Diversion of qualitative characteristics of text complexity (a case study of literary and popular science PIRLS texts)

Chulpan R. Ziganshina, and *Tatyana V*. Mazaeva

Kazan Federal University, Naberezhnochelninsky Institute of KFU

Abstract. The study under consideration highlights the multidimensional comprehensive analysis of eight texts of diverse genre attribution applied in the international PIRLS testing during 2001-2011 (PIRLS (Progress in International Reading Literacy Study) is conducted by International Association in assessment of academic achievements IEA. The national coordinator of the research implementation PIRLS in the Russian Federation is FIAQE «Federal Institute of assessment of quality education»). The objective of the study is to substantiate the hypothesis concerning similarities and differences in typological characteristics that aim at evaluating the complexity of popular science and literary texts. The cornerstone of the theory serves the assumption of the text complexity which involves quantitative (the word, sentence, text length) as well as qualitative (narrativity, syntactical simplicity, precision, referential cohesion, semantic cohesion) characteristics of the text [14]. The study determines that literary texts display a wider diversity of syntactical structures and a higher narrative degree than popular science texts under similar length and readability conditions. The precision indicators of popular science and literary texts are manifested in approximately the same range whereas referential and semantic cohesion represent a broad range of fluctuations in both cases.

1 Introduction

The complexity of the text as a scientific challenge is the study object of a number of disciplines such as Language Education, Psycholinguistics, applied Linguistics [1], which pursue their specific objectives as a part of research area. While investigating the issue the acknowledgement of three major group characteristics typical of all above-mentioned areas is mainstream: linguistic characteristics of the text, semantic (cognitive) complexity of the text and specific peculiarities of the reader [2]. The linguistic approach to the complexity text issue determines the significance of those text characteristics which distinguish it as information-semantic entity [3, p.520; 4, p. 280], preserving communicatively notional sense. It stands to reason that pragmatic goals can specify reduction or fragmentation of the communicatively notional sense of the text under study. For instance, in consultative discourse the authentic (original) text is conventionally modified [5; 6, p. 257] causing a change in a number of lexical and syntactical characteristics of the text [7]. This type of modeling focuses on the reduction of complexity of the authentic text and its ranging according to linguistic and cognitive abilities of the prospective reader.

Contemporary science acknowledges the assumption that the complexity of the text is defined by the combination of qualitative and quantitative characteristics [8, p. 165]. Quantitative characteristics entail the number of sentences, the average word length, the average sentence length, the index of lexical diversity (TTR, Type Token Ratio) and some others [9]. In terms of quantitative characteristics they calculate the readability of the text, the index, the relation of the given text to the academic level or the time span of formal education of the reader [8]. Contemporary formulas of readability such as Gunning Fox Index, Flesch Reading Ease, Flesch-Kincaid Grade that make calculations on the basis of only two characteristics –the word length and sentence length- are used in assessment of the 'readers' address' i.e. target audience [10].

Programs and services designed for the automatic processing of the natural language such as TextInspector [15] or Compleat Lexical Tutor [16] also enable to measure the readability rate of the texts. TextInspector is distinguished by a wide range of calculated text metrics, the clarity of the received data and the impressive data base supported by the Corpus of Contemporary American English (COCA) [17] and the British National Corpus (BNC) [18].

2 Materials and methods

All the readability formulas are calculated depending on typological linguistic signs [8], since typological metrics used in readability formulas differ considerably in different languages. However, there is a viewpoint that readability formulas should be genre-dependent and have amendments and stipulations depending on the particular genre text characteristics in every language. Unfortunately, the given research topic remains an open-ended problem although a vast array of scientific inquiry has been devoted to its solution [11]. The researches managed to substantiate that the text complexity is determined to a great extent not so much by qualitative text differences as by quantitative text differences such as text cohesion, its precision, narrativity, etc. [11].

