5. *Элькин Д.Г.* Восприятие времени. – М.: Изд-во Академии педагогических наук РСФСР, 1962. – 310 с.

APPROACHES TO THE VALUES AND ANTI-VALUES MEANING DESCRIPTION

Chronopsycholinguistics approach to the values and anti-values study allows us to describe the values components meaning in a certain historical period and its influence on the language consciousness images perception and interpretation in different cultures. The consciousness is studied at 3 levels (the public consciousness official level, the intermediate level of social consciousness, the common consciousness level).

Key words: universal values, anti-values, semantic components, content, chronopsycholinguistics, language consciousness, time, chronopsychology.

А.А. ЧУРУНИНА, М.И. СОЛНЫШКИНА, Э.В. ГАФИЯТОВА, А.А. ЗАИКИН

(Казанский (Приволжский) федеральный университет)

СЛОЖНОСТЬ ТЕКСТА КАК ФУНКЦИЯ ЛЕКСИЧЕСКИХ ПАРАМЕТРОВ (НА МАТЕРИАЛЕ УЧЕБНЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ)

Представленное исследование осуществлено в рамках проекта по исследованию параметров академических текстов различного уровня и их влияния на сложность данных текстов. Предметом исследования явилась сопоставительная сложность текстов учебников по обществознанию для общеобразовательной школы. Тексты из семи учебников были оценены на основе десяти количественных и четырех лексических параметров, на основание чего было установлено, что лексическое разнообразие, частотность и

абстрактность лексики, а также количество терминов являются параметрами сложности текста с высокой степенью прогнозирования.

Ключевые слова: абстрактность, сложность текста, лексическое разнообразие, частотность.

Оценка лингвистической сложности русскоязычных текстов как научная проблема имеет более чем пятидесятилетнюю историю [Колмогоров, 1965; Оборнева, 2006] и продолжает сохранять теоретическую значимость вследствие недостаточной разработанности алгоритма определения «целевого читательского адреса» и отсутствия валидированного списка параметров, определяющих сложность текста. Научная парадигма данной области знаний выделяет три основных лексических параметра, влияющих на сложность академического текста: коэффициент лексического разнообразия, частотность и абстрактность [McNamara, 2014].

Активные дискуссии относительно расширения списка параметров, определяющих сложность текста, ведутся более ста лет [Solnyshkina, 2020].

Читабельность определяет уровень удобочитаемости текста и рассчитывается исключительно на основе количественных параметров: 1) среднего количества слов в предложении; 2) среднего количества слогов в слове. Адаптированная для русскоязычных текстов формула была апробирована в многочисленных исследованиях, подтверждающих ее надежность [Solnyshkina, 2018]: ФК (SIS) = 208,7 – 2,6 × СДП – 39 × СДС, (1) где СДП – это средняя длина предложения в словах, а СДС – это средняя длина слова в слогах. Однако количественные характеристики текста не позволяют объективно сравнивать тексты, т.к. они не отражают всех характеристик, которые также влияют на читабельность текста.

При сравнении текстов по сложности Д. Байбер [Biber, 2006] указывает, что повышение сложности сопровождается ростом числа существительных, снижением количества глаголов, большей степенью номинализации глаголов и прилагательных, а также ростом количества абстрактных существительных и длинных слов.

Индекс лексического разнообразия (TTR) широко используется при анализе сложности текстов с 1957 года, когда его впервые ввел М. Темплин [Templin, 1957]: $\mathrm{TTR} = \frac{\text{word types}}{\text{word tokens}}$, (3) где ‘word types’ – это уникальные, т.е. не повторяющиеся, слова текста, а ‘word tokens’ – это общее количество слов в тексте. Исследования, проведенные в начале 2000-х, доказали, что распределение уникальных лексических единиц в тексте не является линейным для корпусов различных размеров, т.к. слова имеют тенденцию повторяться: чем больше корпус, тем больше повторяющихся слов в нем содержится. Таким образом, лишь относительно небольшое количество уникальных слов будет увеличиваться наряду с увеличением объема корпуса, что обуславливает ряд дополнительных проблем при сравнивании корпусов разных размеров. Ввиду вышеуказанных сложностей было предложено рассчитывать TTR для текстовых отрывков, объем которых не превышает 1000 слов [Biber, 2006].

