УДК 808.2:159.937

Шпаковский Ю. Ф., старший преподаватель (БГТУ)

ИНФОРМАЦИОННАЯ СИСТЕМА ПО ОБРАБОТКЕ И АНАЛИЗУ ТЕКСТОВ, ОЦЕНКЕ ИХ ТРУДНОСТИ ДЛЯ ЧИТАТЕЛЕЙ

В статье предлагается разработать информационную систему по обработке, анализу текстов, их оценке с точки зрения трудности восприятия читателями. Система позволит обрабатывать и анализировать различные виды текстов; автоматизировать морфологический и синтаксический анализ текстов; оценивать трудность текстов для читателей; отбирать материалы, оптимальные для восприятия читателями. Проект предлагается также внедрить во всемирную сеть Интернет.

In the article it is offered to develop information system on processing, the analysis of texts, their estimation from the point of view of difficulty of perception for the future readers. The system will allow to process and analyze various kinds of texts; to automate morphological and parse of texts; to estimate difficulty of texts for the future readers; to select materials, optimum for perception readers. The project is offered to introduce also in an all-peace network the Internet.

Введение. Применение математических методов в лингвистике давно уже стало нормой, однако трудности, с которыми сталкиваются исследователи, до сих пор остается реальностью. Сложность состоит, прежде всего, в неоднозначности языковых единиц, функционирование которых в тексте определяется взаимодействием самых разнообразных лингвистических и экстралингвистических факторов. При попытке перевести языковую абстракцию в математическую исследователь часто наталкивается на то, что «исследуемое явление оказывается практически неразложимым на дискретные единицы, обнаружение которых — необходимое условие успешной статистической обработки» [1].

Однако поиск методов, которые позволят максимально отразить тончайшие нюансы семантико-грамматических связей в тексте, приблизит нас к сути вопроса о восприятии и понимании читателями информации.

В эпоху компьютерных технологий методы математической лингвистики получили новую перспективу развития. Поиск решения проблем лингвистического анализа все активнее реализуется теперь на уровне информационных систем. Вместе с тем автоматизация процесса обработки языкового материала, предоставляя исследователю значительные возможности и преимущества, неизбежно выдвигает перед ним новые требования и задачи.

Цель статьи — разработка информационной системы по обработке, анализу текстов, их оценке с точки зрения трудности восприятия для будущих читателей.

Для этого необходимо выполнение следующих задач:

- статистическая обработка и анализ различных видов текстов;
- анализ существующих методов для определения трудности восприятия текстов для читателей, проведение экспериментов, обработка и анализ результатов;
- разработка комплекса компьютерных программ для автоматизации морфологического и синтаксического анализа текстов;
- оценка трудности текстов для будущих читателей;
- разработка программных средств для поиска материалов, оптимальных для восприятия читателями;
- разработка универсальной программной среды, предоставляющей возможность обрабатывать и анализировать текст, оценивать его трудность для читателя, хранить полученную информацию в виде баз данных.

Основная часть. Современные информационные технологии открывают новые пути развития для статистической обработки текстов. В настоящее время появилась возможность обрабатывать большие объемы текстовой информации за приемлемое время. В данном случае под обработкой понимаются такие действия, как разбиение исходного текста на структурные единицы (слово, предложение, абзац), выделение морфологических и синтаксических признаков единиц текста, количественная обработка результатов.

Задачи автоматизированной или автоматической обработки текстов встают в таких прикладных задачах, как автоматическое индексирование и реферирование текстов, системы извлечения текстовой информации, статистическая обработка специальных текстов, машинный перевод, многоязыковая генерация текстов, автоматизированная оценка трудности текстов, извлечение знаний из больших массивов информации, задачи интеллектуального поиска, задачи установления авторства и др.

Особую роль в методиках автоматической и автоматизированной обработки текстов играют параметры, которые выделяются для текстов. Как было отмечено выше, выбор различных параметров (морфологических, синтаксических, лексических) является сложной проблемой. На начальном этапе нами

предлагается использовать следующие морфологические параметры:

1. Часть речи (имя существительное, имя прилагательное, имя числительное, местоимение, глагол, причастие, деепричастие, наречие, предлог, союз, частица, вводное слово, модальные слова, междометие).

2. Имя существительное (разряд, категории рода, числа, падежа, одушевленности/неодушевленности, типы склонения).

3. Имя прилагательное (разряд, форма (краткая/полная), степень сравнения, категории рода, числа и падежа).

4. Имя числительное (состав (простые, сложные, составные), разряд (количественные, дробные, собирательные, порядковые), категории рода, числа и падежа).

5. Местоимение (разряд (личные, возвратное, притяжательные, указательные, вопросительные, относительные, отрицательные, неопределенные, определительные), категории рода, числа и падежа).