In this connection it seems rational and topical to expose and contrast qualitative characteristics of the complexity assessment of multi-genre texts and verify the hypothesis that texts of different genres have different qualitative characteristics. The case study of the work is represented by eight English texts borrowed from the tests on Progress in International Reading Literacy Study (PIRLS) in 2001, 2006, 2011 and 2016 (http://www.pirls.org/). The texts were classified into two groups: literary (The Upside-Down Mice (2001), An Unbelievable Night (2006), Enemy Pie (2011), Macy And The Red Hen (2016) and popular-science (Nights Of The Pufflings (2001), Day Hiking (2006), The Giant Tooth Mystery (2011), The Green Sea Turtle's Journey Of A Lifetime (2016)) (Henceforward the texts will be marked by a corresponding code: The Upside-Down Mice (2001) – UDM1, An Unbelievable Night (2006) – UN6, Enemy Pie (2011) – EP11, Macy And The Red Hen (2016) – MRH16, Nights Of The Pufflings (2001) – NOP1, Day Hiking (2006) – DH6, The Giant Tooth Mystery (2011) – GTM11, The Green Sea Turtle's Journey Of A Lifetime (2016) – GST16).

3 Results and discussion

The quantitative characteristics of the text under study were calculated with the help of the automatic analysis program Text Inspector [16] (see Table 1)

Table 1. Quantitative characteristics of PIRLS texts

Year	2001	2006	2011	2016	2001	2006	2011	2016
Text type	Literary				Popular science			
Text	UDM1	UN6	EP11	MRH16	NOP1	DH6	GTM11	GST16
Text length (sentences)	44	58	89	65	52	60	71	78
Text length (words)	531	837	780	913	686	667	850	934
Average sentence length (in words)	12.07	14.43	8.76	14.05	13.19	11.12	11.97	11.97
Index of lexical diversity (TTR)	0.42	0.38	0.34	0.36	0.39	0.43	0.33	0.39
Flesch-Kincaid Grade	4.07	6.36	2.70	4.43	4.98	4.60	5.46	4.19

The quantitative data enable to make conclusions that within the testing period from 2001 to 2011 both literary and popular science texts have a tendency towards volume gain which is established by two parameters: the total number of sentences and the number of words. The sentence length of literary texts is maintained in the range of 8.76 to 14.43, while for popular science texts the upper and the lower levels of this parameter are substantially drawn close to each other – from the minimum of 11.12 in the text Day Hiking (2006) to 13.19 in the text Nights of the Pufflings (2001). Flesch-Kincaid Grade for literary texts fluctuates within the range of 2.7 to 6.36, whereas the range for popular science texts is maintained within the limits a little more than 1.20: from 4.19 to 5.46. As Flesch-Kincaid Grade corresponds to the number of years of formal education, it's obvious that popular science texts are more suitable for the fifth graders' age group of readers.

The additional set of parameters, loosely named as qualitative is calculated by online service Coh-Metrix [12] which defines every parameter point on a scale of 0 to 100. (See Table 2)

Table 2. Qualitative characteristics of PIRLS texts

Year	2001	2006	2011	2016	2001	2006	2011	2016
Text type	Literary				Popular science			
Text	UDM1	UN6	EP11	MRH16	NOP1	DH6	GTM11	GST16
Narrativity (%)	**90**	73	96	77	**50**	54	56	56
Syntactical simplicity (%)	**59**	64	96	75	**73**	96	87	75
Word precision (%)	**65**	88	67	97	**99**	72	71	98
Referential cohesion (%)	**23**	39	46	59	**42**	14	55	41
Semantic cohesion (%)	**39**	25	86	51	**84**	81	50	29

Let's consider each of the parameters calculated with the help of Coh-Metrix using the example of two texts: the literary text "The Upside-Down Mice" and the popular science text , «Nights Of The Pufflings» (text metrics are highlighted) (See fig.1)

