УДК 81

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ТЕКСТОВ

И.В. Богословская

Кандидат филологических наук, доцент,
доцент кафедры языковой коммуникации и психолингвистики
e-mail: bakoka@yandex.ru

Уфимский государственный авиационный технический университет

В статье рассматриваются возможные подходы к исследованию компонентов сложности научно-популярного текста, обсуждаются результаты автоматической обработки текстов разной сложности, предложено цветовое наполнение текстов.

Ключевые слова: *сложность текста, научно-популярный текст, компоненты сложности, компьютерная программа, автоматизированная обработка, ассоциативная цветность.*

Проблема сложности текста, к которой исследователи обращаются на протяжении многих лет, до сих пор остается нерешенной. Связано это не с тем, что это проблема глобального характера и она не поддается решению. Прежде всего, ученые не пришли к единому мнению относительно критериев сложности текста. О.С. Разумовский подробно освещает этот аспект проблемы [Разумовский 1999]. Более того, не все исследователи различают понятия трудности, сложности, читабельности. В настоящее время уже существуют математические методы и формулы определения читабельности текста на разных языках [Al-Khalifa, Al-Ajlan 2010]. Также следует отметить тот факт, что в поле зрения исследователей попадают тексты разных жанров, чаще всего ученые обращаются к тексту учебному, что и понятно, поскольку разработка стандартов учебного текста по дисциплинам является актуальной задачей на сегодняшний день.

На протяжении ряда лет мы проводили исследование сложности научно-популярного текста с целью определения универсальных и специфических особенностей процесса понимания текста читателями с учетом их личностных различий [Богословская 2013].

В ходе исследований мы пришли к выводу, что способ организации знаний о мире любого индивида имеет явно выраженную тенденцию к определенному стандарту. Стандартизация структуры тезауруса языковой личности характеризуется, с одной стороны, ее универсальностью у разных членов говорящего на одном языке коллектива и, с другой стороны, специфичностью способов субъективации и индивидуальной

фиксации. Трактовка языковой личности предполагает рассмотрение вариативной, переменной части ее (личности) картины мира, «специфической для данной личности и неповторимой», при условии, что «базовая, инвариантная часть картины мира, единая и общая для целой эпохи, нам известна» [Караулов 2004: 37–38].

Специфика психолингвистической трактовки значения слова предполагает изучение слова как единицы ментального лексикона и средства доступа к единой информационной базе человека, как сложного продукта перцептивно-когнитивно-аффективной переработки индивидом его опыта познания и общения, что позволяет совместить вариативные части двойственной медиативной функции значения слова с двукомпонентным состоянием нормы [Залевская 2005].

В рамках психолингвистического подхода изучение динамических аспектов семантики слова предполагает обращение к индивиду, к семантическим процессам, происходящим в его индивидуальном сознании. Спиралевидная модель семантического развития, разработанная Т.М. Рогожниковой, позволяет моделировать семантическую динамику не только с учетом личностных различий, но и с учетом наличия или отсутствия влияния контекста [Рогожникова 1986].

Научно-популярный текст функционирует одновременно с научным и отражает новое знание опосредованно. Научный текст не прямо трансформируется в научно-популярный, а служит средством понимания при постижении сущности определенной теоретической концепции, которое затем воплощается в соответствующий научно-популярный текст [Богословская 2002; Новиков, Богословская 2003].

Параметры, используемые разными исследователями при оценке текста, связаны с обработкой сенсорной, перцептивной информации, относятся к поверхностной структуре. Уровни смысловой структуры текста также имеют свои компоненты сложности, влияющие на процесс понимания или непонимания. В ходе экспериментального исследования были выявлены такие компоненты. Компонентами сложности понимания научно-популярного текста будем называть текстовые явления, появление которых влечет за собой возникновение зон непонимания.

Таким образом, в данной статье предлагается другое направление, в котором мы будем определять сложности на глубинном уровне. В частности, это идентификация, немотивированная тема и имплицитность. В силу своей значимости и сложности доминирующим элементом является имплицитность. Вторую позицию занимает немотивированная тема, которая связана с имплицитностью и идентификацией [Богословская 2003].

Анализ полного массива ассоциативных реакций, полученных в ходе эксперимента, показал, что структуры ассоциативных полей у экстравертов и интровертов имеют как универсальные, так и

специфические черты, и это в свою очередь свидетельствует об универсальных и специфических особенностях семантического развития слова у интровертов и экстравертов [Богословская 2013].

