



Computational Linguistics & Reading Comprehension Text Complexity: Implications for the EFL Teacher

Trisevgeni Liontou

Greek Ministry of Education (Greece)

tliontou@enl.uoa.gr

Abstract

This paper reports on a 1-year longitudinal study that aimed at investigating the relationship between intermediate EFL learners' development of reading and writing proficiency while delineating a range of linguistic features present in a set of 30 reading comprehension texts and 350 written essays. Data (reading comprehension exam scores and written essays) were collected from a junior high school in Athens, Greece and 70 EFL students aged 14-15 years old took part in the study. By making use of advanced Computational Linguistics and Automated Machine Learning systems an attempt has been made to find the relationship, if any, between a set of 135 text variables, students' L2 lexical growth and the readability level of pertinent texts. The rationale of this study is closely linked to Weir (2005: 292) and Alderson et al. (2004: 11) words of caution that "although the CEFR attempted to describe language proficiency through a group of scales composed of ascending level descriptors, it does not contain any guidance, even at a general level, of what might be simple in terms of structures, lexis or any other linguistic level". In fact, the argument that the CEFR is intended to be applicable to a wide range of different languages offers little comfort to the EFL teacher, who has to select texts or activities uncertain as to the lexical breadth of knowledge required at a particular level within the CEFR. The findings of the present research provide practical guidance to EFL teachers, test designers and material developers as to the range of lexicogrammatical features an EFL learner of an expected level of language ability might be able to handle for a successful reading comprehension performance along with the linguistic strategies s/he might have developed for the production of more elaborate written texts.

1. Introduction

Analyzing writing development as a function of grade level is important in elementary and middle school children because the developmental patterns are strongest at a young age and the opportunity to develop successful interventions is most likely [4,13,17]. Along with writing development, Carrell (1987: 21) emphasized the need for EFL reading teachers and materials developers to establish reliable ways of matching the difficulty of reading materials to foreign language readers, since, if materials are too easy, students are unchallenged and bored, and no learning occurs. On the other hand, if materials are too difficult, students become frustrated and withdrawn, and again no learning occurs (ibid: 21). Motivated by the above literature, this paper reports on a one-year longitudinal study that aimed at investigating the relationship between intermediate EFL learners' development of reading and writing competence while delineating a range of linguistic features present in a set of 30 reading comprehension and 350 written essays.

2. Literature Review

The current study is closely linked to earlier findings of research on reading comprehension performance, according to which many variables such as text topic and structure, text language, background knowledge and task type, can have an impact on either the reading process or product and need to be taken into account during test design and validation [16]. In addition to text readability, a common approach to assessing writing quality is through the analysis of linguistic features that characterize proficient writing [13]. Assessing linguistic features allows researchers to make links between text properties important in writing quality such as cohesion [8], elaboration [13], abstractness [10], sophistication [13], and diversity of ideas [12, 13, 19]. Despite the considerable advances that have been made in exploring and understanding the various aspects of foreign language acquisition and reading performance, the available research has been rather unsuccessful in clearly defining and, most importantly, in prioritizing those text features that have a direct impact on text complexity and need to be accounted for during the text selection and item design process. Although readability formulas have been extensively applied in the field of foreign language teaching and testing, numerous researchers have pointed to their serious limitations and repeatedly stressed the need for a more in-depth analysis of text features, in order to better define what sort of text a learner of a given level of language ability should be expected to be able to process. Alderson et al. (2004: 11) and Weir (2005:



292) also stressed that many of the terms in the CEFR remain undefined and argued that difficulties arise in interpreting it because "it does not contain any guidance, even at a general level, of what might be simple in terms of structures, lexis or any other linguistic level". Therefore, according to Alderson *et al.*, the CEFR would need to be supplemented with lists of grammatical structures and specific lexical items for each language teachers and item writers to make more use of it. By making extensive use of advanced Computational Linguistics and Machine Learning systems that only recently became available, the current research has, thus, been designed to fill this void and further add to our present state of knowledge on EFL language acquisition by investigating the relationship between intermediate EFL learners' development of reading and writing proficiency while delineating a range of linguistic features present in a set of 30 reading comprehension texts and 350 written essays. In order to explore these issues, the following research question was formed:

Research Question 1: Is there a significant relationship between intermediate EFL students' written production quality and reading comprehension performance scores?

