Towards a Text Classification Profile per Level of English Language Competence: Intermediate vs. Advanced Reading Comprehension Texts

Dr. Trisevgeni Liontou
Ministry of Education
Greece

Abstract

The aim of the present study was to examine the existence of any statistically significant lexicogrammatical differences between intermediate and advanced reading comprehension exam texts of the Greek State Certificate of English Language Proficiency national exams in order to better define text complexity per level of competence. Through Coh-Metrix 2.1, LIWC2007, VocabProfile 3.0, CLAN suite of programs, CPIDR 3.0, Gramulator and Text Analyzer 135 text indices were estimated with regard to thirty-four authentic intermediate reading comprehension texts used between November 2003 and November 2011 examination periods and twenty-nine advanced texts used between April 2005 and May 2012 examination periods. The statistical analysis revealed significant differences between intermediate and advanced reading comprehension texts for a specific number of text features such as word, sentence and paragraph length, levels of word frequency, lexical richness and text entropy, proportion of passive sentences and presence of words with rich conceptual content. The main outcome of the research is the Text Classification Profile that includes a qualitative and quantitative description of lexicogrammatical features pertinent in intermediate and advanced reading comprehension exam texts.

Keywords: text complexity, readability, text classification, automatic text analysis

1. Introduction

Although the issue of text readability is long-standing and has a venerable research tradition, its impact on foreign language assessment has garnered increased attention over the last decade, while the number of studies on exam validity and reliability has also increased. Yet, most well-established international exam systems have failed to provide sound evidence of their text selection processes (Bachman *et al*., 1988: 128; Chalhoub-Deville & Turner, 2000: 528; Fulcher, 2000: 487). In fact, while reviewing the main characteristics of three respected international tests, Chalhoub-Deville and Turner (2000: 528-30) stressed the lack of adequate documentation regarding how the level of text difficulty is determined, and which processes for text selection are applied, with a view to establishing internal consistency of these tests and equivalence across parallel test forms. According to them (ibid: 528-9) making such information available to the public is mandatory, in order to help all interested parties make informed evaluations of the quality of the tests and their ensuing scores. Chalhoub-Deville and Turner concluded by pointing to the fact that especially nowadays, when language ability scores are used to make critical decisions that can affect test-takers' lives, developers of large-scale tests have the responsibility not only to construct instruments that meet professional standards, but also to continue to investigate the properties of their instruments over the years and make test manuals, user guides and research documents available to the public to ensure appropriate interpretation of test scores (ibid: 528-9; Khalifa & Weir, 2009: 17).

2. Background to the Study

The current study is closely linked to earlier findings of research on reading assessment, according to which many text variables such as topic and structure, word frequency and concreteness, propositional density and syntactic complexity can have an impact on either the reading process or product and need to be taken into account during test design and validation (Alderson, 2000: 81, Oakland & Lane, 2004: 247). Although a lot of research has been conducted in the field of second language acquisition with specific reference to ways of reading and text processing strategies, Alderson (2000: 104) stressed language testers' lack of success "to clearly define what sort of text a learner of a given level of language ability might be expected to be able to read or define text difficulty in terms of what level of language ability a reader must have in order to understand a particular text".

Such information would be particularly useful in providing empirical justification for the kinds of reading texts test-takers sitting for various language exams are expected to process, which to date have been arrived at mainly intuitively by various exam systems (Alderson, 2000: 104; Fulcher, 1997: 497; Lee & Musumeci, 1988: 173; Oakland & Lane, 2004: 243). Fulcher (1997: 497) also drew on the importance of text difficulty or text accessibility as a crucial but much neglected area in language testing. For him, defining text complexity is critical for test developers to become aware of the range of factors that make texts more or less accessible, in order to be able to select reading texts at appropriate levels for inclusion into the reading test papers of their examinations (ibid: 497). He further stressed that research in this area is particularly pertinent as text difficulty is re-emerging as an area of great concern not only in language teaching and materials writing but also in the testing community (ibid: 497). Echoing Fulcher, Freedle and Kostin (1999: 3) postulated that, ideally, a language comprehension test 'should' assess primarily the difficulty of the text itself; the item should only be an incidental device for assessing text difficulty. Especially in relation to reading tests, it has been shown that text variables can have a significant effect on both test item difficulty and test scores, regardless of the test method employed (Alderson, 2000: 61; Carr, 2006: 271; Davies & Irvine, 1996: 173; Frazier, 1988: 194; Freedle & Kostin, 1999: 5). In fact, although the research literature is full of evidence that text difficulty is one of the most important factors in reading comprehension, many researchers are still resorting to readability formulas or their own experience for assigning reading levels to texts (Oakland & Lane, 2004: 244; Shokrpour, 2004: 5). However, readability formulas have been widely criticized by both L1 and L2 language researchers for limiting their scope of analysis on rather basic text features, such as word and sentence length, and failing to take into account a number of additional factors that contribute to text difficulty, such as word abstractness and density of information (Bailin & Grafstein, 2001: 292; Carr, 2006: 282; Crossley *et al*., 2008: 476; Spadorcia, 2005: 37; Wallace, 1992: 77).