Анализируя выявленные нами стратегии, которые применялись испытуемыми в процессе понимания текста, мы можем констатировать, что набор стратегий у представителей разных психотипов характеризуется неоднородностью. Успешность процесса понимания достигается различием в использовании экстравертами и интровертами стратегий понимания, в частности стратегии локальной когерентности, продукционных стратегий, стратегий вывода. Кроме того, мы пришли к не менее важному выводу о том, что для экстравертов характерно доминирование процесса генерализации, интроверты же характеризуются доминированием процесса дифференциации.

Таким образом, выдвинув гипотезу о существовании компонентов сложности на глубинном уровне и доказав ее экспериментальным путем, мы подошли к возможности рассмотрения следующего этапа – автоматизированного анализа научно-популярного текста.

Исследования, проводимые Уфимской психолингвистической школой, позволили обозначить наиболее перспективные направления для декодирования суггестивного потенциала вербальной модели и измерения «***силы слова»*** [Рогожникова 2014]. Среди компьютерных программ анализа слова и текста, выполненных молодыми учеными Уфимской психолингвистической школы под руководством профессора Т.М. Рогожниковой, следует отметить такие, как БАРИН, БЮРГЕР, БАТЫР, СЧЕТОВОД. Исследование звуко-цветовых соответствий проводится с целью выяснения закономерностей сложного комплекса психофизиологических процессов, которые лежат в основе человеческого представления о мире в его языковой манифестации [Рогожникова 2015]. Особый интерес исследователей вызывает изучение проблем звуко-цветовой ассоциативности восприятия звуков речи и цветовых закономерностей организации текста, значимости фонетического значения слова и психологических средств воздействия текста. Остановимся подробно на той программе, которую мы будем использовать в нашем исследовании.

Компьютерная программа по автоматизированному анализу слова и текста БАРИН была разработана коллективом авторов (руководитель проекта Т.М. Рогожникова, программист С.А. Воронков, аспиранты Н.В. Ефименко, Р.В. Яковлева) в 2011 году. Группа ученых построила и описала цветовую матрицу звукобукв русского языка. Программа БАРИН написана на языке C# в интегрированной среде обработки MicrosoftVisualStudio 2008. Программный продукт включает в себя несколько модулей: модуль интерфейса; общий модуль для анализа текста, предполагающий работу с любым письменным текстом на

русском или английском языках; модуль определения динамики цветового наполнения текста; модуль построения спиралевидной модели цветового образа текста; модуль художественно-компьютерной интерпретации (художественный образ) звуко-цветовых соответствий в тексте; модуль статистики. С помощью этой программы можно рассчитывать частотность звукобукв русского и английского текстов, вывести результаты в таблицу, определить цветовое наполнение текста на основании рассчитанной частотности звукобукв, представлять результаты в виде графиков и диаграмм, сохранить в файл, выполнить художественную интерпретацию звуко-цветовых соответствий в тексте с сохранением в файл.

Мы воспользуемся этой программой для сравнения научно-популярных текстов разной сложности. Формат статьи не позволяет продемонстрировать все полученные данные, потому мы предлагаем на этом этапе рассмотреть частотность гласных звукобукв текстов, относящихся к разным уровням сложности, а именно: легкий, средний и сложный. Легкий текст называется «Сатурн обошел Юпитер по количеству спутников». Рассмотрим соотношение частотности звукобукв в тексте и в языке. Первая таблица показывает сравнительный анализ гласных звукобукв по частотности употребления в текстах и в языке.

В таблице применяются следующие сокращения: Л – легкий текст, СР – средней сложности, СЛ – текст сложный. Данные представлены в процентном соотношении.

Таблица1

Звукобуква	Частотность в тексте			Частотность в языке
	Л	СР	СЛ	
[и]	7,89	7,04	7,62	6,56
[у]	3,82	1,98	2,26	2,49
[ы]	2,47	2,21	2,55	2,04
[ю]	0,63	0,64	0,75	0,61
[о]	10,01	11,59	10,08	11,2
[е]	7,16	9,52	8,37	8,48
[а]	7,64	7,91	6,82	8
[я]	1,45	2,02	3,14	1,98
[э]	0,15	0,41	0,38	0,38
[е:]	0	0	0	0,06

Анализируя данные, представленные в таблице 1, можно сказать, что в текстах разного уровня сложности наблюдаются различия в использовании гласных. В частности, в легком тексте у четырех звукобукв показатели выше, чем в языке, у пяти звукобукв – ниже; в тексте средней

сложности мы видим, что семь показателей выше, чем в языке и только два – ниже.

