Алгоритмизация и программирование. Актуальные проблемы программной инженерии. Материалы 87-й научно-технической конференция профессорско-преподавательского состава, научных сотрудников и аспирантов, 31 января-17 февраля 2023 г.

8. Новикова И.В., Смелова В.В., Шиман Д.В. Планирование валового объема продукции инновационно-промышленного кластера. Управление информационными ресурсами: материалы XIX Международной научно-практической конференции, Минск, 23 марта 2023 г./Академия управления при Президенте Республики Беларусь. – Минск, 2023. –С. 368-370.

УДК 81'33

М.И. Солнышкина
Казанский (Приволжский) федеральный университет
Казань, Россия

ПОДХОДЫ К ОЦЕНКЕ СЛОЖНОСТИ ТЕКСТА

Аннотация. *Дискурсивная комплексология как междисциплинарная область знаний, нацеленная на выявление сложности текста, имеет в качестве объектов исследования лингвистические параметры текста, чтение как когнитивный процесс, а также языковую личность читателя и его способность извлекать информацию из текста. В работе представлены три основных подхода к оценке сложности текста.*

M. Solnyshkina
Kazan (Volga Region) Federal University
Kazan, Russia

APPROACHES TO TEXT COMPLEXITY ASSESSMENT

Abstract. *Discourse complexology as an interdisciplinary field of knowledge, aimed at identifying text complexity, studies text parameters, reading as a cognitive process, reader as a linguistic personality and his ability to process text information. The paper presents and exemplifies main approaches to text complexity assessment.*

Дискурсивная комплексология как наука о сложности текста занимается изучением триады ТЕКСТ – ЧИТАТЕЛЬ – ЧТЕНИЕ, в которой каждый из объектов взаимосвязан и зависим. Несмотря на множество концепций, выдвинутых относительно сложности текста, единой теории в современной науке не выработано. Отдельной

проблемой в рамках дискурсивной комплексологии является и разграничение экстенсионала терминов, функционирующих в данной области знаний: читабельность, удобочитаемость, читаемость, сложность, трудность, понятность текста. В условиях динамичного развития новых форм передачи знания и типов текстов (нелинейных, поликодовых) сформировать устойчивую типологию текстов и зависимых от нее подходов к оценке их сложности становится все более проблематично. Доминирующими в рамках современной научной парадигмы являются три подхода к оценке сложности текста: параметрический, критериальный, а также «датацентричный», лежащий в основе машинного обучения.

В основе параметрического подхода лежит гипотеза о том, что текст как инвариантная единица в рамках дискурса одного типа, одной сферы функционирования, предметной области или уровня сложности может быть описан при помощи ограниченного списка параметров. Классификация этих параметров включает дескриптивные, морфологические, лексические, синтаксические и дискурсивные. Для текстов каждого отдельного типа можно найти уникальный набор параметров, который отличает один тип текста от другого. Например, текст учебника по истории для определенной возрастной группы имеет свой набор значений параметров, отличный от метрик текста научной статьи. Каждый из типов как инвариант объективируется в бесконечном количестве вариантов. При этом сторонники параметрического подхода, признавая его ограниченность, указывают на идиоматичность (эмерджентность) текста как системы, т.е. свойств целостности, не присущие составляющим его элементам. Эмерджентность есть одна из форм проявления принципа перехода количественных изменений в качественные, не позволяющая сводить все смыслы текста к смыслам его составляющих. Принято считать, что идиоматичность свойственна прежде всего художественным текстам [1], сложность которых может быть рассчитана только на основе не параметрического, но критериального подхода. При формализации анализа художественных текстов особое внимание ученые уделяют объему имплицитной информации в тексте или «количеству инференций», а также доли «эксплицитной и имплицитной информации». Когнитивная параметризация текста осуществляется на основе расчета информационных характеристик текста, т.е. объема передаваемой текстом информации. Единицей измерения в этом случае является количество пропозиций и субпропозиций текста [2].