In a nutshell, 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, nevertheless, 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. Weir (2005: 292) further acknowledged that, although the Common European Framework of Reference for Languages (CEFR) attempted to describe language proficiency through a group of scales composed of ascending level descriptors, it failed to provide specific guidance as to the topics that might be more or less suitable at any level of language ability, or define text difficulty in terms of text length, content, lexical and syntactic complexity. In fact, according to Weir, the argument that the CEFR is intended to be applicable to a wide range of different languages "offers little comfort to the test writer, who has to select texts or activities uncertain as to the lexical breadth of knowledge required at a particular level within the CEFR" (Weir, 2005: 293). Alderson *et al*. (2004: 11) 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 for item writers or item bank compilers to make more use of it.

3. Aim of the Study

The aim of the present study was to examine the existence of any statistically significant lexicogrammatical differences between intermediate (B2) and advanced (C1) reading comprehension exam texts of the Greek State Certificate of English Language Proficiency national exams (KPG) in order to better define text complexity per level of competence. Although it is beyond the scope of the present paper to provide a detailed description of the KPG English Language exam it is worth mentioning that it is a relatively new multi-level multilingual suite of national language examinations developed by teams of experts from the foreign language departments of the National and Kapodistrian University of Athens and the Aristotle University of Thessaloniki. The exams are administered by the Greek Ministry of Education, making use of the infrastructure available for the Panhellenic university entrance exams. Despite being in its infancy, KPG is rapidly gaining acceptance as a high-stakes exam in Greece and, because of its official recognition by the state, it can influence one's future prospects for employment and education. Exams are administered twice a year and, since November 2003, more than 500,000 test-takers have taken part in the English language exams.

The level of the reading comprehension texts has been broadly defined in the Common Framework of the KPG exams according to which "the B2 Level exams are designed to test at an Independent User level the candidates' abilities to use English in order to understand the main ideas of texts of *average difficulty* on various topics, including abstract ideas or specialized information that requires some technical knowledge" whereas "the C1 Level exams are designed to test at a Proficient User level the candidates' abilities to understand relatively long texts and of a *high level of difficulty*" (Common Framework of the KPG examinations, 2003: 6). However, it has not yet been possible to define, based on empirical evidence, the readability level of texts and the specific lexicogrammatical features that could be more appropriate to the intended audience i.e. prospective B2 or C1 test-takers. The current research has, thus, been designed to fill this void and further add to our present state of knowledge on EFL text difficulty in general. In order to explore these issues, the following research question was formed:

1. Are there any statistically significant lexicogrammatical differences between intermediate and advanced reading comprehension exam texts? If yes, which text variables can better predict text difficulty variation between these two levels of English language competence?

4. Research Methodology

***Linguistic Text Analysis*:** The 135 text variables analyzed in the present research were chosen for both practical and theoretical reasons. First, from the practical standpoint of comparability, it was important to establish whether particular features existed, whose presence in the KPG English language reading comprehension exam texts might have introduced construct irrelevant variance into test scores. If this turned out to be the case, then steps could be taken to incorporate such factors into subsequent revisions of the text selection guidelines. In addition, given that previous research had failed to produce a definite set of quantifiable text variables, no decision was a priori made in terms of their expected significance. In relation to text analysis a combined model based on Systemic Functional Grammar and additional text features was adopted (see Appendix 1 for the complete list of text indices). To be more specific, the presence of cohesive ties created by referencing, conjunction and lexical cohesion as well as that of nominal group structure, grammatical intricacy and lexical density was explored. Moreover, the occurrence of surface text features, such as number of words, sentences and paragraphs per text, word frequency, lexical diversity, propositional density, proportion of passive sentences, negations, phrasal verbs and idioms per text along with estimates from four well-known readability formulas, namely the Flesh Reading Ease Index, the Dale-Chall Readability Index, the Fry Readability Index and the Gunning-Fog Index, was determined.