Другой параметр, напрямую влияющий на сложность текста – это частотность лексики: чем больше в тексте использовано частотных слов, тем более легким он является для восприятия читателем.

Индексы частотности для русского языка, зафиксированные в Словаре частотности [Sharov, 2009], успешно используются для оценивания сложности текста [Zinsmeister, 2015; Solnyshkina, 2019].

В случае недоступности списков частотной лексики или недостаточного описания частности в рамках корпуса, исследователи останавливают свой выбор на более простых

для измерения параметрах текста, позволяющих определить его сложность, например, количестве терминов. Р.В. Майер вводит новое понятие сложности текста – дидактическую сложность, – которая основывается как на количестве терминов в тексте, так и на количество математических символов и информационную плотность текста [Mayer, 2016].

Еще одним параметром, определяющим сложность текста, является степень абстрактности [Solovyev, 2019; Xu, 2020; Borghi, 2016].

Исследование осуществлено с целью выявления зависимости сложности учебного текста от кластера лексических параметров, включающих лексическое разнообразие, количество терминов, частотность лексики и абстрактность.

Материал исследования составили 70 учебных текстов, извлеченных из учебников серии «Обществознание, 5 – 11 классы» [Боголюбов 2012 – 2014] общим объемом более 160 тыс. словоупотреблений. Расчеты лексических параметров и сложности текстов осуществлены при помощи онлайн сервиса RusAC [Solnyshkina, 2019], для выявления взаимосвязи параметров использовался коэффициент Спирмена.

На первом этапе исследования на основе Русского академического корпуса [Solovyev, 2019] нами был составлен Русский корпус текстов по обществознанию (РКТО). Данный этап также включал подсчеты, нацеленные на выявление объема РКАТО; объема каждого учебника в составе РКАТО и длины десяти отрывков из каждого учебника. Каждому отрывку из учебников был присвоен свой код, указывающий на класс, название предмета и номер отрывка в выборке. Например, ‘5SS1’, где отрывок 1 из учебника 5 класса по обществознанию (‘SS’ - ‘Social Science’).

Таблица 1. Объем и структура Русского корпуса текстов по обществознанию

Класс	Учебник	Выборка
5	10083	1008
6	10135	1013
7	11226	1122
8	24027	2402
9	21184	2118
10	38440	3844
11	52803	5280
РКТО	**167898**	**2398**

Проблема репрезентативности корпуса связана не только с его объемом, но и их жанровым однообразием [Biber, 2006]. Русский корпус текстов по обществознанию, используемый в данном исследовании, определяется как репрезентативный, поскольку он представляет одну форму языка, а именно – учебных текст по обществознанию. Длина отобранных текстовых отрывков определяется на основе формуле Д. Байбера [Biber, 2006]: T (s) = T(t) : 20, (4) где T(s) – это количество слов в отрывке, а T(t) – количество слов в учебнике. Размер отрывков варьируется от 1008 словоупотреблений для 5 класса до 5280 словоупотреблений для 11 класса соответственно.

На втором этапе все 10 отрывков из каждого учебника были проанализированы посредством автоматического сервиса обработки текстов RusAc [RusAC, 2008]. Список рассчитываемых параметров включает следующие: 1) общее количество слов в тексте; 2) общее количество слогов; 3) общее количество предложений; 4) среднее количество слов в предложении; 5) среднее количество слогов в слове; 6) количество прилагательных; 7) количество наречий; 8) количество местоимений; 9) количество существительных; 10) количество глаголов; 11) частотность; 12) ФК; 13) индекс абстрактности; 14) индекс лексического разнообразия; 15) количество терминов.

