

УДК 81'42:811.111-26

Каменский Михаил Васильевич

ЛИНГВИСТИЧЕСКАЯ ПЛАТФОРМА GATE КАК СРЕДА АВТОМАТИЗИРОВАННОГО АНАЛИЗА КОГНИТИВНО-ФУНКЦИОНАЛЬНЫХ СВОЙСТВ ДИСКУРСНЫХ МАРКЕРОВ

В статье подводятся итоги исследования соответствия универсальной лингвистической программной среды GATE требованиям разработки поисково-аналитического модуля когнитивно-функционального анализа дискурсных маркеров.

Ключевые слова: дискурсные маркеры, когнитивно-функциональные свойства, лингвистическая платформа, среда GATE, автоматизированный анализ.

Kamensky Mikhail V.

GATE AS A LINGUISTIC PLATFORM FOR AUTOMATED ANALYSIS OF COGNITIVE-FUNCTIONAL PROPERTIES OF DISCOURSE MARKERS

The article contains the results of the research of applicability of the universal linguistic platform GATE for the development of the specialized module that may be used to perform search and analysis of discourse markers as language units with cognitive functions in discourse.

Key words: discourse markers, cognitive-functional properties, linguistic platform, GATE, automated analysis.

На современном этапе развития лингвистической науки важную роль приобретают исследования в области компьютерной лингвистики, в частности, автоматизированного анализа речетекстовой информации на естественных языках (ЕЯ). Проведение подобных исследований направлено на решение практически значимых задач, таких как идентификация дискурсивной принадлежности текста, индексация текстов в поисковых системах с целью обеспечения автоматизированного поиска релевантной текстовой информации в сети Интернет; автоматизация перевода текстов с одного естественного языка на другой с учетом особенностей коммуникативной ситуации, в том числе социальных факторов и др.

Начиная с 1960-х годов, в процессе работы над созданием современных компьютерных систем, решающих интеллектуальные задачи понимания текстов на естественном языке, выдвигались следующие проблемы:

1) проблема представления знаний (А. Тьюринг 1960; Ю. Д. Апресян 1966; А. К. Жолковский, И. А. Мельчук 1967; Дж. А. Фодор 1975; Р. Г. Пиотровский 1979; К. Дейт 1980; Р. Шенк 1980; П. Уинстон 1980; Т. Виноград 1985; А. Н. Баранов 1987; И. М. Кобозева 1987; Дж. Лакофф, М. Джонсон 1987; Ф. Джонсон-Лэрд 1988; Ю. И. Шемакин 1992; А. Е. Кибрик 1992; Д. Норман 1998 и др.);

2) проблема изучения методов автоматизированного анализа текста и их реализации в проектировании и разработке систем (Г. Сэлтон 1973; Б. Ю. Городецкий 1976; Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская 1977; Д. А. Поспелов 1981; У. Н. Фрэнсис 1983; Ю. Д. Апресян 1988; К. Маккьюин 1989; Н. Н. Леонтьева 1990, 2006; Э. В. Попов, А. Б. Преображенский 1990; Л. Н. Беляева, А. С. Герд, И. И. Убин 1996; И. А. Исаев 1996; А. Н. Баранов 1998; М. Г. Мальковский 2000, К. Селезнев 2003 и др.);

3) проблема исследования обработки ЕЯ-текстов и построения их адекватных моделей (Н. Хомский 1961; И. А. Мельчук 1974; Ч. Филлмор 1981; Дж. Катц 1981; Б. Ю. Городецкий

1983, 1989; А. А. Зализняк 1987; Ж. Г. Мошкович 1989; Д. А. Поспелов 1990; И. П. Панков, В. П. Захаров 1996; А. Вежицкая 1993, 1999 и др.);

4) проблема машинного перевода (И. П. Ревзин, В. Ю. Розенцвейг 1964; Р. Якобсон 1985; Дж. Слокум 1989; А. Д. Бакулов, Н. И. Леонтьева 1990; А. Д. Бакулов, Н. И. Леонтьева, З. М. Шаляпина 1990; Ю. Н. Марчук 1983, 2000 и др.).

Исходя из представленного выше обзора можно заключить, что стремительное развитие компьютерных технологий обработки текстов на естественных языках вызвано жизненной необходимостью решения сложных, трудно формализуемых задач, связанных с различными сферами деятельности человека.

В настоящем исследовании для решения задачи анализа когнитивно-функциональных свойств дискурсных маркеров автором разработана поисково-аналитическая система «Когнитивно-функциональный анализатор дискурсных маркеров», основанный на применении регулярных выражений как современной технологии автоматизированного анализа текста.

История появления регулярных выражений относится к середине 1950-х годов и тесно связана с научными работами по теории автоматов и теории формальных языков. Само понятие регулярных выражений впервые было введено С.К. Клини [1, с. 17–27], одним из основоположников теоретической компьютерной науки. В дальнейшем важный вклад в популяризацию регулярных выражений был внесен К. Томпсоном (1968), установившим, что регулярные выражения удобно использовать для задания шаблонов поиска строк в длинных текстах. По существу, регулярное выражение преобразуется в конечный автомат, производящий поиск совпадений строк с заданными шаблонами. С целью построения такого конечного автомата К. Томпсоном был разработан специальный алгоритм, в настоящее время известный как «построение Томпсона».