The KPG English Reading Dataset: Thirty-four authentic B2 reading comprehension texts used between November 2003 and November 2011 examination periods and twenty-nine C1 texts used between April 2005 and May 2012 examination periods were chosen for analysis. For texts to be appropriate for comparisons and avoid any test-method effects only those reading passages that contained ten multiple choice reading comprehension questions with three options (A, B or C) per item were considered appropriate for further analysis. Finally, these two levels of competence were chosen for reasons of practicality since, when the research began, they were the only ones available and had attracted a great number of test-takers.

Automated Text Analysis Tools: Advances in Computational Linguistics and Machine Learning systems have made it possible to go beyond surface text components and adopt more theoretically sound approaches to text readability, 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 (Graesser *et al*., 2004: 193). In the present study *Coh-Metrix 2.1*[1], *Linguistic Inquiry and Word Count 2007* (LIWC)[2], *VocabProfile 3.0*[3], *Computerized Language Analysis* (CLAN) suite of programs[4], *Computerized Propositional Idea Density Rater 3.0* (CPIDR)[5] and *Gramulator*[6] were used to estimate the 135 text variables.

[1] Available online at http://cohmetrix.memphis.edu/cohmetrixpr/index.html
[2] Available online at http://www.liwc.net
[3] Available online at http://www.lextutor.ca/vp
[4] Available online at http://childes.psy.cmu.edu.
[5] Available online at http://www.ai.uga.edu/caspr.
[6] Available online at https://umdrive.memphis.edu/pmmccrth/public/software/software_index.htm

5. Results & Discussion

Once the analysis of text characteristics per level was completed, independent sample t-tests[7] were carried out in order to explore and further determine the significance of existing differences between intermediate and advanced reading comprehension texts and thus answer the first research question (*Are there any statistically significant lexicogrammatical differences between intermediate (B2) and advanced (C1) KPG reading comprehension exam texts?)*. The analysis revealed that texts used at the C1 level included a significantly higher number of words (t=-6.642, df=61, r=0.67, p<0.001) than their B2 counterparts and were also characterized by significantly longer words in terms of average number of syllables per 100 words (t=-2.163, df=61, r=0.27, p=0.035), 6-letter words (t=-2.016, df=60.468, r=0.25, p=0.048), 11-letter words (t=-2.236, df=61, r=0.27, p=0.029) and 14-letter (t=-2.533, df=61, r=0.31, p=0.014) words, as well as a significantly higher number of sentences (t=-2.764, df=61, r=0.33, p=0.008), all of which could have contributed to increased text difficulty. Regarding ***word frequency***, an important measure of text difficulty as there is increasing research evidence that high-frequency words are normally read more quickly and are more easily understood than infrequent ones (Brown, 1993: 277; Laufer, 1997: 266; Nation, 1993: 120), significant differences were found for the sum of the five British National Corpus frequency bands and the sub-group of first and second most frequent 1000 words. To be more specific, advanced texts included a significantly lower proportion of the 5,000 most frequent word families present in the BNC corpus (t=2.322, df=61, r=0.28, p=0.024) and of those occurring in the first 2,000 of the same corpus (t=2.183, df=61, r=0.27, p=0.033).

These findings might indicate that more advanced texts are expected to be characterized by the progressively higher presence of less frequently used words. Moreover, the mean proportion of words from Coxhead's Academic Word List in B2 texts was 4.13%, whereas the adjusted mean frequency for content words and adjusted minimum logarithmic frequency scores for each sentence, obtained through Coh-Metrix, were 2.22 (SD=0.15) and 1.12 (SD=0.29) respectively, which could indicate that a limited number of rare content words were present in B2 texts. On the other hand, C1 texts contained a slightly higher proportion of academic words (Mean=4.85, SD=1.72) and less frequent content words in the CELEX database (Mean=2.15, SD=0.15), which could have made comprehension comparatively more difficult, since rarer content words would need to be decoded and integrated within the same discourse context (Brown, 1993: 277; Laufer, 1997: 266; McDonald & Shillcock, 2001: 296; McNamara *et al*., 2010: 306). The relatively low percentage of academic words in texts of both levels may be attributed to the fact that most texts were sourced from everyday newspapers and magazines and were in their majority narrative (69.3%) rather than scientific ones (31.7%). In addition, this may suggest that statistically significant vocabulary differences across the two levels could be drawn, should test designers become more alert to such features as the academic and technical word lists and take them into account during their text selection process, since C1 texts are addressed to advanced candidates who are expected to be able to understand specialized articles and longer technical instructions, even when they do not relate to their field of expertise (KPG C1 Exam Specifications, 2007: 12).