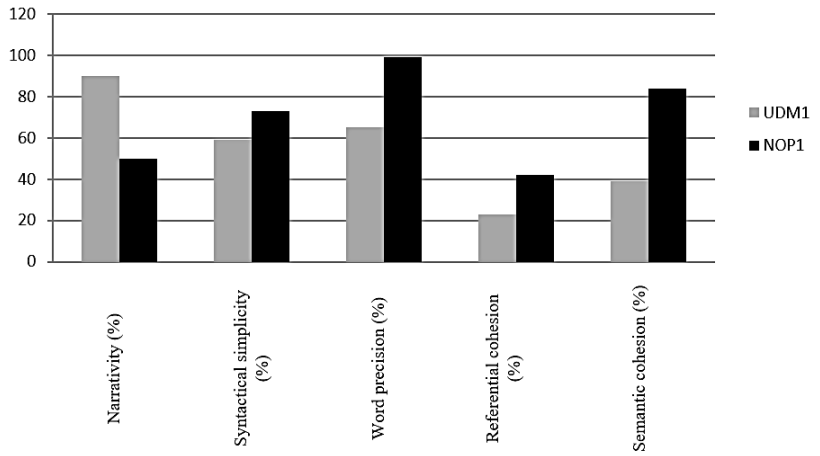

Fig. 1. Qualitative characteristics of the text complexity.

The 'narrativity' index represents the occurrence of the plot characters, events, space-time in the text. Narrativity as a rule suggests the occurrence of colloquial well-used vocabulary, which is familiar to the reader [13]. This parameter is considered to be a reliable indicator of knowing the vocabulary and the world.

Non-narrative texts touch upon less familiar topics, they are traditionally on the opposite end of the continuum. As we can see in Table 2 and Fig.1 the narrativity of the literary texts considerably exceeds the index of popular science texts. Let's compare as an example two extracts from the literary and popular science texts: «*That night when the mice came out of their holes and saw the mousetraps on the ceiling, they thought it was a tremendous joke. They walked around on the floor, nudging each other and pointing up with their front paws and roaring with laughter*» (The Upside-Down Mice, 2001) (Underlining marks verbs and non-finite verb forms). «*Every year, black and white birds with orange bills visit the Icelandic island of Heimaey. These birds are called puffins. They are known as "clowns of the sea" because of their bright bills and clumsy movements. Puffins are awkward fliers during takeoffs and landings because they have chunky bodies and short wings*» (Nights of the Pufflings, 2001).

Syntactical simplicity embraces a cluster of characteristics such as the number of words in a sentence, the number of words before the predicate and the number of mainstream syntactical constructions familiar to the reader [13]. The higher this characteristics is the simpler the text: the syntax of the text Nights of the Pufflings (73) is less complicated for readers' comprehension than the text The Upside-Down Mice (59). The illustrations given above demonstrate a significantly smaller gamut of syntactical constructions and fewer words from the beginning of the sentence to the predicate in the popular science text.

High precision rate implies that the text contains linguistic units that are able to conjure up mental images with facility, i.e the images which are easy to process and comprehend. Abstract words represent notions which are hard to visualize therefore they appear to present challenges to the readers. The precision of the popular science text under consideration is extremely high – NOPI – 99%, while in the literary text UDM1 betrays a smaller precision ratio – 65%. However generally precision ranges of the literary and popular science texts under study do not differ notably: 65-97 for literary texts versus 71-99 for popular science texts.

Referential cohesion is realized through lexical repetition (reiteration) and replacement of items, the latter being represented mostly by pronouns. The text with high referential cohesion contains words and ideas which are repeated in the sentences as well as in the whole

text, building clear-cut semantic chains that make the text coherent to the reader. The extract of the text UDM1 includes two semantic chains that encompass the following linguistic units: the mice, they, they, each other, their; (2) the mousetraps on the ceiling, it.

The extract of the popular science text NOP1 has a high referential cohesion of the semantic chain "puffins" that includes nine speech units: birds, these birds, puffins, they, clowns of the sea, their, puffins, filers, they. The data indicate that the parameter range of referential cohesion of different texts is wide enough: from 14 to 55 for popular science texts and 36 for literary texts. It is possible to make an assumption that the cohesion degree depends to a large extent on the writer's style, yet it can be deliberately increased to decrease the text complexity degree.