В рамках параметрического подхода были разработаны все формулы читабельности, имеющие в своей основе, как правило, только

две переменные: длина слова и длина предложения. В современной парадигме дискурсивной комплексологии традиционно осуществляется анализ знательно большего количества параметров текста, чем просто длина слова и предложения. Рассчитываются (а) дескриптивные (длина текста / абзаца в слогах, словах, предложениях), (б) морфологические (средняя длина слова в морфемах, доля знаменательных частей речи, доля служебных слов и др.), (в) лексические (лексическое разнообразие текста, количество абстрактных слов, доля высокочастотной лексики, доля различных пластов лексики и др.), (г) синтаксические (средняя длина предложения, количество простых, сложносочиненных, сложноподчиненных, придаточных предложений, количество слов до сказуемого, количество модификаторов в именных группах, минимальное расстояние преобразования и др.), (д) дискурсивных (количество кореферентных наименований, скреп, глубинная связность и др.). Параметрический подход вполне оправдан при анализе текстов официально-делового и научного стилей, например документов, инструкций, научных статей, объем коннотативных смыслов в которых минимален. Именно поэтому материалом в прикладных работах при оценке лингвистической и когнитивной сложности текста выступают преимущественно информационные тексты, используемые в образовании, сфере обслуживания, медицине, строительстве, армии и проч.

Взгляд на текст как систему аддитивных параметров явился фундаментом не только параметрического подхода к лингвистическому анализу, но и толчком к созданию ряда современных пакетов программ автоматического анализа текста. Из наиболее известных укажем на TextInspector, ReaderBench, Coh-Metrix, SourceRater, Lexile framework. Лучшие из них, например, Coh-Metrix или ReaderBench рассчитывают метрики сотен параметров, сгруппированных в следующие кластеры: 1. Параметры лексического и морфологического уровней: количество слогов, часть речи, частота слов, конкретность, образность, многозначность. 2. Синтаксис: структурная сложность, количество модификаторов именных конструкций, количество слов перед сказуемым, синтаксическое сходство между предложениями. 3. Референциальная связность: лексические и синтаксические повторы, латентно-семантический анализ, лексическое разнообразие. 4. Связность ситуационной модели: метадискурсивы, каузальные и интенциональные глаголы, каузальная и интенциональная связность, синтаксический параллелизм, логические операторы. Метрики каждого из параметров валидированы

для решения конкретных задач. Например, функционал Coh-Metrix валидирован для определения жанра и жанровой «чистоты», выявления соответствия текста году (уровню) обучения, различий в связности текстов, написанных авторами различного происхождения и проч. [3].

Параметризация текстов на русском языке и установление референтных значений для текстов различных типов является в настоящее время исследовательской нишей. Интенсивные работы в этой области ведутся в рамках отдельных отечественных научных школ. Автоматизация лингвистического анализа текста на русском языке для носителей языка и иностранцев, изучающих русский язык, осуществлена группой исследователей Государственного института русского языка им. А.С. Пушкина, разработавших портал Текстометр (textometr.ru/).

Аналогичный проект – RuLingva (rulex.kpfu.ru) – запущен в Казанском федеральном университете. RuLingva производит оценку читабельности текста, ранжируя тексты в соответствии с годом обучения. Функционал профайлера включает также извлечение терминов и расчет более 46 параметров и содержит индексы абстрактности, связности, оценку лексической сложности текста с корреляцией по годам обучения. Однако отсутствие валидированных референтных диапазонов для каждого из параметров не позволяет на данном этапе осуществлять комплексную оценку сложности текста с корреляцией на год обучения или объем словарного запаса языковой личности читателя.

Оценка синтаксической трудности восприятия текстов законодательных актов осуществляется учеными Высшей школы экономики (lawreadability.hse.ru/about/) на основе критериального подхода, учитывающего время чтения и количество правильных ответов читателей. При этом важно указать, что сам критериальный подход в отличие от параметрического нацелен на оценку (не)готовности читателя к восприятию и пониманию текста. Именно поэтому данных подход предполагает не расчеты лингвистических индексов текста, а оценку индивидуальных или групповых характеристик, влияющих на объем извлекаемой из текста информации. Групповые характеристики фиксируются при расчете трудности для лиц с определенным уровнем образования, т.е., например, для «среднестатистического» второклассника или студента математического факультета. Оценку трудности текста производят на основе результатов теста по тексту, изложения по прочитанному тексту, заполнения пропусков, воссоздания текста из фрагментов, экспертной оценки и др. Таким образом, в качестве критериев

используют количество правильных ответов или объем воспроизведенной информации. При этом дополнительно осуществляется психодиагностическое тестирование читателя, включающее мотивированность, оперативную память, объем словарного запаса, общую осведомленность и проч. [4].