In the present research, five metrics of ***readability*** were calculated, namely the Flesch Reading Ease Index, the Flesch-Kincaid Grade Level (0-difficult, 100-easy), the Dale-Chall Grade Level, the Gunning-Fog Index (6-easy, 20-hard) and the Fry Readability Graph. It is notable that, despite the serious limitations of readability formulas, data analysis revealed a significant difference between B2 and C1 texts in relation to the employed indices, all of which rated B2 texts as less difficult than those used at the C1 level. This finding could be an indication that, despite their apparent simplicity, readability formulas do seem to align with KPG test designers' perception of text difficulty and might be of some practical usefulness to them during the text selection and validation process.

[7] Homogeneity of group variances per text variable was assessed using Levene's Test for Equality of Variances (p>.05). The alpha level of 0.05 was corrected for multiple tests using the Holm-Bonferroni adjustment. In addition to **t**, **df** and **p** values, the effect size was estimated as $r = \sqrt{\frac{t^2}{t^2 + df}}$.

The magnitude of the effect was considered "small" for values lower than 0.3, "medium" for values ranging from 0.3 to 0.5 and "large" for values above 0.5.

To be more specific, the B2 texts were found to have a mean score of 57.88 (SD=9.18, Min=38.77, Max=73.89) in the Flesch Reading Ease Index, a 10.54 (SD=2.24, Min=6.9, Max=15.20) in the Gunning-Fog Index and a Flesch-Kincaid Grade Level of 9.54 (SD=1.89, Min=5.59, Max=12.00), a Dale-Chall Grade Level of 9.57 (SD=1.90, Min=5.50, Max=14.00) and a Fry Level of 9.47 (SD=2.00, Min=6.00, Max=14.00). In comparison, C1 texts were generally of an advanced level with a mean score of 50.80 (SD=10.74, Min=33.29, Max=73.49) in the Flesch Reading Ease Index and were assigned higher grade levels by all relevant formulas with a Flesch-Kincaid Grade Level of 10.77 (SD=1.71, Min=6.33, Max=12.00), a Dale-Chall Grade Level of 11.29 (SD=2.69, Min=5.50, Max=16.00), a Gunning Level of 12.8 (SD=2.87, Min=8.00, Max=20.51) and a Fry Level of 11.01 (SD=2.24, Min=7.00, Max=15.00).A series of independent samples t-tests showed that all of these differences were statistically significant (i.e. Flesch Reading Ease Index: t=2.819, df=61, p=0.006; Flesch-Kincaid Grade Level: t=-2.679, df=61, p=0.009; Dale-Chall Grade Level: t=-2.957, df=61, p=0.004; Gunning-Fog Index: t=-3.467, df=61, p<0.001; Fry Readability Graph: t=-2.899, df=61, p=0.005), but within each group reading texts received different and at times contradictory ratings regarding their level of complexity. For example, the text entitled "Protecting the environment in Greece" that was used in May 2006 B2 examination was rated as the most difficult (15.2) of all B2 texts by the Gunning-Fog Index, while Dale-Chall categorized it as fairly easy (7.5).

Such a discrepancy of readability ratings within and across levels raises the question of which, if any, specific formula should be given priority, not only in interpreting present results but most importantly in selecting and classifying new texts. This question can be partly answered by comparing the magnitude of the effect of each variable, which appears to be slightly higher for the Gunning-Fog Index (r=0.40), followed by the Dale-Chall (r=0.35) and Fry (r=0.34) ones, whereas the Flesch Reading Ease Index (r=0.33) and the Flesch-Kincaid Grade Level (r=0.32) reflect a lower ability to distinguish between B2 and C1 KPG reading levels. Without doubt, although such measures may serve to highlight potential problems regarding text complexity, final decisions on the suitability of test material will inevitably need to weigh the impact of more complex text features and take into serious account a wider range of syntactic, semantic and cognitive aspects pertinent to the purposes of the specific exam.