3. Methodology

In order to investigate the relationship between intermediate EFL learners' development of reading and writing proficiency a range of linguistic features present in a set of 30 reading comprehension texts and 350 written essays were analysed. Data (reading comprehension exam scores and written essays) were collected from a junior high school in Athens, Greece and 70 EFL students aged 14-15 years old took part in the one-year longitudinal study. The corpus of 350 essays consisted of 5 written essays per student, whereas participants also processed 30 multiple-choice Reading Texts with 5 questions per text which resulted in a total of 150 comprehension questions per student and 10,500 reading comprehension answers for the whole set. Advances in Computational Linguistics and Machine Learning systems have made it possible to go beyond surface text components, focusing on a wider range of "deep" text features that take into account semantic interpretation and the construction of mental models and can, thus, offer a principled means for test providers and test-takers alike to assess this aspect of test construct validity [7]. In the present study *Coh-Metrix 3.0*, *Linguistic Inquiry and Word Count 2007* (LIWC), *VocabProfile 3.0*, *Computerized Language Analysis* (CLAN) suite of programs, *Computerized Propositional Idea Density Rater 3.0* (CPIDR) and *Gramulator* were used to estimate the 135 text variables.

4. Discussion

An important part of the present research involved the investigation of relationships between mean reading comprehension scores per text (mean scores of the 5 reading comprehension multiple-choice questions pertinent to each text based on the performance of the 70 intermediate EFL language learners that had taken part in the research) and textual features present in the 30 intermediate reading comprehension texts. In order to determine the contribution of the 135 text indices to mean reading performance per text, Pearson correlation coefficients were estimated (Normality of distribution per variable was assessed using the One-Sample Kolmogorov-Smirnov Test ($p > .05$)). Data analysis showed that, mean reading performance per text of the total number of 70 intermediate EFL test-takers correlated with three specific text variables, that is, Lexical Density ($r = -.326$, $p < 0.05$), syntactic structure similarity across all sentences ($r = .536$, $p < 0.05$), and words expressing agreement ($r = .456$, $p < 0.05$). Although moderate, these correlations might be taken to indicate that intermediate EFL test-takers scored lower in texts that were more lexically dense and higher when pertinent texts contained a higher proportion of syntactically similar sentences and lexical items showing agreement. Moreover, data analysis showed a negative correlation between individual intermediate EFL test-takers' mean scores in each set of multiple-choice reading comprehension questions and two text variables, i.e. proportion of past tenses ($r = -.645$, $p < 0.05$) and frequency of words occurring in the fourth and fifth band of the BNC corpus ($r = -.874$, $p < 0.05$). Being rather high, this effect might reflect intermediate EFL learners in processing less frequent words compared to the first and second thousand more frequent ones of the BNC corpus, while it further supports the view that vocabulary knowledge in general and word frequency in particular could affect reading comprehension [3, 6, 11, 15, 20], since less frequent words might not form part of intermediate EFL users' repertoire at this stage of their language learning process. The negative impact of past tenses on exam performance could also be treated as an indication of increased difficulty from the part of the readers, who might have struggled to disentangle relationships between present and past events or follow a series of past actions [14]. On the other hand, the higher proportion of verbs in future tenses ($r = .460$, $p < 0.05$) and present tenses ($r = .637$, $p < 0.05$) was found to correlate with mean reading performance scores in a significantly positive way, i.e. the higher the percentage of such features in intermediate reading texts, the higher the percentage of correct responses to pertinent questions. This finding could be taken to suggest that both features facilitate text processing as they better define time relations and further