Таблица 2. Параметры сложности текста

Параметр / класс	5	6	7	8	9	10	11
Кол-во слов	1008,30	1013,50	1122,60	2402,70	2118,40	3844,00	5280,30
Кол-во слогов	2510,10	2505,10	3006,70	6561,60	5956,30	11101,70	15705,50
Кол-во предложений	85,10	92,00	107,90	209,30	188,80	297,50	384,10
Слов/ Предложений	11,90	11,08	10,52	11,66	11,32	12,98	13,77
Слог/ слов	2,49	2,47	2,68	2,73	2,81	2,89	2,98
Прилагательные	129,30	122,80	152,40	360,80	329,60	648,60	954,40
Наречия	47,90	46,40	43,50	107,00	73,30	138,40	185,20
Местоимения	98,20	99,20	116,20	243,40	219,80	394,00	560,50
Сущ-ные	345,20	330,70	422,20	906,60	846,60	1517,20	2162,50
Глаголы	168,90	178,20	184,70	329,90	278,80	468,50	602,40
Частотность	134,31	143,79	128,27	117,97	116,18	113,10	104,65
ФК	6,66	6,26	7,24	7,97	8,30	9,34	10,12
Абстрактность	-1,69	-1,89	-1,89	-2,00	-1,50	-2,17	-2,05
TTR	0,63	0,63	0,63	0,54	0,52	0,51	0,47
Термиы	18,90	18,00	38,10	60,00	124,20	91,50	167,00

Результаты проведенного исследования представлены на рисунках 1-4.

На рис. 1 продемонстрирован стабильный рост показателей индекса ФК с 6,26 (6 класс) до 10,22 (11 класс). Однако график иллюстрирует также, что в большинстве случаев показатели читабельности ниже уровня подготовки заявленной целевой аудитории.

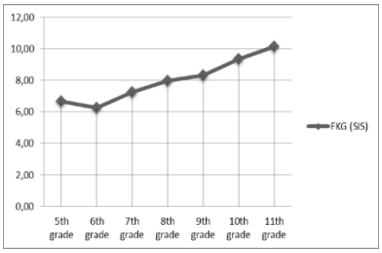

Рис. 1. Индекс Флеша-Кинкейда (SIS)

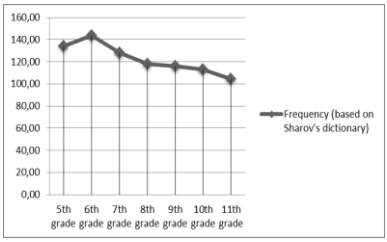

Рис. 2. Показатели частотности

Показатели частотности исследуемых текстов понижаются с повышением их уровня сложности (см. Рис. 2). Данный показатель варьируется от 143,79 для 6 класса до 104, 65 для 11 класса.

Индекс лексического разнообразия для нормализированных текстовых отрывков (1000 слов) находится в пределах от 0,61 до 0,64 (см. Рис. 3). Это указывает на то, что 61% - 64% всех используемых в тексте слов являются уникальными, что рассматривается как средние значения для подобного вида текстов [Solnyshkina, 2018].

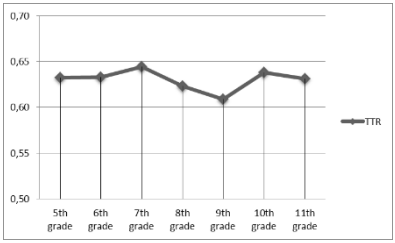

Рис. 3. TTR для нормализованных текстов

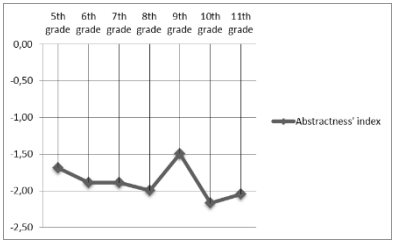

Рис. 4. Показатели индекса абстрактности

Индексы абстрактности также указывают на общее повышение сложности текстов (см. Рис. 4). Значительные флуктуации метрики объясняются тем, что абстрактность повествования не всегда достигается исключительно за счет роста доли абстрактных лексических единиц в тексте. Ее рост связан также с дистрибуцией частей речи. Согласно полученным данным, доля прилагательных в тексте варьируется между 0,128 (5 класс) и 0,180 (11 класс). При этом средняя доля наречий в тексте уменьшается и находится в пределах от 0,047 (5 класс) до 0,035 (11 класс). Похожим образом изменяется доля существительных и местоимений: она постепенно увеличивается, находясь в пределах от 0,342 (5 класс) до 0,409 (11 класс) для существительных и от 0,097 (5 класс) до 0,106 (11 класс) для местоимений соответственно. Доля глаголов уменьшается, изменяясь с 0,167 до 0,114 от 5 класса к 11, соответственно. Таким образом, абстрактность увеличивается от класса к классу не только из-за увеличения количества абстрактной лексики, но и из-за изменений в морфологической дистрибуции.