Using *CPIDR 3.0*, ***propositional idea density*** scores, that is, the number of propositions divided by the number of words, were obtained for each of the 63 texts in our dataset. Although the mean incidence of ideas contained in B2 texts was much lower (Mean=189.08, SD=47.21) than that of ideas contained in C1 texts (Mean=277.51, SD=57.14), the independent samples t-test revealed no significant differences between B2 and C1 level texts regarding their overall propositional idea density. This surprising finding needs to be interpreted with caution since, contrary to existing research evidence that propositional density has a direct effect on readability (Graesser *et al*., 1997: 163; Kintsch *et al*., 1975: 212; Kintsch & Vipond, 1979: 329; Long &Baynes, 2002: 228; White, 2011: 85), more advanced KPG texts did not appear to contain a statistically significant higher proportion of ideas, a text feature that could have added to their complexity by increasing the amount of information needed to be processed by short-term memory. Apart from sounding a cautionary note towards that direction, it is worth pointing out that, despite the fact that propositional density values were scattered indicating lack of normal distribution within each level of competence, 161 ideas per text was the most frequent value for B2 level texts and 214 for the C1 ones, which could reflect a tendency for C1 texts to be more loaded in ideas, possibly as a result of their increased length.

Furthermore, the ***lexical richness*** of B2 and C1 level texts was measured through ten indices, i.e. percentage of word families along with percentage of types and tokens per family based on the British National Corpus, lexical diversity optimum D value measured through *vocd*, MAAS and MTLD values obtained via Gramulator, percentage of lexical density (LD) that refers to the proportion of content words to the total number of running words calculated through LIWC2007, and percentage of hapax legomena (unique types), dis legomena (words that occur only twice in the same text) and text entropy estimated using Stylometrics. The analysis of frequencies across levels showed that although B2 texts contained a higher percentage of unique words (Mean=41.51, SD=4.52) than their C1 counterparts (Mean=37.92, SD=3.88), they also included a higher percentage of tokens (Mean=0.57, SD=0.14) and types (Mean=0.32, SD=9.91) per word family, which indicates that words pertaining to the same family were most often repeated within each text and could, thus, have been easier for readers to process if they were familiar with the word stem. Moreover, both D (Mean=100.87, SD=22.34) and MTLD (Mean=103.05, SD=23.83) values were lower for B2 texts, which might be taken to reflect the less diverse vocabulary present in the specific discourse.

The complexity of C1 texts in terms of the range of vocabulary they contained was further supported by their lower percentage of dis legomena (Mean=6.15, SD=1.78) and higher mean scores on lexical density (Mean=49.87, SD=4.42) and entropy (Mean=2.23, SD=0.05). Taking the analysis a step forward, the statistical significance of these differences was further investigated by carrying out a set of independent samples t-tests. The alpha level of 0.05 was corrected to 0.005 using the Holm-Bonferroni adjustment model for multiple tests, while homogeneity of group variances per text variable was assessed using Levene's Test for Equality of Variances ($p > .05$). Significant differences between B2 and C1 texts were found for five of the ten lexical richness metrics. To be more specific, B2 texts contained a significantly higher proportion of dis legomena ($t=2.998$, $df=61$, $r=0.36$, $p=0.004$) along with tokens ($t=6.091$, $df=61$, $r=0.61$ $p<0.001$) and types ($t=6.086$, $df=61$, $r=0.61$, $p<0.001$) per family, whereas C1 texts were found to be significantly richer according to the lexical density ($t=-3.575$, $df=42$, $r=0.51$, $p=0.001$) and entropy ($t=-5.810$, $df=61$, $r=0.60$, $p<0.001$) measurements. These findings are considered particularly useful in providing a valid way of discriminating between texts of varying levels of lexical richness, while, at the same time, they make possible the selection of the most appropriate variables by reducing their number from ten to five.

The extent of ***abstractness*** in B2 and C1 KPG reading texts was measured through four indices, i.e. concreteness of content words, minimum concreteness of content words, mean verb and noun hypernym values. No significant differences across levels were found for all four measures, with their mean scores being almost identical. This could indicate the lack of subject specificity in the KPG reading dataset and the selection of texts of general interest that do not take an abstract, theoretical approach to topics, but are rather based on concrete, tangible information. Without doubt, this finding comes in agreement with the design of a general exam that is addressed to a wide range of EFL language users and assesses language competence in different discourse environments. At the same time, it needs to be pointed out that the inclusion of more abstract words at higher levels of language proficiency could make the comprehension process more demanding (Cacciari & Glucksberg, 1995: 291; Crossley *et al*., 2009: 322; Dufty *et al*., 2006: 1253; Salsbury *et al*., 2011: 352; Schwanenflugel *et al*., 1997: 545), should the exam specifications include such a requirement.