В настоящее время для лингвистического анализа и расчетов сложности текстов все чаще и чаще применяется так называемый «датацентричный» подход, лежащий в основе машинного обучения. Методы машинного обучения способны одновременно принимать во внимание множество параметров без разделения влияния «веса» каждого параметра на сложность, т.е. алгоритм работает как «черный ящик». Пользователь не получает информации, почему тексту присвоен тот или иной уровень сложности, каков вес в него тех или иных параметров. Особенностью этого подхода является и отсутствие необходимости генерирования математической модели типа текста. Сложность использования этого подхода состоит в необходимости иметь объемный корпус данных – ранжированных, т.е. нормированных для определенного читательского адреса, текстов для обучения «машины» алгоритму. В работе Лапошиой А.Н. и др. [5] этот подход применен для оценки сложности текстов, адресованных изучающим русский язык как иностранный. В качестве мерила сложности была использована Общеевропейская шкала уровней владения языком A1-C2 (CEFR).

Таким образом, современная парадигма дискурсивной комплексологии располагает тремя основными подходами для оценки сложности текста и готовности читателя к восприятию и воспроизведению прочитанного: параметрическим, критериальным и «датацентричным», т.е. использующим методы машинного обучения.

Список использованных источников

1. Оборнева И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. ... канд. пед. наук. М., 2006.

2. Солнышкина М.И., Мартынова Е.В., Андреева М.И. Пропозициональное моделирование для оценки информативности текста / М.И. Солнышкина, Е.В. Мартынова, М.И. Андреева // Ученые записки Национального общества прикладной лингвистики № 3 (31), 2020. – С.47-57.

3. McNamara, D., Graesser, A., McCarthy, P. and Cai, Z. (2014). Automated Evaluation of Text and Discourse with Coh-Metrix, Cambridge University Press, Cambridge, UK. DOI: 10.1017/CBO9780511894664

4. Солнышкина М.И., Гафиятова Э.В. Методика проведения лингвистического эксперимента: к вопросу об определении словаря языковой личности // Russian Journal of Humanities. 2018. Vol. 10, Is. 3-3. P. 275-292.

5. Лапошина А.Н., Веселовская Т.С., Лебедева М.Ю., Купрещенко О.Ф. Автоматическое определение сложности русского текста как иностранного // Компьютерная лингвистика и интеллектуальные технологии: материалы Международной конференции «Диалог». 30 мая – 2 июня 2018 г. М.: РГГУ, 2018. Вып. 17. С. 396-406.

УДК 519.816

Д.С. Соловьев
Тамбовский государственный университет имени Г.Р. Державина
Тамбов, Россия

РАЗРАБОТКА ПОДХОДОВ К ПОВЫШЕНИЮ ОБЪЕКТИВНОСТИ РЕЗУЛЬТАТОВ ПРИ СИНТЕЗЕ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Аннотация. *В статье рассматриваются проблемы объективизации при синтезе систем поддержки принятия решений. Обращается внимание на три основные проблемы: нормализацию данных, коэффициенты значимости критериев и мультивариантность результатов. Для их решения предлагается осуществлять: выбор наиболее объективного метода нормализации, определение согласованных коэффициентов значимости критериев и расчет коэффициентов компетентности участников группового голосования.*

D.S. Solovjev
Derzhavin Tambov State University
Tambov, Russia

DEVELOPMENT OF APPROACHES TO IMPROVE OBJECTIVITY OF RESULTS IN DECISION SUPPORT SYSTEMS SYNTHESIS

Abstract. *The article discusses the problems of objectivization in the synthesis of decision support systems. Attention is drawn to three main problems: normalization of data, significance coefficients of criteria, and multivariate results. To solve them, it is*