Среднее количество терминов также увеличивается с увеличением класса. При этом флуктуации графика 10 класса могут фиксировать этап повторения, заложенный авторами учебника (см. Рис. 5). Это также может быть связано с особенностями тематики данных учебников. Учебник для 9 класса фокусируется на политике, а для 11 класса – на экономике и социальной стратификации, учебник для 10 класса фокусируется об общих нормах, приятых в обществе, и деятельности человека, которые можно объяснить без использования специфичной лексики. Общее количество терминов варьируется между 243 (6 класс) и 2844 (10 класс). Среднее количество терминов увеличивается от 18 (6 класс) до 167 (11 класс).

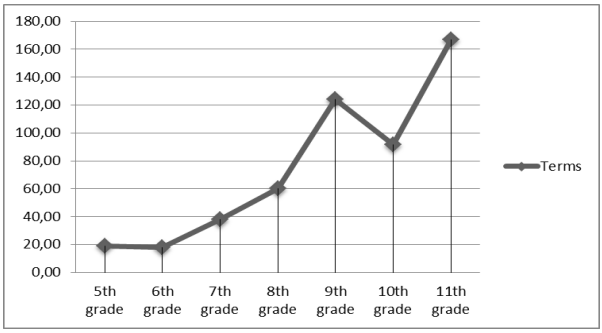

Рис. 5. Среднее количество терминов

Мы также провели анализ корреляций по Спирману, чтобы выявить статистическую зависимость указанных параметров.

Таблица 3. Показатели дисперсионного анализа (Spearman)

Параметр	**P-показатель**	**Параметр**	**P-показатель**
Кол-во слов	0,96	Сущ-ные	0,93
Кол-во слогов	0,93	Глаголы	-0,96
Кол-во предложений	0,96	Частотность	-0,96
Слова/предложения	0,57	ФК (SIS)	0,96
Слоги/слова	0,96	Абстрактность	-0,61
Прил.	0,96	TTR	-0,86
Нареч.	-0,86	TTR (1000 слов)	-0,25
Местоимен.	0,82		0,93

Сопоставление данных анализа подтвердили корреляции для следующих параметров: количество слов, количество слогов, количество предложений, среднее количество слогов в слове, доля прилагательных, доля наречий, доля местоимений, доля существительных, доля глаголов, частотность лексики, ФК и TTR. Статистически значимые параметры имеют p-показатель <0,05 (см. Табл. 3).

Комплексный анализ 14 основных параметров семи учебников по обществознанию, осуществленный при помощи онлайн-сервиса RusAC, разработанного для

обработки и оценивания текстов на русском языке, выявил статистическую зависимость сложности текста от частотности лексики, уровня лексического разнообразия, абстрактности и количества терминов. Результаты данного исследования могут быть использованы для определения соответствия учебных текстов и текстов контрольно-измерительных материалов целевой аудитории. Расширение списка параметров текста, определяющих его сложность, рассматривается как перспектива настоящего исследования.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00807. Составление корпуса было реализовано при финансовой поддержке РНФ, грант № 18-18-00436.