Syntactic complexity was investigated through a number of metrics that assessed the syntactic composition of sentences and the frequency of particular syntactic classes in each text. To be more specific, one set of analysis included nine text variables that signal logical or analytical difficulty, i.e. percentage of all connectives present in a text with them being further divided into positive and negative additive, temporal, causal and logical ones. Descriptive statistics showed that B2 texts contained a slightly higher proportion of positive (Mean=36.54, SD=14.28) and negative (Mean=9.73, SD=7.19) additive connectives as well as a higher percentage of positive (Mean=20.96, SD=7.06) and negative (Mean=0.95, SD=1.73) causal connectives, all of which could have facilitated comprehension by clarifying the relationships among ideas and providing a clear structural pathway for the readers to follow (Britton *et al*., 1982: 51; Caron *et al*., 1988: 309; Geva, 1992: 731; Zadeh, 2006: 1). On the other hand, C1 texts included more positive (Mean: 10.22, SD: 4.17) and negative (Mean: 0.58, SD: 1.02) temporal connectives, along with a higher proportion of positive (Mean: 19.49, SD: 6.40) and negative (Mean: 11.61, SD: 5.48) logical connectives. Although the independent sample T-test revealed only one statistically significant difference (positive temporal connectives: $t=-3.013$, $df=61$, $r=0.36$, $p<0.001$), the overall differences mentioned above may be taken to reflect a general tendency for C1 texts to entail more cognitively demanding processes, given their higher percentage of logical arguments and the frequent shift in time sequence.

Regarding ***syntactic structure similarity***, data analysis revealed no statistically significant differences across levels, although sentences of the B2 texts appeared to have more features in common both across (Mean=0.085, SD=0.02) and within paragraphs (Mean=0.093, SD=0.02). The final set of analysis explored the frequency of seven additional syntactic variables, i.e. percentage of negations, passive sentences, conditional operators, noun-phrase constituents, modifiers per noun phrase, verb-phrases per word and words before the main verb, whose increase has been reported to have a direct impact on text difficulty (Charrow, 1988: 93; Dufty *et al*., 2006: 1254; Gorin, 2005: 351; Kaup, 2001: 960; Kemper, 1987: 323; Kirschner *et al*., 1992: 546; Nagabhand *et al*., 1993: 900). In the KPG reading dataset, C1 texts generally included a higher percentage of these syntactic constituents, which comes in agreement with KPG exam specifications that C1 test-takers should demonstrate their ability to comprehend extensive, complex and more linguistically-demanding written texts.

In order to check the statistical significance of the above mentioned differences, independent samples t-tests were carried out for each set of variables, namely the nine indices measuring connectives, the three indices measuring syntactic structure similarity, and the seven indices measuring specific syntactic classes. The alpha level of 0.05 was corrected to 0.005 for the first set, 0.016 for the second set and 0.007 for the third one using the Holm-Bonferroni adjustment model for multiple tests, whereas homogeneity of group variances per text variable was assessed using Levene's Test for Equality of Variances ($p > .05$). The analysis revealed statistically significant differences between B2 and C1 texts for two specific syntactic measures, i.e. proportion of passive sentences ($t=-3.058$, $df=61$, $r=0.36$ $p=0.003$) and positive temporal connectives ($t=-3.053$, $df=61$, $r=0.36$ $p=0.003$). In other words these two features were found to make a significant contribution to distinguishing texts according to their syntactic complexity and could even be used to obtain a numerical cut-off point between proficiency levels, with passive sentences covering approximately 18% (SD=0.08) of the total number of words in C1 texts, as opposed to 11% (SD=0.09) in B2 texts, whereas positive temporal connectives accounted for 9% of running words in advanced texts and 7% in lower level ones. Most importantly, such a statistical procedure has made possible the significant reduction from nineteen to two in the number of syntactic variables that could be of practical usefulness and should be given priority in future investigations on syntactic complexity across various levels of language performance.

