Evaluating Text Complexity and Flesch-Kincaid Grade Level

Marina I. Solnyshkina [1], Radif R. Zamaletdinov [2], Ludmila A. Gorodetskaya [3] & Azat I.

Gabitov [4]

Abstract

The article presents the results of an exploratory study of the use of T.E.R.A., an automated tool measuring text complexity and readability based on the assessment of five text complexity parameters: narrativity, syntactic simplicity, word concreteness, referential cohesion and deep cohesion. Aimed at finding ways to utilize T.E.R.A. for selecting texts with specific parameters we selected eight academic texts with similar Flesch-Kincaid Grade levels and contrasted their complexity parameters scores to find how specific parameters correlate with each other. In this article we demonstrate the correlations between text narrativity and word concreteness, abstractness of the studied texts and Flesch – Kincaid Grade Level. We also confirm that cohesion components do not correlate with Flesch –Kincaid Grade Level. The findings indicate that text parameters utilized in T.E.R.A. contribute to better prediction of text characteristics than traditional readability formulas. The correlations between the text complexity parameters values identified are viewed as beneficial for developing a comprehensive approach to selection of academic texts for a specific target audience.

Keywords: *Text complexity, T.E.R.A., Syntactic simplicity, Narrativity, Readability, Texts analysis.*

Introduction

The modern linguistic paradigm comprising achievements of "psycholinguistics, discourse processes, and cognitive science" (Danielle et al., 2011) provides both a theoretical foundation, empirical evidence, well-described practices and automated tools to scale texts on multiple levels including characteristics of words, syntax, referential cohesion, and deep cohesion. The scope of applications of such tools is enormous: from teaching practices to cognitive theories of reading and comprehension. One of the tools, T.E.R.A., Coh-Metrix Common Core Text Ease and Readability Assessor, an automated text processor developed in early 2010s by a group of American scholars of The Science of Learning and Educational

---

[1] Prof, Kazan Federal University - Kazan, mesoln@yandex.ru
[2] Prof, Kazan Federal University - Kazan, director.ifmk@gmail.com
[3] Prof, Lomonosov Moscow State University - Moscow, lgorodet@gmail.com
[4] Pst. Grad, Kazan Federal University - Kazan, gabit.azat@gmail.com

Technology (SoLET) Lab, directed by Dr. Danielle S. McNamara, has already been successfully applied in two Russian case studies conducted by A.S. Kiselnikov (Solnyshkina, Harkova & Kiselnikov, 2014). As the research shows, it is by all means under-used in modern Russian academic practices in general and in the area of teaching English as a foreign language in particular. Addressing the gap, we demonstrate how T.E.R.A. can be applied in academic practices and how a limited number of text parameters in all their varieties, are significant in selecting texts for academic purposes.

Methods

The data for the study were collected from "Review" Chapters marked A in Spotlight 11 approved by the Ministry of Education of the Russian Federation and recommended for English language teaching in the 11th grade of Russian public schools. All the texts compiled in the corpus are the texts used to test students' reading skills in the classrooms. Their length varies from 323 words in Text 3A to 494 words in Text 7A with the mean of 395 words. The readability of the texts selected fall into the scope of the target audience, i.e. Russian high school graduates, and vary between indices 8 and 9 of Flesch-Kincaid Grade Levels (see Table 1 below). We measured the complexity parameters of the 8 selected texts with the help of T.E.R.A. and consecutively contrasted two texts with the highest and lowest scores of each complexity parameter to identify the correlation between a particular index and Flesch-Kincaid Grade Level.

Except for the Flesch-Kincaid Grade Level, T.E.R.A. available on the public website computes five complexity parameters of texts, i.e. syntactic simplicity, abstractness/concreteness of words, narrativity, referential cohesion, deep cohesion. Thus, T.E.R.A. provides detailed information of how logically connected the text is, what functions make the texts more or less grammatically cohesive, what are the dependencies between one part of the text and another for each analyzed text, the program assigns definite values thus indicating the position of a particular text among other texts assessed and stored in the database (T.E.R.A. Coh-Metrix Common Core Text Ease and Readability Assessor). A user can view texts and their complexity indices in T.E.R.A online library.