Самый сложный текст (следует отметить, что все тексты равны по количеству знаков) характеризуется равным количеством частотности употреблений в языке и тексте только одной звукобуквы. Кроме этого, мы видим, что четыре звукобуквы имеют показателя выше и четыре – ниже, чем в языке. Звукобуква [е:] во всех трех текстах не употребляется.

Компьютерная программа «БАРИН» также позволяет нам представить ассоциативную цветность звукобукв в виде ассоциативных спиралей. Структура каждой спирали включает ядро и периферию. Ведущий цвет находится в ядерной зоне, образуется он благодаря цветности тех звукобукв, которые имеют самый высокий процентный показатель частотности в тексте. Цветовой окрас периферии напрямую зависит от цветности редких и единичных реакций.

Следующие рисунки демонстрируют нам цветовую наполняемость каждого текста. Начнем с легкого текста под названием «Сатурн обошел Юпитер по количеству спутников» (рис. 1).

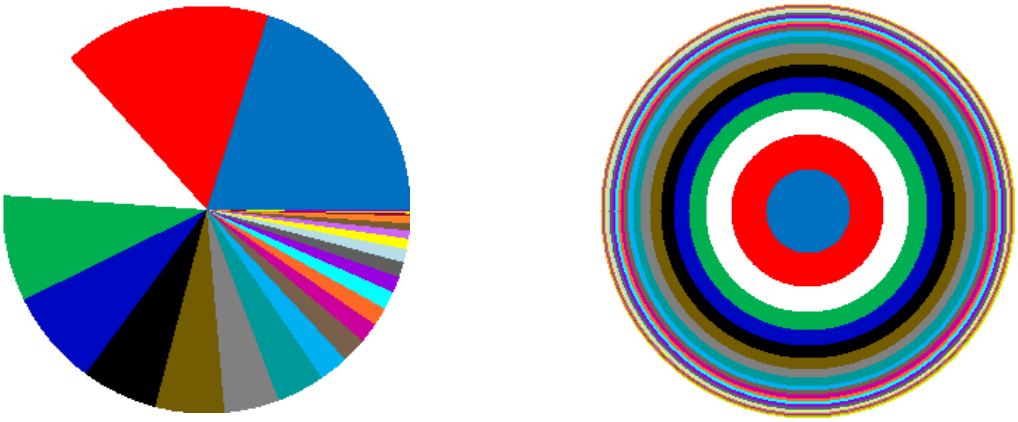

Рис. 1. Ведущий цвет и цветовая спираль легкого текста

На рисунке 2 представлены ведущий цвет и цветовая спираль текста средней сложности, который называется «На межзвездной комете Борисова нашли следы воды» .

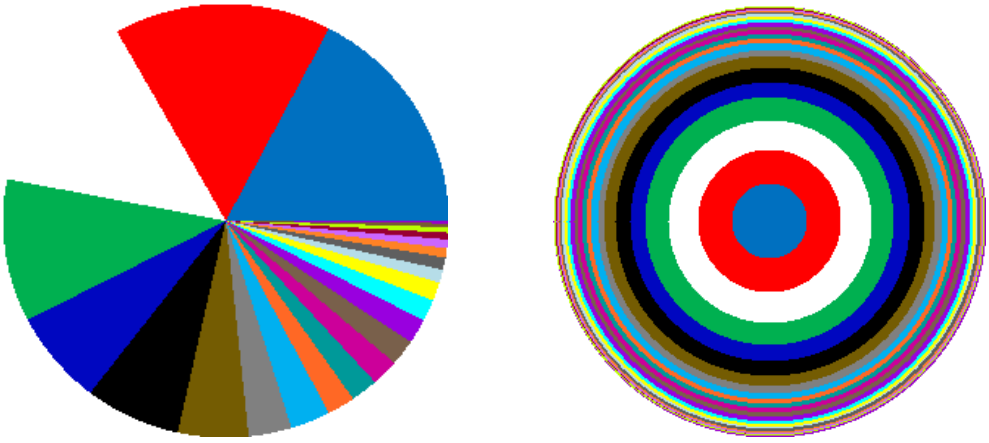

Рис. 2. Ведущий цвет и цветовая спираль текста средней сложности

Сложный текст называется «Голубых бродяг назвали предшественниками магнитаров», и его цветовая спираль показана на рисунке 3.

