КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 81'322.2

БЛОК РАЗБОРА ПРОГРАММНОГО КОМПЛЕКСА «СПЛекСИс»

А. А. Кретов, Н. В. Огаркова, О. А. Березовская, Е. В. Долбилова

Воронежский государственный университет

В работе представлена структура прграммного комплекса, предназначенного для проведения лингвистических исследований, направленных на выделение параметрического ядра языка. Рассмотрен «Блок разбора» указанного программного комплекса: сформулированы требования, которым он должен удовлетворять, описана предварительная обработка входных данных. Разработан программный блок для определения настроек словаря, отвечающий за формирование и работу со структурой словарной статьи, а также с общими характеристиками словаря.

ВВЕДЕНИЕ

Большая часть языков представлена исключительно бумажными двуязычными словарями, которые нередко являются если не единственными источниками информации о данных языках, то, как правило, единственными практически доступными источниками такой информации. Поэтому двуязычный словарь является основой лингвистических исследований, направленных на выделение параметрического ядра языка. Оно необходимо для исследования языка и получения его лексико-семантических характеристик.

Работа со словарем-книгой занимает много времени, и нет возможности какого-либо анализа данных, например, подсчета статистик, проведения выборок и т.д. В связи с этим в научно-методическом центре компьютерной лингвистики ВГУ в течение нескольких лет ведется работа по созданию электронных словарей, в состав которых входит не только база данных слов и значений, но и набор средств, позволяющих анализировать эти данные. Созданы автоматический македонско-русский словарь и система CompLex, но они не поддерживают функцию формирования параметрического ядра, необходимого для проведения лингвистических исследований, а кроме этого работают с закрепленной моделью словарной статьи (СС). Разрабатываемый программный комплекс «СПЛекСИс (Система для Проведения ЛЕКсико-Семантических Исследований)» предназначен для изучения языка и проведения его лексико-семантического анализа. При его реализации следует учесть все недостатки ранее разработанных систем.

Двуязычный словарь является основой лингвистических исследований, направленных на выделение параметрического ядра языка, которое необходимо для его анализа и получения лексико-семантических характеристик. Предполагается, что входной информацией для разрабатываемого программного комплекса будут являться файлы, с отсканированным текстом бумажного словаря. На выходе — база данных с занесенной в нее информацией. Затем с данными в базе можно работать (просматривать, редактировать и т.д.) или использовать для проведения лексико-семантического анализа. На рис. 1 представлена структура такого программного комплекса.

Прежде, чем приступить к разбору словарной статьи, необходимо выполнить ряд предварительных работ, которые представляют собой длительный и трудоемкий процесс, направленный на исправление ошибок. Часто эти ошибки обусловлены неправильной работой блока распознавания специализированных программ, предназначенных для сканирования и распознавания текстов, например, FineReader.

1. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ОТСКАНИРОВАННОГО ТЕКСТА СЛОВАРЯ

Ошибки, возникающие после распознания текстов специализированными программами, бывают различными: например, скобки

