чердаке» в переводе В. Маянца.

На первом этапе работы необходимо получить текст в электронном виде и привести его в удобный для морфологического парсера вид. На этапе преформатирования используется программа xMarkup, написанная на языке *icon programming language*. Параметры работы программы содержатся в специальных модулях с расширением *.par*. Для преформатирования нужного нам текста следует использовать поставляемый модуль *lib.par* [1].

Это позволит автоматически избавиться от служебных символов, таких как каретка конца абзаца, выравнивание текста несколькими пробелами, использование дефиса вместо тире и других неточностей, осложняющих дальнейшую работу с текстом. Важно отметить, что, несмотря на хорошие результаты работы программы, текст всё равно следует проверить вручную, чтобы избежать непредвиденных ошибок.

По умолчанию программа устанавливается в папку по адресу *C:/Program Files*, к которой может потребоваться дополнительный доступ в операционной системе windows 7 для записи файлов.

На втором этапе текст необходимо разбить на отдельные слова с их последующим морфологическим анализом и присваиванием грамматических категорий. Для этой процедуры используются модули *_para_num.par, _extr_para_sent_num.par, _extr_words.par*, которые разбивают текст на абзацы, предложения и отдельные слова. Для автоматизации процесса с модулями поставляется исполняемый файл *segment_text.bat*, который запускает их в автоматическом режиме. Его нужно запускать с правами администратора, для более высоких привилегий пользователя в операционной системе Windows 7. Также следует изменить его кодировку, для читабельности кириллических символов в англоязычной системе. Делается это с помощью программы Notepad++.

После этого мы получаем файл со списком всех слов в тексте, который необходимо распарсить с помощью морфологического анализатора от компании Яндекс mystem [2]. Программа предусматривает работу из командной строки, но для облегчения процесса можно воспользоваться ещё одним предустановленным файлом *lemmatize.bat*, который автоматически запустит модули *mystem1.par* и *mystem2.par*, а также сам исполняемый файл *mystem.exe*. Все файлы должны находиться в одной папке, по умолчанию *C:/Program Files/xmwin/bin*.

Полученный файл также необходимо перекодировать и запускать с правами администратора. Для работы файла список словоформ должен находиться в файле *in.txt*. Его можно создать вручную или выгрузить все полученные на предыдущем шаге значения с помощью поставляемой в пакете базы данных *minicorpus* командой «выгрузить слова для лемматизации». В результате пользователь получит текстовый файл с перечнем всех лемм и их грамматических значений.

Такой формат неудобен для дальнейшей работы, так как содержит лишние значения, которые не нужны для построения частотного словаря. Чтобы изменить вид выходного файла, нужно разобраться с параметрами программы mystem. Для этого нужно отредактировать файл *lemmatize.bat*, в котором находится строка *mystem -cgli in.txt out.txt*. Для минимизации вывода данных следует оставить ключи *ci*. Если нужно оставить информацию о частях речи, менять ключи не надо, но придётся избавляться от лишней информации другими путями.

Mystem приведёт все словоформы к начальной форме согласно своим морфологически словарям. В ситуациях с неразрешенной омонимией будет предложено два или более вариантов разбора. На данном этапе полученные значения необходимо внести в таблицу Excel, сверить вручную и проставить части речи. Для этой задачи удобно использовать программу Nooj, в которую можно импортировать текст оригинала. Nooj позволяет находить слова и словосочетания по заданному поиску и показывает их левое и правое окружение, что облегчает снятие омонимии по контексту.

После разбора омонимии и исправления некоторых ошибок парсера все полученные значения следует вынести в сводную таблицу в программе Excel. Совпадающие по форме слова будут автоматически суммированы, после чего их можно будет упорядочить по алфавиту или частотности. Точно так же следует поступить со столбиком части речи [3].

Для построения обратного частотного словаря необходимо воспользоваться сторонним сервисом http://olegon.ru/pr/flip.html, позволяющим автоматически переписать все слова справа налево, тем самым их можно отсортировать по убыванию по последним буквам слова, а затем с помощью того же сервиса вернуть в исходное положение, тем самым получив обратный частотный словарь.

Последующая работа с частотным словарём может заключаться в построении таблиц и графиков по статистическим выкладкам, таким, как самая распространённая буква в тексте, наиболее часто встречаемая буква в первой позиции, соотношение лемм и словоформ и другие. Частотный словарь может применяться для лингвистических задач, связанных с обработкой большого массива информации.

Данная методика не претендует на звание самой правильной, а всего лишь даёт направление для работы и последующего улучшения процесса создания словаря. В конечном итоге, этот процесс должен стать практически полностью автоматическим, за исключением снятия сложных форм омонимии, которую необходимо снимать вручную.

Литература

1. *Баранов А. Г.* Методика лингвистического анализа текста с помощью xMarkup, mystem и minicorpus.mdb // А. Г. Баранов // [Электронный ресурс]. – 2015. Режим доступа: http://minicorpus.narod.ru/files/guide.htm. – Дата доступа: 20/04/2015.
2. Документация программы mystem // [Электронный ресурс]. — 2015. Режим доступа: https://tech.yandex.ru/mystem/doc/usage-examples-docpage/. — Дата доступа: 21/04/2015.
3. *Елисеева О. Е.* Создание частотного словаря словоформ с помощью инструментов Microsoft Word и Excel. // О. Е. Елисеева // [Электронный ресурс]. – 2015. Режим доступа: http://it.lang-study.com/sozdanie-chastotnogo-slovarya-word-excel/. – Дата доступа: 21/04/2015.
4. Частотный словарь русского языка / под ред. Л.Н. Засориной. – М.: Русский язык, 1977. – 936 с.