6. Глагол (форма (спрягаемые и неспрягаемые формы глагола, инфинитив), категории лица, вида, залога, переходности, наклонения, времени, спряжения).

7. Причастие (категории переходности, возвратности, вида, залога, времени, рода, числа, падежа, форма (у страдательных причастий наличие как полной, так и краткой формы)).

8. Деепричастие (категории возвратности, вида).

9. Наречие (словообразовательная структура (непроизводные, производные), разряд по значению (определительные, обстоятельственные, знаменательные, местоименные), изменяемость, степень сравнения (сравнительная, превосходная), способ образования степени сравнения).

10. Модальные слова (разряд по значению (выражающие уверенность говорящего в реальности сообщения и выражающие значение возможности, предположения, вероятности сообщаемого)).

11. Слова категории состояния (разряды слов по образованию (слова на -о, соотносимые с наречиями и краткими формами прилагательных (свежо, приятно); слова, этимологически связанные с существительными (пора, время); слова, которые не находят соответствий в других частях речи современного русского языка (нельзя, можно)), по значению (качественные, модальные)).

12. Предлог (разряды по происхождению (первообразные, производные), по структуре (простые, составные), по отношениям (пространственные, временные, объектные, причинные, целевые)).

13. Союз (разряды по происхождению (непроизводные, производные), по структуре (простые, составные), по употреблению (одиночные, повторяющиеся, двойные), по синтаксической функции (сочинительные, подчинительные); сочинительные по значению (соединительные, противительные, разделительные, пояснительные, присоединительные), подчинительные по значению (временные, изъяснительные, причинные, следствия, уступительные, сравнительные, целевые, условные)).

14. Частица (разряды по строению (первообразные, непервообразные; простые, составные (расчленяемые, нерасчленяемые); по функциям (указательные, уточнительные, ограничительные, усилительные, отрицательные, вопросительные, формообразующие)).

15. Вводное слово (значение (модальное значение, обычность совершаемого, указание на источник сообщения, отношение к способу выражения мысли, призыв к собеседнику, связь и последовательность мыслей, эмоциональная оценка, экспрессивный характер высказывания)).

16. Междометие (происхождение (первообразные, производные), выражающие наши чувства и волеизъявления (эмоциональные, повелительные)).

Кроме того, выделены синтаксические параметры, связанные со сложностью организации текста (процент числа простых, сложных, сложносочиненных, сложноподчиненных, придаточных (определительных, дополнительных, обстоятельных) предложений, процент числа причастных и деепричастных оборотов).

При анализе отдельно будут выделены компоненты, не связанные с грамматическими признаками: длина слов и предложений, разнообразие словаря, трудность межсловных связей, процент цитат, идиом, антропонимов, иностранных слов, фразеологизмов, сокращенных слов, аббревиатур, звукоподражательных слов, неязыковых символов и др.

В целом выделено более 200 текстовых параметров. Далее будет разработан комплекс компьютерных программ для их автоматизированной обработки.

На втором этапе исследуются основные методы по определению трудности восприятия текстов читателями.

Выбор методов для определения трудности восприятия текста — ответственная и непростая задача. От этого будет зависеть объективность конечного результата, поэтому особое внимание в работе следует уделить выбору наиболее достоверного и надежного метода для определения понимания текста. По результатам экспериментов можно будет судить о трудности восприятия того или иного текста.

В настоящее время описаны и проанализированы различные методы определения трудности восприятия и понимания текста [2]: по-

становка вопросов по содержанию текста, метод заполнения пробелов, метод выбора резюме, экспертные оценки трудности текста, метод воссоздания авторского текста из фрагментов, метод парных сравнений (основные методы); интонирование, пересказ, обобщение содержания текста, составление плана или схем текста, скорость чтения, метод расшифровки предложений, метод отсроченного воспроизведения (вспомогательные методы).

Анализ показал [2–5], что наиболее надежными являются метод заполнения пробелов, экспертные оценки трудности текста и метод воссоздания авторского текста из фрагментов. В дальнейшем будут продолжены исследования по выявлению наиболее эффективных методов и проведены эксперименты по определению трудности восприятия различных текстов.

Анализ взаимосвязи между степенью понимания текста и его статистическими параметрами — последняя важнейшая задача исследования. Для достижения цели предлагается создать базу данных с ответами испытуемых относительно трудности различных текстов, с одной стороны, и базу данных с морфологическими и синтаксическими параметрами — с другой.

В качестве первого шага для анализа взаимосвязи можно предложить матрицу частот парной встречаемости выделенных классов слов. Для получения такой матрицы следует выбрать систему грамматических классов, достаточно детально описывающую особенности языка и стиля конкретного произведения; затем перекодировать последовательность слов текста в последовательность соответствующих обозначений грамматических классов и подсчитать частоты парной встречаемости для каждой пары классов.