Широкое распространение разработанной поисковой технологии обусловило тот факт, что регулярные выражения в современной компьютерной науке и прикладной лингвистике стали *de facto* стандартом описания шаблонов поиска текстовой информации. Регулярные выражения получили распространение в программах для обработки текстов в UNIX (утилита-фильтр *grep*, редактор *sed*), проникли в текстовые редакторы, производные от *ed*, а также в язык программирования Perl. В настоящее время большинство программ, работающих с регулярными выражениями, используют один из вариантов нотации К. Томпсона.

К достоинствам применения регулярных выражений с целью автоматизированной идентификации дискурсных маркеров в связном тексте на естественном языке можно отнести сравнительно высокую скорость обработки данных и высокую точность идентификации найденных языковых конструкций в тексте в случае полного соблюдения заданных формальных критериев поиска. Так, если некоторый дискурсный маркер может быть описан исключительно в структурно-синтаксических терминах (например, по особенностям его синтагматики в предложении или по иным позиционным критериям), то для его автоматизированной идентификации достаточно применения одного или нескольких регулярных выражений, описывающих структурно-синтаксические свойства исследуемого дискурсного маркера в некотором минимальном контексте.

Вместе с тем, очевидным представляется тот факт, что только лишь структурно-синтаксический анализ не может стать единственно необходимым и достаточным критерием автоматизированной идентификации значительной доли дискурсных маркеров. Это объясняется тем, что статус такой языковой единицы как дискурсный маркер определяется в большей мере ее когнитивно-функциональными, нежели структурно-синтаксическими свойствами, что осложняет поиск дискурсных маркеров и идентификацию их когнитивных свойств на основе технологии регулярных выражений.

В современной прикладной лингвистике повышенный интерес проявляется к так называемым технологиям нечеткого поиска в тексте [2], лежащим в основе нечетких регулярных выражений. Классический подход к определению нечетких регулярных выражений базируется на постулировании задачи нечеткого поиска текстовой информации следующим образом: «по заданному слову найти в тексте или словаре размера n все слова, совпадающие с этим словом (или начинающиеся с этого слова) с учетом k возможных различий» [2]. Исходя из поставленной таким образом задачи, нечеткие регулярные выражения определяются как расширение ре-

гулярных выражений, позволяющее учесть варианты трансформации исходного поискового шаблона, выходящие за рамки четко заданных формальных критериев. К данного рода трансформациям относятся, в том числе, опечатки в отдельных словах или пропуск слов [3].

Следует отметить, что описанная выше трактовка нечетких регулярных выражений является недостаточно широкой для целей и задач автоматизированного лингвистического анализа текста, в том числе для автоматизации идентификации дискурсных маркеров в связном тексте и определения их когнитивно-функциональных характеристик. Она представляется недостаточной, так как дополняет классическое определение регулярных выражений исключительно способами идентификации пропущенных или дополнительных элементов, но не позволяет учесть лингвистическую специфику исследуемого явления, в том числе грамматические характеристики искомых языковых единиц (например, их частеречную принадлежность), а также их соотнесенность с определенным семантическим полем. Следовательно, данная трактовка не позволяет выйти на грамматический и семантический уровни автоматизированного анализа текста, необходимые для выявления дискурсных маркеров, которые не могут быть описаны исключительно в терминах синтагматики.

Исходя из вышесказанного, нам представляется целесообразным расширить определение нечетких регулярных выражений, используемых в автоматизированном лингвистическом анализе текста, с учетом последних научных достижений в области прикладной лингвистики. Для настоящего исследования считаем целесообразным принять за обобщенное рабочее определение нечетких регулярных выражений, учитывающее специфику их применения в качестве методологической основы автоматизированного анализа когнитивно-функциональных аспектов использования дискурсных маркеров в корпусе текстов на естественном языке, следующее определение.

Нечеткое регулярное выражение – это регулярное выражение, расширенное путем использования в качестве дополнительного значимого критерия поиска некоторой лингвистической (лексико-семантической, грамматической), экстралингвистической (когнитивной, социальной, культурной) и/или количественной информации, позволяющей идентифицировать и дифференцировать определенные текстовые фрагменты, совпадающие по формальным синтаксическим признакам и различающиеся по иным основаниям, как принципиально различные.

Из этого следует, что автоматизированный когнитивно-функциональный анализ дискурсных маркеров может быть эффективно осуществлен только с применением специализированной программной платформы, реализующей как полнофункциональную возможность поиска текстовой информации по регулярным выражениям, так и возможность учета ряда дополнительных лингвистических, экстралингвистических и количественных параметров.