In the present research, ***referential cohesion*** was measured through eight indices, i.e. proportion of anaphoric references between adjacent sentences and for constituents mentioned up to five sentences earlier in the text, word stem and argument overlap between adjacent sentences and across paragraphs, proportion of content words -with no deviation in their morphological forms- overlapping between adjacent sentences and percentage of pronouns. Although no statistically significant differences were found between the two levels regarding these specific metrics, preliminary descriptive analysis showed a general tendency for B2 texts to contain a higher proportion of pronouns (Mean=8.81, SD=4.03) and anaphoric references between adjacent sentences (Mean=0.30, SD=0.19) and across the range of five sentences (Mean=0.15, SD=0.11), whereas C1 texts included a higher proportion of sentence pairs that shared one or more arguments (Mean=0.40, SD=0.12), content words (Mean=0.75, SD=0.01) and word stems (Mean=0.36, SD=0.16). The higher percentage of semantic similarity between C1 sentences was further confirmed by the scores obtained through Latent Semantic Analysis (LSA), since C1 texts received higher scores in all three measures of LSA cosines between adjacent sentences (Mean=0.193, SD=0.08), across all sentences (Mean=0.173, SD=0.08) and across paragraphs (Mean=0.292, SD=0.14). This surprising finding comes in contrast with our expectations and evidence from previous research that, given its facilitative effect on comprehension, overlapping of word units would be more extensive in lower level texts, whereas a higher density of pronouns is pertinent to advanced texts (Crossley & McNamara, 2009: 124; Douglas, 1981: 101; Field, 2004: 121; Horning, 1987: 58; Rashotte & Torgesen, 1985: 186; Rayner *et al*., 2004: 50-51). Without doubt, such an inconsistency may be taken to suggest that more explicit and statistically significant differences across levels could be drawn, should test designers become more alert to such in-depth text features and take them into account during the text selection process, especially when assessing the complexity of texts addressed to proficient English language users.

The degree of ***causal***, ***temporal***, ***spatial*** and ***intentional*** relations in B2 and C1 reading texts was investigated through four relevant indices provided by Coh-Metrix. The preliminary descriptive analysis showed a relatively higher percentage of causal (Mean=0.84, SD=0.36) and spatial (Mean=0.49, SD=0.08) networks in C1 texts, whereas a relatively higher proportion of temporal (Mean=0.83, SD=0.10) and intentional (Mean=14.37, SD=8.03) particles characterized B2 texts. This could indicate that the C1 reading dataset consisted of more complex texts with increased causal mechanisms and stories with an action plot, while the B2 reading dataset was likely to contain a higher number of simple stories and other forms of narrative with a clear time connection between events. Although no significant differences were found between B2 and C1 texts with regard to these cohesion measures, a more in-depth analysis of verb tenses, whose presence in each text was estimated using *LIWC2007*, revealed a statistically significant difference for past and present tenses per level of competence. To be more specific, the statistical analysis showed that B2 texts included a significantly higher proportion of present tenses ($t=3.009$, $df=61$, $r=0.36$ $p=0.004$), whereas past tenses were more frequent in C1 texts ($t=-2.756$, $df=61$, $r=0.33$ $p=0.008$). The higher incidence of past tenses in more advanced texts could be interpreted as an indicator of text complexity, as previous research has already shown that such tenses may have a negative impact on the ease and speed of text processing (Nagabhand *et al*., 1993: 900).

Finally, in order to explore the existence of any additional differences between B2 and C1 texts, a final set of variables that have been reported in reading literature to have an impact on text difficulty were subjected to thorough statistical analysis. The variables included in this data set referred to the percentage of idioms, phrasal verbs and Greek cognates as well as that of main and auxiliary verbs, adverbs and prepositions, articles and numbers in KPG reading texts. Our expectations that more idiomatic language would be present in advanced texts were not confirmed, since the results of the independent samples t-tests showed no significant differences between the two levels. However, a closer look at mean scores did reflect a tendency for B2 texts to contain a slightly higher proportion of phrasal verbs (Mean=1.32, SD=0.69), whereas more idioms (Mean=1.35, SD=0.49) were present in C1 texts. This lack of difference should be interpreted with caution at this stage of the research, given that a more in-depth analysis on the nature of idioms and phrasal verbs present in KPG test source texts might reveal significant differences within and across levels regarding the figurative or literal meaning of such expressions.

6. Implications of the Study

One of the most important outcomes of the within and across levels analysis of the lexicogrammatical features present in KPG reading comprehension exam texts has been the detailed description of their linguistic characteristics that has ultimately led to the creation of a Text Classification Profile per level of competence (B2/C1). The newly created profile (see Table 1) could be of practical use in various exam batteries' attempt to qualify the communicative descriptors included in their exam specifications with the interrelated linguistically articulated features pertaining to two different levels of language proficiency and, thus, describe the linguistic qualities of the texts that test-takers at each of these two levels must be able to handle for a successful exam performance. At the same time, the proposed Text Classification Profile could increase the construct validity of pertinent exams and serve as a yardstick for item writers to select future reading texts by taking into account the average linguistic profile of intermediate and advanced reading comprehension texts. For the sake of simplicity, the profile presented below includes basic qualitative information regarding the discoursal and lexicogrammatical characteristics of pertinent texts. Nevertheless, interested parties could also refer to Appendices 2 and 3 for the complete list of quantifiable text features.