Table 1

Complexity Parameters of Texts 1 A - 8 A

Text	Narrativity	Syntactic simplicity	Abstractness/ Concreteness	Referential Cohesion	Deep Cohesion	Flesh – Kincaid Grade Level
1A	79%	34%	36%	39%	81%	8,20
2A	77%	65%	39%	37%	99%	7,40
3A	92%	54%	70%	40%	74%	6,50
4A	69%	65%	73%	24%	94%	7,30
5A	80%	55%	78%	13%	94%	6,20
6A	75%	51%	14%	20%	94%	9,70
7A	84%	63%	33%	9%	95%	7,50
8A	30%	36%	80%	22%	42%	9,50

According to McNamara and Graesser (2012) *narrativity* depends on the mean of verbs per phrase, presence of common words and overall story-like structure. To ensure high readability of a text, researchers recommend to use a large number of dynamic verbs in a relatively small variety of time forms, which makes the sentences syntax similar and reduces the number of words in front of the main verb. In texts with a high narrative value, fewer unique nouns and more pronouns create similar combinations of sentences. T.E.R.A. assesses

Syntactic simplicity of a text is measured based on three measured parameters, i.e. the average number of clauses in sentences throughout the text, the number of words in the sentence, and the number of words in front of the main verb of the main sentence (McNamara & Graesser, 2012). Texts with lower number of clauses, fewer words per sentence and fewer words before the main verb will have a higher syntactic simplicity value. The correlation of the parameter with the above mentioned indices was conveniently verified in the research pursued by a group of Russian scholars on the materials of Unified State Exam in English (EGE), which is a matriculation exam in the educational system of the Russian Federation (Solnyshkina, Harkova & Kiselnikov, 2014).

Abstractness/concreteness of words as it comes from the name, shows the proportion of concrete words to abstract ones in a text (McNamara & Graesser, 2012; Waters & Russell, 2016). Assessing a text abstractness/concreteness T.E.R.A does not provide any instrument to verify abstractness/concreteness of separate words. However, its developers refer potential inquirers to the Medical Research Council (MRC) Psycholinguistic Database, containing

150,837 words with 26 specific linguistic and psycholinguistic attributes (Brysbaert, Warriner & Kuperman, 2014; MRC Psycholinguistic Database; Erbilgin, 2017; Tarman & Baytak, 2012). The scores are derived based on human judgments of word properties such as concreteness, familiarity, imageability, meaningfulness, and age of acquisition (MRC Psycholinguistic Database). The resource acquires a word a rank in the list of 'less' or 'more' concrete/abstract words. As the tool assesses the word family tokens only and neglects the context of a word, MRC Psycholinguistic Database, as it is admitted by the developers and researchers 'is not without limitations" (McNamara & Graesser, 2012).

Referential cohesion is a measure of the overlap between words in the text, formed with the help of similar words and ideas transmitted by them (McCarthy et al., 2006). When sentences and paragraphs have similar words or ideas, it is easier for the reader to establish logical connections between them. If a text is cohesive its ideas overlap thus providing a reader with explicit threads connecting parts of the text. In adjacent sentences the threads are manifested by co-referencing words, anaphora, similar morphological roots, etc. For example in Text 1A we find repetitions of the word *child*, semantic overlap in the words *country – China, child – family, an only child – one child*: "I am an only child because, in 1979, the government in my country introduced a one-child-per-family policy to control China's population explosion" (Text 1A).

Deep cohesion reflects the degree of logical connectives between sentences, but in this case it is revealed by measuring different types of words that connect parts of the text (McNamara & Graesser, 2012). There are different types of connectives: temporal, causal, additive, logical. Examples of these words are *after, before, during, later, additionally, moreover, or*. These elements of the text help to link together events, ideas and information of the text, forming the reader's perception. For example: "The good news, *however*, is that you CAN deal with stress *before* it gets out of hand! *So*, take control and REMEMBER YOUR A-B-Cs." (Text 2A).