clarify the intended meaning on behalf of the writer. Thus, these variables appear to be related to text complexity in a more subtle way, through their positive contribution to the elucidation of apparently confusing messages. In order to better define the relationship between written production linguistic complexity and mean reading performance scores on each text Pearson correlation coefficients were estimated and statistical analysis showed that higher mean reading performance correlated with higher syntactic structure complexity ($r=.647$, $p<0.05$), higher propositional density ($r=.574$, $p<0.05$), higher lexical diversity ($r=.835$, $p<0.05$), higher proportion of less frequent words ($r=.854$, $p<0.05$) and higher proportion of less concrete words ($r=.480$, $p<0.05$). These results could help us draw a profile of intermediate EFL learners' L2 growth regarding their competence in processing and producing written information with low-performers facing difficulties when exposed to less frequent words or when the number of pronouns in a text made reference a more cognitively demanding process to them. On the other hand, high performers seemed to take advantage of word stem repetition and even try to guess the meaning of unfamiliar words by making use of derivational rules and contextual information while they were also able to produce more elaborate texts with more sophisticated words, more complex sentence structure and fewer cohesive features in text as a function of grade level. The findings of the study provide practical guidance to EFL teachers, material developers and test designers as to the kind of linguistic strategies intermediate learners develop as a function of their grade level and suggestions to consider when designing classroom curricula, writing skills textbooks and exam papers.

References

- [1] Alderson, C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.
- [2] Alderson, C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). *The development of specifications for item development and classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening: Final report of The Dutch CEF Construct Project*. Unpublished Working Paper. Lancaster: Lancaster University
- [3] Anderson, R. & Freebody, P. (1983). Vocabulary Knowledge. In R. Ruddell, M. Ruddell & H. Singer (Eds.), *Theoretical Models and Processes of Reading* (pp. 343-371). Newark: International Reading Association.
- [4] Carrell, P. (1987). Readability in ESL. *Reading in a Foreign Language*, 4, 1, 21-40.
- [5] Crossley, S., Salsbury, T., McNamara, D. & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28, 4, 561-580.
- [6] Graesser, A., McNamara, D., Louwerse, M. & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers*, 36, 2, 193-202.
- [7] Halliday, M. & Hasan, R. (1976). *Cohesion in English*. New York: Longman.
- [8] Hillocks, G. (2002). *The testing nap: How state writing assessments control learning*. New York, NY: Teachers College Press.
- [9] Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, 19, 2, 255-271.
- [10] McCarthy, P. & Jarvis, S. (2010). MTLTD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42, 381-392.
- [11] McNamara, D., Crossley, S. & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication*, 27, 57-86.
- [12] Nagabhand, S., Nation, P. & Franken, M. (1993). Can Text be too Friendly? *Reading in a Foreign Language*, 9, 2, 895-907.
- [13] Nation, P. (2001). How many high-frequency words are there in English? In M. Gill, A. Johnson, L. Koski, R. Sell & B. Wårvik (Eds.), *Language, Learning and Literature: Studies Presented to Håkan Ringbom* English Department Publications 4 (pp. 167-181). Åbo Akademi University: Åbo.
- [14] Oakland, T. & Lane, H. (2004). Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests. *International Journal of Testing*, 4, 3, 239-252.
- [15] Perfetti, C., & McCutchen, D. (1987). Schooled language competence: Linguistic abilities in reading and writing. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics* (pp. 105-141). Cambridge, UK: Cambridge University Press.
- [16] Weir, C. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22, 3, 281-300.
- [17] Weston, J., Crossley, S., McCarthy, P. & McNamara, D. (2011). Number of words versus number of ideas: Finding a better predictor of writing quality. *Proceedings of the 24th International Florida Artificial Intelligence Research Society*.
- [18] Zhang, L. & Anual, S. (2008). The Role of Vocabulary in Reading Comprehension: The Case of Secondary School Students Learning English in Singapore. *RELC Journal*, 39, 1, 51-76.