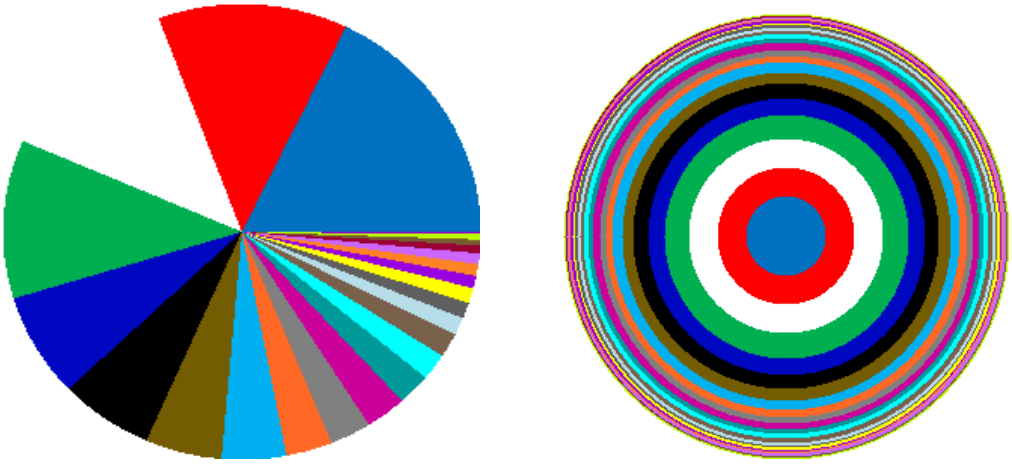

Рис. 3. Ведущий цвет и цветовая спираль сложного текста

Количественный анализ цветовой наполняемости текстов показал, что легкий текст включает 21 цвет, а средний и сложный – по 23 цвета. Качественный анализ демонстрирует определенные различия в периферийной зоне всех трех текстов.

Следует отметить, что исследование находится еще в стадии развития и пока еще рано делать какие-либо выводы, но предварительные результаты показали возможные пути решения проблемы сложности научно-популярного текста.

Библиографический список

Богословская И.В. Соотношение научно-популярного текста с научным текстом // Лингво-методические проблемы обучения

иностранным языкам в вузе: материалы межвузовской научно-методической конференции. Уфа: БГУ, 2002. С. 19–22.

Богословская И.В. Имплицитная информация как возможная сложность понимания научно-популярного текста // Лингводидактические и культурологические аспекты обучения иностранным языкам в вузе: межвузовский сборник научных трудов. Уфа, 2003. Вып. 3. С. 25–29.

Богословская И.В. «Живой» смысл и «мертвая» буква: монография // Уфа: Уфим. гос. авиац. техн. ун-т, 2013. 195 с.

Залевская А.А. Психолингвистические исследования. Слово. Текст. Избранные труды. М.: Гнозис, 2005. 543 с.

Караулов Ю.Н. Русский язык и языковая личность. М.: Едиториал УРСС, 2004. 264 с.

Новиков А.И., Богословская И.В. Научно-популярный текст в его соотношении с текстом научным // Когнитивное моделирование в лингвистике: сб. докладов. Varna, 2003. С. 346–356.

Разумовский О.С. Оптимология. Ч.1: Общенаучные и философско-методические основы. Новосибирск: Изд-во ИДМШ, 1999. 285 с.

Рогожникова Т.М. О спиралевидной модели развития значения слова у ребенка // Психолингвистические проблемы семантики и понимания текста. Калинин: Калинин. гос. ун-т, 1986. С. 100–105.

Рогожникова Т.М. Автоматизированный анализ вербальной информации как декодирование суггестивного потенциала языковой системы // Вестник Уфимского государственного авиационного технического университета. 2014. Т. 18. № 2 (63). С. 113–124.

Рогожникова Т.М. Инструменты анализа суггестивных ресурсов языковых явлений // Когнитивная психология: методология и практика. Коллективная монография / Под науч. ред. В. М. Аллахвердова и др. СПб.: Изд-во ВВМ, 2015. С. 233–244.

Al-Khalifa H.S., Al-Ajlan A.A. Automatic Readability Measurements of the Arabic Text: An Exploratory Study // The Arabian journal for science and engineering, 2010. 35 p.