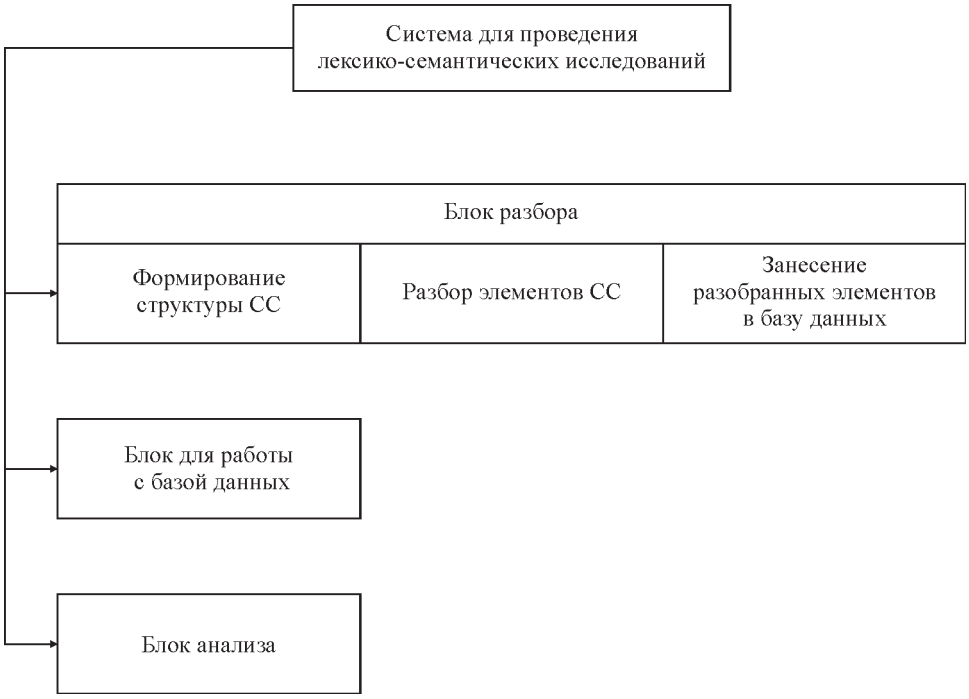

Рис. 1. Структура программного комплекса

вида «(» и «)» определяются как «<» и «>» и наоборот, точка — как запятая, двоеточие — как точка с запятой, английская буква *f* заменяется на символ «/» или «//» и т.д. Такие ошибки можно исправить, внимательно проверив соответствие бумажных и электронных версий словаря.

Но даже после такой проверки сканированной версии словаря остается ряд ошибок, которые невозможно обнаружить визуально: ошибки, связанные с одинаковым написанием букв обоих алфавитов, например, *pena* (русский «ре**п**а») и *pena* (английский «pe**n**a»), *m* (русский «**т**») и *m* (английский «**m**») и т.д. Для обнаружения таких ошибок необходимо использование программных средств.

В некоторых случаях текст словарной статьи не содержит ошибок, но программа не может принять единственное решение о разборе, т.е. определить, с каким элементом имеет дело. Например, в каталано-русском словаре может встретиться следующее сочетание помет *n pr*, которое может означать либо «имя собственное», либо совокупность двух помет: «средний род» и «прочее». В данном случае программа должна сформировать два варианта разбора, а пользователь — выбрать правильный.

При реализации подобных задач были опробованы различные подходы к разбору элементов словарной статьи, например, в македонско-русском автоматизированном словаре элементы заносились в базу по мере распознавания, и при обнаружении ошибки возникала необходимость ее исправления и перезаписи данных в БД. В программе CompLex была реализована функция проверки, которая позволяла выявлять минимальный набор ошибок, но выводы о правильности разбора элементов словарной статьи можно было сделать только после занесения информации в базу данных. Для этого необходимо было просмотреть все данные в таблицах. Если информация в них была некорректна, то возникала необходимость очистки таблиц и самостоятельного отыскания местоположения ошибок в тексте, а затем перезаполнения базы данных. В связи с этим процесс обработки словаря был очень долгим. После анализа неудобств разработанных программных продуктов были сформулированы требования, которым должна удовлетворять подпрограмма разбора элементов СС:

1) занесение данных в базу производить только после исправления ошибок разбора;

2) отображать местоположение ошибок в тексте и предоставлять пользователю возможность их исправлять;

3) в случае успешного разбора статьи пользователю должна быть предоставлена возможность сразу увидеть результат;

4) предлагать варианты в случае неоднозначного разбора элементов.

Одним из способов, удовлетворяющих перечисленным требованиям, является подпрограмма раскраски элементов СС: каждому элементу словарной статьи приписывается определенный цвет, и по мере разбора элементы окрашиваются в соответствующие цвета. После того, как файл с СС раскрашен, и проверены ошибки, элементы словарных статей заносятся в базу данных.

Такой вариант реализации удовлетворяет всем перечисленным требованиям, кроме того, если словарная статья сложная, и не представляется возможным выполнить ее раскраску программно, пользователь может раскрасить текст словарной статьи, выделяя элементы статьи и им назначая (приписывая) определенные цвета.

После проведения разбора элементов словарной статьи (раскраски ее элементов) может потребоваться несколько дополнительных действий.

— Замена встречающихся в тексте символов на соответствующие им элементы словарной статьи, например, в словарной статье

abat *m* аббат. || **fer ~** *(abatre)* <разрушить>.

знак «~» следует заменить на **abat**.

— Удаление незначащих элементов: после раскраски в тексте СС могут появиться элементы, которые не используются при формировании ядер (например, комментарии и пояснения).

Эти элементы можно удалить.

Таким образом, разбор файлов состоит из нескольких этапов:

1) исправление ошибок, которые можно обнаружить при внимательном прочтении текста;

2) раскраска текста и исправление ошибок, обнаруженных программно;

3) занесение раскрашенных и отредактированных элементов словарной статьи в базу данных.

2. ПРОГРАММНЫЙ БЛОК «РАЗБОР СЛОВАРЯ»

Блок «Разбор словаря» предоставляет функции формирования структуры СС, разбора ее элементов и занесение их в базу данных. Этот блок является самым сложным компонентом комплекса, поскольку составители словарей-книг не имеют единого мнения по поводу модели СС. Если программа работает с закрепленной моделью, то нет никакой возможности обрабатывать словари с другими моделями статьи. В связи с этим возникла идея попытаться создать программный комплекс, позволяющий обрабатывать словари с разными структурами СС и требующий при этом минимальных модификаций кода.