Так, многофункциональная лингвистическая аналитическая среда General Architecture for Text Engineering (GATE), предназначенная для массовой обработки корпусной текстовой информации, включает в основной установочный пакет набор алгоритмических модулей, позволяющих с высокой точностью осуществлять выявление ключевых грамматических параметров слов и их сочетаний в тексте.

Технология Java Annotation Patterns Engine (JAPE), являющаяся неотъемлемым компонентом GATE, основана на применении регулярных выражений для описания текстовых блоков, подлежащих идентификации и аннотированию, однако шаблонные описания поисковых образцов на основе регулярных выражений могут быть расширены за счет лингвистической информации, получаемой на предварительном этапе автоматизированного анализа текстов, входящих в исследуемый корпус. В частности, автоматизированный анализатор ANNIE, также входящий в установочный комплект GATE, способен проводить предварительный лексико-семантический и синтаксический разбор текстов, отмечая частеречную принадлежность лексем, их семантическую принадлежность к определенным лексическим полям и т. п. [4].

В силу вышеназванных причин в нашем исследовании в качестве программной платформы, служащей основой реализации авторского алгоритмического модуля когнитивно-функционального анализа дискурсных маркеров, выбрана и применена многофункциональная лингвистическая аналитическая среда GATE. Разработка на ее основе нечетких регулярных выражений потребовала применения диалекта языка регулярных выражений, используемого в языке программирования Java. Данная специфика реализации алгоритма объясняется тем, что

названный язык программирования лежит в основе как непосредственно платформы GATE, так и специализированных средств ее модульного расширения.

Как отмечалось выше, регулярные выражения, используемые для поиска и анализа дискурсных маркеров в связном тексте, являются нечеткими и, как следствие, предполагают анализ как собственно структурно-синтаксических конструкций, так и дополнительной лексико-семантической, морфологической, а также экстралингвистической и количественной информации об анализируемом тексте. В связи с этим целесообразность выбора лингвистической аналитической среды GATE и лежащего в ее основе диалекта регулярных выражений продиктована двумя критериями: 1) функциональными возможностями механизма поиска текстовой информации по структурно-синтаксическим основаниям с применением регулярных выражений на основе диалекта Java; 2) потенциалом алгоритмического расширения регулярных выражений с применением специализированных лингвистических, экстралингвистических и количественных метаданных о тексте.

Соответствие лингвистической платформы первому критерию свидетельствует о наличии полнофункциональной аналитической основы идентификации лексем и их сочетаний по формальным структурно-синтаксическим критериям, в то время как ее соответствие второму критерию доказывает применимость платформы для эффективной разработки нечетких регулярных выражений когнитивно-функционального анализа дискурсных маркеров.

Исследование функциональных возможностей механизма поиска текстовых данных с применением регулярных выражений, описывающих позиционные структурно-синтаксические характеристики дискурсных маркеров в тексте, позволило прийти к следующему выводу: диалект языка регулярных выражений Java, лежащий в основе лингвистической аналитической платформы GATE, обладает полным набором функциональных возможностей, необходимых для идентификации структурно-синтаксических параметров дискурсных маркеров в тексте.

Итак, изучение потенциала алгоритмического расширения регулярных выражений GATE с применением специализированных лингвистических, экстралингвистических и количественных метаданных о тексте показало, что отмеченная выше технология аннотирования текста с применением лингвистических алгоритмов JAPE, являющаяся неотъемлемым компонентом среды GATE, предоставляет необходимый и достаточный функциональный минимум для разработки нечетких регулярных выражений. Разрабатываемые на основе данной технологии поисковые выражения позволяют учесть метаданные о лингвистических, экстралингвистических и количественных характеристиках текстов, в том числе устанавливаемых методом автоматизированного анализа с применением аналитических ресурсов GATE. Данные ресурсы являются важными для автоматизации идентификации дискурсных маркеров и их когнитивно-функциональных характеристик. Их значимость объясняется тем, что идентификация в тексте полифункциональных дискурсных маркеров, принадлежащих нескольким функциональным слотам с различающимися иллокутивным потенциалом и когнитивной нагрузкой, значительно затруднена и в некоторых случаях невозможна при использовании традиционных регулярных выражений. Вместе с тем, она становится выполнимой с применением лингвистической, экстралингвистической и количественной информации о контексте, внедряемой в состав нечетких регулярных выражений.

Следовательно, универсальная лингвистическая аналитическая среда GATE может служить платформой для разработки поисково-аналитической среды когнитивно-функционального анализа дискурсных маркеров.

Литература

Клини С. К Представление событий в нервных сетях и конечных автоматах // Сб. «Автоматы». М.: Изд-во иностр. лит., 1956. С. 17–27.

Нечёткий поиск в тексте и словаре [Электронный ресурс]. URL: <http://habrahabr.ru/post/114997/> (дата обращения: 02.09.2013).

Расширенные регулярные выражения – поиск в строках [Электронный ресурс]. URL: http://www.solarix.ru/for_developers/api/regexex.shtml (дата обращения: 02.09.2013).

GATE: General Architecture for Text Engineering [Электронный ресурс]. URL: <http://gate.ac.uk> (дата обращения: 02.09.2013).