The deep semantic cohesion of the text is expressed through lexical means that establish causal-resultative, spatial, temporal and other types of relations [13]. The text with low degree of cohesion is more challenging for comprehension, since it lacks the part of logical connections, which the reader has to restore. Topic and comment can present a specific issue when the succession of introduction of new information is broken [13]. For example, the extract of the English text NOP1 can have two possible variants of translation of the third sentence into Russian: (1) «Из-за ярких клювов и неуклюжих движений их ещё называют «клоунами моря». (2) «Их ещё называют «клоунами моря» из-за ярких клювов и неуклюжих движений». The second variant should obviously be accepted as a preferable one as it introduces the comment (They are known as "clowns of the sea") after the repetition of the topic (because of their bright bills). Similarly to referential cohesion, the indices of deep semantic cohesion of the texts under consideration do not demonstrate apparent dependence and are allegedly determined primarily by the writer's style rather than the genre of the text.

4 Conclusion

Thus we have every reason to reach a conclusion that the material under study displays palpable difference in qualitative characteristics of the texts belonging to different genres. Having approximately similar lexical diversity and readability literary texts represent a higher degree of narrativity. Popular science texts are distinguished by a higher degree of syntactical simplicity in contrast to literary texts. Their precision is maintained almost at the same level. As far as referential and deep semantic cohesion are concerned, the texts under consideration do not exhibit a clear-cut tendency. It should be admitted that the conducted research doesn't settle all the complexity of the problem of the differences in qualitative characteristics that influence the complexity of literary and popular science texts. The perspectives of the research appear to be the increase in the volume of the material as well as extension of genre scope of the investigated texts of different complexity degree.

Acknowledgements

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. M. I. Solnyshkina, E.V. Harkova, M.B. Kazachkova, Journal of Language & Education, **6(1)**, 103 (2020)
2. V. Solovyev, M. Solnyshkina, V. Ivanov, I. Batyrshin, Journal of Intelligent and Fuzzy Systems, **36(5)**, 4553 (2019)

3. N.S. Bolotnova, *Filologicheskiy analiz teksta: ucheb. posobie* (2009)
4. N.S. Valgina, *Text theory* (2003)
5. A. A. Sabinina, News RGPU im. A. I. Gertzena, 97, 222 (2009)
6. V.N. Semerdzhidi, *Semiotic regularities of the functioning of the phenomena of paralinguistics in didactic texts: on the material of the Russian and English languages* (2008)
7. M.I. Solnyshkina, R.R. Zamaletdinov, L.A. Gorodetskaya, A.I.Gabitov, Journal of Social Studies Education Research, **8(3)**, 238 (2017)
8. I. V. Oborneva, *Automated assessment of the complexity of educational texts based on statistical parameters* (2006)
9. M. Templin, *Certain language skills in children* (1957)
10. Marina Solnyshkina, Valery Solovyev, Vladimir Ivanov, and Andrey Danilov, CMLS 2019, Kazan, Russian Federation, 2303, 1 (2019)
11. V. Ivanov, M. Solnyshkina, V. Solovyev, Computational linguistics and intellectual technologies: Based on the materials of the annual international conference "Dialogue", 17(24), 276 (2018)
12. Coh-Metrix, http://cohmetrix.com/
13. M.I. Solnyshkina, A.S. Kiselnikov, Bulletin of the Volgograd State University, Series 2, Linguistics, **1(25)**, 99 (2015)
14. Danielle McNamara, Arthur C. Graesser, Max M Louwerse, Zhiqiang Cai, Journal of the Psychonomic Society, **36(2)**, 193 (2004)
15. TextInspector, https://textinspector.com/
16. Compleat Lexical Tutor, https://www.lextutor.ca/
17. Corpus of Contemporary American English (COCA), https://www.english-corpora.org/coca/
18. British National Corpus (BNC), https://www.english-corpora.org/bnc/