Таким образом блок «Разбор словаря» состоит из следующих подзадач:

— формирование структуры словарной статьи (возможность задавать, редактировать, сохранять и загружать созданную структуру СС);

— разбор элементов словарной статьи (подпрограмма раскраски текста СС);

— занесение разобранных элементов в базу данных (сохранение элементов СС, необходимых для лингвистических исследований в БД).

Программный блок, отвечающий за реализацию подзадачи формирования словарной статьи, должен предоставлять возможность формировать общие характеристики словаря, графически задавать сохранять структуру словарной статьи, сохранять эти данные в формат XML и загружать их. Кроме этого, указанный программный блок должен позволить по настроенной модели СС строить внутренние структуры, по которым в дальнейшем будет производиться разбор элементов словарной статьи. На рис. 2 представлена общая схема обработки данных блока формирования структуры словарной статьи и задания общих характеристик словаря.

ЗАКЛЮЧЕНИЕ

В результате был разработан и реализован программный блок для определения настроек словаря, включающий в себя средства для создания проекта словаря, позволяющего сохранять, загружать и использовать информацию об общих характеристиках словаря (списках помет и множествах допустимых символов) и о структуре словарной статьи. На основании сформированной структуры в дальнейшем планируется производить разбор элементов СС для словарей, обладающих различными структурами СС. А затем заносить разобранные элементы в спроектированную БД и предоставлять функции обработки данных и проведения различного вида выборок и подсчета статистик, необходимых для лексико-семантических исследований.

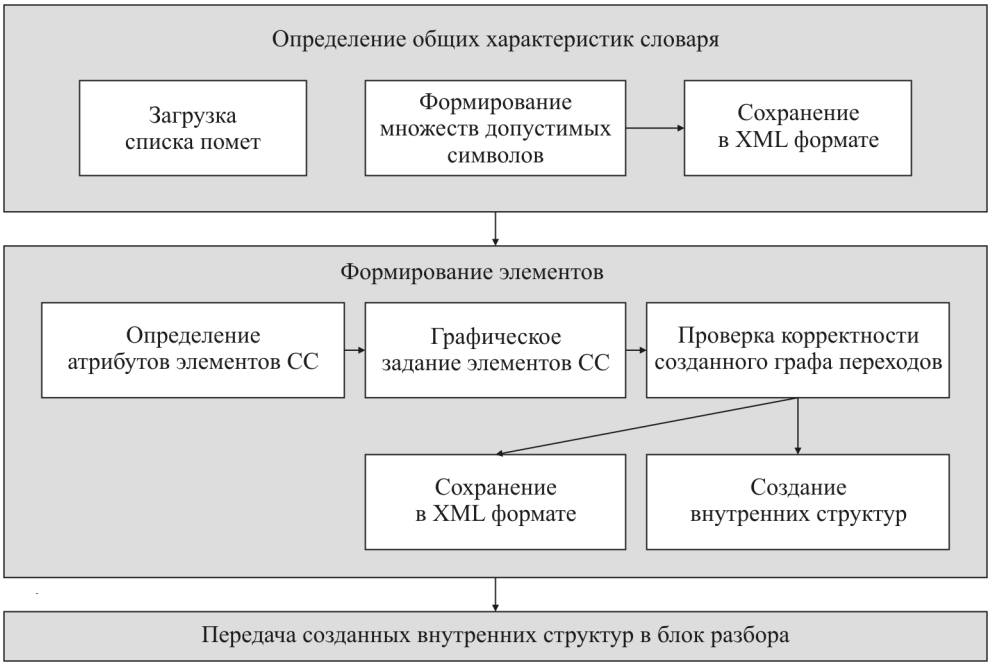

Рис. 2. Схема обработки данных

СПИСОК ЛИТЕРАТУРЫ

1. *Ахо А.* Теория синтаксического анализа, перевода и компиляции. Синтаксический анализ / А. Ахо, Дж. Ульман ; пер. с англ. В. Н. Агафонова ; под ред. В. М. Курочкина. — М. : Мир, 1978. — 616 с.

2. *Караулов Ю.Н.* Лингвистическое конструирование и тезаурус литературного языка / Ю. Н. Караулов // М. : Наука, 1981. — 366 с.

3. *Гасфилд Д.* Строки, деревья и последовательности в алгоритмах: информатика и вычислительная биология / Пер. с англ. И. В. Романовского. — СПб: Невский диалект. БХВ-Петербург, 2003 — 654 с.