We also utilized an online tool Text Inspector to measure lexical diversity of every text studied. Lexical Diversity is viewed by the authors as "the range of different words used in a text" (McCarthy & Jarvis, 2010). Text Inspector assesses VOCD (or HD-D) and MTLD. As the texts in the corpus studied are of about the same length, i.e. about 400 words their lexical diversity metrics are viewed as reliable, not sensitive to the length of the texts studied. The Lexical Diversity tool used by Text Inspector is "based on the Perl modules for measuring

MTLD and voc-d developed by Aris Xanthos" (Text Inspector). "MTLD is performed two times, once in left-to-right text order and once in right-to-left text order. Each pass yields a weighted average (and variance), and the two averages are in turned averaged to get the value that is finally reported (the two variances are also averaged). This attribute indicates whether the reported average should itself be weighted according to the potentially different number of observations in the two passes (value 'within_and_between'), or not (value 'within_only')" VOCD method implies random selection of "35, 36, …, 49, and 50 tokens from the data, then computing the average type-token ratio for each of these lengths, and finding the curve that best fits the type-token ratio curve just produced <…>. The parameter value corresponding to the best-fitting curve is reported as the result of diversity measurement. The whole procedure can be repeated several times and averaged" (Text Inspector). Lexical Diversity of Text 6A (393 words) computed with Text Inspector is 134.75 (VOCD), 116.61 (MTLD) which is viewed as relatively high (Text Inspector).

Results

To determine the impact of each of the parameters computed by T.E.R.A. on the Flesch-Kincaid Grade Level and identify correlations between variables of Coh-Metrix, we measured texts indices of 8 texts from Spotlight 11 (2009) and contrasted vocabulary and grammar of the texts with minimum and maximum values of narrativity, syntactic simplicity, word concreteness, referential and deep cohesion. The results of T.E.R.A. processing are presented in Table 1.

It was decided to exclude Text 8 from further analysis based on the assumption that as its narrativity score twice as low as those of the other texts (30% vs 69% - 92%) and it may lead to a considerable bias in the research outcomes. Text 8A portrays four sights and is mostly descriptive. Consider an excerpt from Text 8A: *Otherwise known as The Lost City of the Incas', Machu Picchu is an ancient Incan city located almost 2,500 metres above sea level in the Andes Mountains in Peru. Machu Picchu is invisible from below* (Spotlight 11, Text 8A). As it is shown in the example above, the author uses mostly stative verbs (know, be, etc.) in contrast to Text 3A with the highest narrativity index in the corpus of the texts studied, i.e. 92%, in which the verbs used are mostly dynamic: *arrived, gone, checking, had taken, reported, caught.* The sentences are short and easy to understand: *Burglars recently broke into our house while we were sleeping upstairs! My sister and I heard a noise, so we woke up our dad, who called the police* (Spotlight 11, Text 3A).The genre also reflects on Concreteness/Abstractness and Deep Cohesion indices:

all narrative texts prove to be more concrete and cohesive than the contrasted descriptive text. Both Deep cohesion (42%) and Referential Cohesion (22%) indices of Text 8A are significantly lower than the corresponding parameters of all other texts (Table 1 above).

T.E.R.A. also discriminated the texts which were otherwise similar but had different scores on *Syntactic Simplicity*. As we see in Table 1, Syntactic Simplicity in Texts 1A and Text 2A differ significantly with 34% and 65% respectively. The corresponding Flesh – Kincaid Grade levels differ in 1.2., Deep Cohesion – 17%, while the rest of the parameters are only 2 % - 3% different. Text 1 A presenting the theme "family" serves a good example of low Syntactic simplicity score. It contains simple syntactic structures, 27 sentences of which are in the Present Simple tense, there are no participial or gerundial constructions either. Its lexical diversity is only 91.66 (VOCD), 84.80 (MTLD). All these make the text less challenging to process by the reader than Text 2A which is at the opposite end of the continuum: with 30 infinitives, 10 gerundial constructions, 7 verbs in the Present Simple tense, five past participles. Cf. *"In stressful situations, the nervous system causes muscles to tense, breathing to become shallow and adrenaline to be released into your bloodstream as your body gets ready to beat challenges with focus and strength"* (Spotlight 11, Text 2A). The lexical diversity of Text 2A is also much higher than in Text 1A: 101.26 (VOCD), 100.80 (58 LD). Thus, we may provisionally conclude that Syntactic simplicity does not much correlate with Flesh – Kincaid Grade Level.