С помощью теории графов можно формализовать алгоритм. По матрице частот парной встречаемости строится граф сильных связей. Две вершины графа (т. е. два грамматических класса) соединяются дугой, если частота встречаемости данной пары грамматических классов равна или превосходит заданный порог. Очевидно, что чем больше величина порога, тем меньше вершин и дуг содержит граф сильных связей. Граф можно обозначить как $G_A(X, Y)$, где X — множество вершин (грамматических форм), а Y — множество дуг (сильных связей грамматических форм).

Далее заданное по определенному критерию число дуг можно объединять в узлы. Другими словами, узел — это такая вершина графа, в которую входит более чем α (заданное число) дуг. Например, при атрибуции текстов используют отношение числа общих для двух сравниваемых текстов «узлов» к суммарному количеству узлов для данных текстов.

В нашей работе возникает архиважный вопрос: как влияют (и влияют ли вообще) частоты парной встречаемости грамматических классов, количество вершин, дуг и узлов на восприятие читателя? И если да, то есть ли оптимальные значения для различных категорий читателей с разным уровнем образования?

Будем исходить из предположения, что при чтении какого-либо текста наибольшее количество информации может быть получено в том случае, если читатель обладает определенным запасом навыков и умений, необходимых для чтения данного текста, и уровень его знаний соответствует тому уровню, на который рассчитан этот текст. При оптимальном сочетании сложности текста и уровня подготовленности читателя (при всех прочих равных условиях) данный текст будет содержать максимальное количество информации для данного читателя. Таким образом, количественная характеристика того, насколько соответствует сложность текста уровню подготовленности читателя — одна из вполне определенных оценок количества информации, содержащейся в данном тексте, относительно некоего тезауруса, соответствующего уровню знаний, необходимых для чтения, и навыкам, которыми обладает читатель.

Таким образом, созданный с помощью программных средств языково-стилистический «портрет» любого текста можно будет сравнивать с «портретом» читателя, составленным на основе его ответов относительно трудности восприятия текстов. Поиск оптимального «портрета» текста позволит извлекать наибольшее количество информации из него.

Поиск оптимальных текстов с точки зрения их трудности для читателей позволит использовать информационные каналы более эффективно.

Всемирная сеть Интернет уже давно и достаточно активно используется в научной среде как средство коммуникации и неограниченного доступа к информационным ресурсам. Последнее время всемирная сеть находит применение и как средство публикации научных трудов. Появились веб-сайты различных научных сообществ, высших учебных заведений, тематические страницы по разным научным дисциплинам, а также сайты научных журналов. Уже сегодня начинают появляться виртуальные лаборатории, где не только студенты, но и все желающие могли бы проводить различные эксперименты, ставить опыты, выбирая параметры и модели экспериментов по своему усмотрению. Более того, в последние годы появляются виртуальные кафедры и виртуальные университеты. Определенная часть образования принимает дистанционную форму. Таким образом, потоки информации возрастают в геометрической прогрессии.

Выходом из этой ситуации является разработка интеллектуальных программных средств, регулирующих поиск информации, оптимальной для восприятия определенной категорией читателей. Все это также возможно на основе создания указанных баз данных.

Заключение. Разработка информационной системы и расширение ее потенциальных возможностей будут продолжены. Наиболее актуальными задачами на данный момент являются: поиск новых параметров, наиболее статистически показательных и значимых для решения поставленной цели; поиск эффективных методов для анализа статистических параметров; выявление наиболее действенных методов определения трудности восприятия текстов; поиск методов для анализа взаимосвязи между степенью понимания текста и его статистическими параметрами. Предстоящие сбор и обработка данных покажут, насколько верны эти усилия.

В любом случае компьютерные технологии позволяют по-новому подойти к решению са мых важных и актуальных вопросов: «Что такое информация?», «Что такое текст?», «Как извлечь максимальное количество информации из текста?».

Литература

1. Ахманова, О. С. О принципах и методах лингвостилистического исследования / О. С. Ахманова [и др.]. — М., 1966.

2. Шпаковский, Ю. Ф. Новая классификация методов определения понимания текста / Ю. Ф. Шпаковский, М. М. Невдах // Труды БГТУ. Сер. IX, Издат. дело и полиграфия. – 2007. — Вып. XV. — С. 100–104.

3. Методика исследования восприятия информации: сб. науч. тр. / под ред. Б. М. Фирсова. — Л.: НИИ ААВ АПН СССР, 1972. — 152 с.

4. Перовский, Е. И. Методическое построение и язык учебника для средней школы / Е. И. Перовский // Известия АПН РСФСР. — 1955. — Вып. 63. — С. 3–139.

5. Микк, Я. А. Методика измерения трудности текста / Я. А. Микк // Вопросы психологии. — 1975. — № 3. — С. 147–155.

Поступила 02.04.2010