Литература

1. *Боголюбов Л.Н., Иванова Л.Ф., Виноградова Н.Ф., Городецкая Н.И.* и др. Обществознание. 5-11 класс: учеб. для общеобразоват. учреждений. – М.: Просвещение, 2012-2014.
2. *Колмогоров А.Н.* Три подхода к определению понятия количества информации // Проблемы передачи информации. − 1965. − Т. 1. − № 1. − С. 3-11.
3. *Оборнева И. В.* Автоматизированная оценка сложности учебных текстов на основе статистических параметров : дис. … канд. пед. наук: 13.00.02. – Москва, 2006. − 165 с.
4. *Biber D.* University Language: A corpus-based study of spoken and written registers. – John Ben. Publishing Co. Amsterdam, 2006. – 271 p.
5. *Borghi A. M. and Zarcone E.* Grounding Abstractness: Abstract Concepts and the Activation of the Mouth. Front. Psychol. 7:1498. doi: 10.3389/fpsyg.2016.01498
6. *Ivanov V.V, Solnyshkina M.I, Solovyev V.D.* Efficiency of text readability features in Russian academic texts // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. – 2018. – Iss.17. – Pp. 267-283.
7. *Mayer R.V.*, Assessing the complexity of academic text in natural sciences // Modern Education. – 2016. – № 4. – Pp. 56-64.

8. *McNamara D.S., Graesser A.C., McCarthy P.M. & Cai Z.* Automated evaluation of text and discourse with Coh-Metrix. – Cambridge, MA: Cambridge University Press, 2014. – 289 p.
9. RusAC. – URL: http://tykau.pythonanywhere.com (Дата обращения: 21.10.2020).
10. *Sharov S.A., Lyashevskaya O.N.* An introduction to frequency dictionary for Russian frequency language. M.: Azbukovik, 2009. – 21 p.
11. *Solnyshkina M., Solovyev V., Ivanov V.& Danilov A.*, Studying Text Complexity in Russian Academic Corpus with Multi-Level Annotationin // CEUR Workshop Proceedings. – Vol. 2303, International Workshop on Computational Models in Language and Speech, CMLS. – Kazan, 2019. – Pp. 1-11.
12. *Solnyshkina M.I., Harkova E.V., Kazachkova M.B.* The Structure of Cross-Linguistic Differences: Meaning and Context of 'Readability' and its Russian Equivalent 'Chitabelnost'// Journal of Language & Education. – 2020. – № 6 (1). – Pp. 103-119.
13. *Solovyev V., Andreeva M., Solnyshkina M., Zamaletdinov R., Danilov A. and Gaynutdinova D.* Computing Concreteness Ratings of Russian and English Most Frequent Words: Contrastive Approach, 12th International Conference on Developments // eSystems Engineering (DeSE). – Kazan, 2019. – Pp. 403-408.
14. *Templin M.* Certain language skills in children. –Minneapolis: University of Minnesota Press, 1957. – 208 p.
15. *Xu X., Li J.* Concreteness/abstractness ratings for two-character Chinese words // MELD-SCH. PLOS ONE. doi:10.1371/journal.pone.0232133 June 22, 202013/ 16
16. *Zinsmeister H., Birzer S., Batinić D.* LeStCor: Levelled Study Corpus of Russian. – URL: http://lestcor.org/about_the_project (Дата обращения 21.10.2020).

LEXICAL FEATURES OF TEXT COMPLEXITY: THE CASE OF RUSSIAN ACADEMIC TEXTS

The work presented in this paper is a part of an ongoing project that investigates academic text features indicative of its complexity at different grade levels. In this study we examine comparative complexity of Social science texts used in Russian

secondary and high schools. Based on the metrics of ten descriptive and four lexical features assessed for seven classroom textbooks we claim lexical diversity, frequency, abstractness and the number of terminological units to be statistically significant predictors of text complexity.

Key words: abstractness, text complexity, lexical diversity, frequency.

Л.А. ТЮКИНА
(Ярославский государственный технический университет)

В.Н. БАБАЯН
(Ярославское высшее военное училище противовоздушной обороны)

М. ЛАЗОВИЧ
(Университет Хильдесхайма, Германия)

ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ ЮМОРИСТИЧЕСКОГО ДИАЛОГИЧЕСКОГО ДИСКУРСА (НА МАТЕРИАЛЕ НЕМЕЦКО-, АНГЛО- И РУССКОЯЗЫЧНОГО БЫТОВОГО АНЕКДОТА)

Исследование затрагивает общие, типичные характеристики для анализа текста анекдота в трех лингвокультурах – немецкой, английской и русской. Исследование построено на большом корпусе текстов бытового анекдота, представляющего собой юмористический диалогический дискурс.

Ключевые слова: анекдот, Witz, joke, юмористический диалогический дискурс, анализ