The texts chosen for the contrastive analysis of *Word Concreteness* are Texts 5A and Text 6A with Flesh – Kincaid Grade Levels of 6.20 and 9.70, respectively. These two texts have radically different Flash-Kincaid Grade levels (3.5 grade difference), but similar scores of Narrativity, Syntactic Simplicity, Deep Cohesion. However, the critical difference lies in the Concreteness/ Abstractness of the words with values of 78% and 14% for Texts 5A and Text 6A, respectively. Low word concreteness value indicates the presence of a large number of abstract words in Text 6A. As the theme of Text 6A is the study of alien activities, it contains specific vocabulary: *civilization, intelligent life, signal, screensaver,* etc. The vocabulary of Text 5A, which portrays life of homeless people, consist of predominantly concrete nouns: *benches, doorways, houses*, *hostel, room, streets* etc. Thus, it is obvious that it is mostly Concreteness of Text 5A that decreases its Flesh – Kincaid Grade Level.

Referential Cohesion demonstrates a spike with 40% in Text 3A and falls to 9% in Text 7A. Indices of Narrativity and Syntactic simplicity fluctuate within a narrow range of 8 - 9%,

while Concreteness/Abstractness is distinctively diverse with 70% in Text 3A and 33 % in Text 7A. The statistics also shows little relation between Flesh – Kincaid Grade Level and Referential Cohesion (see Table 1 above).

As lexical diversity is proved to be in inverse proportion to cohesion (McNamara & Graesser, 2012), we also computed Lexical diversity of Texts 7A and 3A. Text Inspector measures lexical diversity of Text 7A to be 145.56 (VOCD), while that of Text 3A to be only 92.48 (VOCD). Based on the scores we can assume that Text 3A contains more words and ideas that overlap across adjacent sentences and the entire text, while Text 7A contains fewer explicit threads that connect the text for the reader. Cf.: *"Fortunately, I was able to identify the mugger from a photo at the police station. He was a well-known criminal in the area, so the police knew where to find him. Anyway, he confessed to the crime, the police arrested him"* (Text 3A). As we can see the connections between the ideas are made with the help of thematic similarity (*the mugger – a criminal – a crime – arrested*), repetition (*the police*), substitution (*the mugger – he – him – he – him*), derivatives (*criminal – crime*). Referential cohesion for Text 7A is low due to the lack of lexical and semantic overlap. Cf.: *"Believe you can climb that mountain, swim that ocean or reach that place, and surely one day you will. There would be no Ford cars, Star Wars, light bulbs or Beethoven symphonies if this was not true!"* (Text 3A). Thus, Text 7A is more challenging for the reader, especially for a non-native speaker. The counterbalance which levels up Flesh – Kincaid Grade Levels of the Texts 3A and Text 7A is Word Concreteness which is much higher in Text 3A (see the Table above).

The texts demonstrating distinctively different *Deep Cohesion* are Texts 2A and Text 8A, which judged from the statistics in Table 1, are also different in the following characteristics: narrativity, syntactic simplicity, word concreteness and referential cohesion. *Deep Cohesion* of Text 2A is extremely high, 99% , which means that the text connections are very dense. It contains 17 temporal connectives, 3 causal, 7 intentional, while Text 8A incorporates 3 temporal connectives, 2 causal, 0 intentional connectives (Gabitov & Ilyasova, 2016). At this stage of the research it is difficult to explain all the correlations between the parameters but the fact that deep cohesion has very little correlation with Flesh – Kincaid Grade Level is obvious.

Discussion

The analysis has showna wide range of possibilities which T.E.R.A. provides for assessing text complexity parameters and their interrelations. By assessing complexity

parameters it discriminated Text 8A from the rest of the texts studied as a text of different genre: as a descriptive text Text 8A demonstrated much lower narrativity score than all the other in the continuum.The question of this text appropriateness as the final reading text in the textbook, though being urgent, is beyond the scope of this paper.

T.E.R.A. also assesses text syntactic simplicity thus providing a user with an instrument to measure three different syntactic indices: the number of clauses, the number of words in a sentence, the number of words before the main verb. The results of this study confirm that syntactic simplicity measured with T.E.R.A. does not much correlate with Flesch - Kincaid Grade Level. However the research demonstrated strong correlation between text concreteness computed with T.E.R.A. and Flesch - Kincaid Grade Level: with all other complexity parameters of two texts being similar, it is word concreteness that shapes the grade level score. As for referential cohesion and deep cohesion scores assessed with T.E.R.A., they go beyond traditional readability formulas, including Flesch - Kincaid Grade Level, i.e. do not correlate with the latter. Two other phenomena discovered are the following: the score of Referential cohesion of all narrative texts in the corpus is below 40% with the mean being 26%, while the Deep cohesion score is above 74% with the mean of 90%.

The complexity parameters measured with T.E.R.A. and the elicited interdependences between the latter and Flesch-Kincaid Grade level provide a good foundation for educators to elaborate an extensive approach to selection of reading texts for academic purposes of different groups of students (Readability Formulas). Several authors have proposed different metric sets to assess similarity and dissimilarity in text complexity, such as adjective per sentence, nouns per sentence, frequency of content words, etc. that can successfully rank academic texts for different age and grade levels (Solovyev, Ivanov & Solnyshkina, 2017).

Conclusion

T.E.R.A. analyses of the text complexity values demonstrated that (1) Narrativity of the texts studied tends to be in inverse ratio to deep cohesion and directly proportional to word concreteness. (2) Concreteness of the studied texts displays strong correlation with Flesch – Kincaid Grade Level and potential to decrease the latter. (3) Syntactic simplicity does not demonstrate much interdependence with Flesch – Kincaid Grade Level. (4) The cohesion components, i.e. referential cohesion and deep cohesion indices, do not correlate with Flesch –

Kincaid Grade Level. The identified correlations between the text parameters values computed by T.E.R.A. are viewed by the authors as beneficial for designing an algorithm to select and modify texts so that they correspond to the cognitive and linguistic level of the target readers.

References

Brysbaert, M., Warriner, A.B. & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46: 904. https://doi.org/10.3758/s13428-013-0403-5.

Danielle, S. McNamara, D.S., Graesser, A.C., Cai, Z. & Kulikowich, J.M. (2011). *Coh-Metrix Easability Components: Aligning Text Difficulty with Theories of Text Comprehension. AERA*. Retrieved from https://www.researchgate.net/publication/228455723_Coh-Metrix_Easability_Components_Aligning_Text_Difficulty_with_Theories_of_Text_Comprehension.

Erbilgin, E. (2017). A comparison of the mathematical processes embedded in the content standards of Turkey and Singapore. *Research in Social Sciences and Technology*, 2(1): 53-74.

Gabitov, A.I. & Ilyasova, L.G. (2016). Use of automated instruments of text analysis to provide proper difficulty level of English language educational materials. *Problems of Modern Pedagogical Education: Pedagogy and Psychology*, 53(3): 101-108.

McCarthy, P.M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42: 381. https://doi.org/10.3758/BRM.42.2.381

McCarthy, Ph.M., Lightman, E.J., Dufty, D.F. & McNamara, D.S. (2006). Using Coh-Metrix to assess distributions of cohesion and difficulty: An investigation of the structure of high-school textbooks. In: *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (190-195). Mahwah: Eribaum.

McNamara, D.S. & Graesser, A.C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In: *Applied natural language processing and content analysis: Identification, investigation, and resolution* (188-205). Hershey, PA: IGI Global.

MRC Psycholinguistic Database. Retrieved from http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm.

Readability Formulas. Free readability tools to check for Reading Levels, Reading Assessment, and Reading Grade Levels. Retrieved from http://www.readabilityformulas.com.

Solnyshkina, M.I., Harkova, E.V. & Kiselnikov, A.S. (2014). Comparative Coh-Metrix Analysis of Reading Comprehension Texts: Unified (Russian) State Exam in English vs Cambridge

First Certificate in English. *English Language Teaching*, 7(12): 65-76.

Solovyev, V., Ivanov, V. & Solnyshkina, M. (2017). *Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics*. In Press.

Tarman, B. & Baytak, A. (2012). "Children's online language learning: A constructionist perspective", Energy Education Science and Technology Part B: Social and Educational Studies, 2012, 4(2) 875-882.

T.*E.R.A. Coh-Metrix Common Core Text Ease and Readability Assessor*. Retrieved from http://129.219.222.66:8084/Coh-Metrix.aspx.

Text Inspector. Retrieved from http://textinspector.com/workflow/B3021C1A-706A-11E7-B233-AB44AFCE53D3.

Waters, S. & Russell, W.B. (2016). Virtually Ready? Pre-service teachers' perceptions of a virtual internship experience. *Research in Social Sciences and Technology*, 1(1): 1-23.