Moreover, word frequency was found to be an important indicator of text complexity, with more advanced reading texts been characterized by the progressively higher presence of less frequently used words. As already presented in the Text Classification Profile, this implies that in intermediate texts word units belonging in the first 1,000 and second 1,000 most frequent families of the BNC corpus are expected to account for approximately 90% of the lexicon. On the other hand, advanced texts are expected to include a lower proportion of highly frequent words (about 85% of their lexis) and increased lexical complexity to be achieved by adding less frequent words from the third (5%), fourth (3%) and fifth (2%) frequency bands of the BNC corpus. These results could provide valuable information to EFL teachers regarding the breadth of vocabulary knowledge their students need to have for successful comprehension of pertinent reading texts. To be more specific, if we follow Nation's suggestion that 95-98% of text lexis is needed for unassisted reading of a range of authentic texts (2006: 70), prospective B2 test-takers might need to familiarize themselves with the words included in the first four frequency bands of the BNC corpus, namely 4,000 word families, while C1 test-takers will need a full coverage of the first five frequency bands, namely 5,000 word families plus some basic knowledge of more technical vocabulary included in Coxhead's academic word list, for processing advanced texts. Without doubt, this recommendation should by no means be treated as a panacea towards successful exam performance, since additional text, task and reader factors will inevitably have an impact on the reading process. Nevertheless, defining the word frequency profile of English reading comprehension exam texts and providing all interested parties with vocabulary lists might prove particularly useful for EFL teachers to consider, especially when designing their classroom curricula or preparing students for relevant exams. For instance, EFL teachers might wish to devote time to the revision and consolidation of words appearing in the first two frequency bands of the BNC corpus, since such a vocabulary-based instruction could provide prospective test-takers with minimum 85%-90% coverage of vocabulary frequently present in the specific KPG texts. The lists could also be a valuable source of information for EFL publishers, who wish to produce preparation materials for various English language exams, but are still selecting texts on intuitive grounds.

In addition, more in-depth linguistic features that could help test designers take even more consistent and informed decisions when distinguishing between intermediate and advanced level reading texts, could include estimates of lexical density and lexical richness, such as the percentage of content words, tokens and types per word family and proportion of dis legomena, with intermediate texts being characterized by a lower lexical density score (30%-40%) and increased repetition of types and tokens. Regarding syntactic complexity, the occurrence of four specific linguistic features, i.e. proportion of past and present tenses, passive sentences and positive temporal connectives, demonstrated significant differences between the two levels and could be used to obtain a numerical cut-off point between them, with passive sentences expected to cover approximately 18-20% of the total number of words in C1 texts as opposed to 10-12% in B2 texts and positive temporal connectives to account for about 9%-10% of running words in advanced texts and 6-7% in lower level ones. Moreover, a higher incidence of past tenses in more advanced texts (4%) could be an indicator of text complexity, while present tenses are expected to be more frequently used in B2 texts (6%).

Against our expectations, no significant differences were found regarding additional B2 and C1 text features especially in relation to syntactic and semantic complexity and text abstractness. This may be taken to suggest that more explicit text differences across levels could be drawn, should test designers become more alert to such features as idioms and phrasal verbs, verb and noun hypernym levels, anaphoric reference and content word overlap and take them into account during the text selection process. Finally, it is worth making test designers aware of the fact that a higher percentage of semantic similarity was noticed across C1 rather than B2 sentences. This surprising finding warrants further investigation, since it comes in contrast with evidence from previous research that, due to its facilitative effect on comprehension, overlapping of word units is expected to be more extensive in lower level texts, whereas a higher density of pronouns is pertinent to advanced texts (Crossley & McNamara, 2009: 124; Douglas, 1981: 101; Field, 2004: 121; Horning, 1987: 58; Rashotte & Torgesen, 1985: 186; Rayner *et al*., 2004: 50-51).

7. Concluding Remarks

Through its investigation of the significant relationships among a range of text variables, the present research attempted to provide evidence regarding the contribution of various lexical indices to text complexity. It also aspired to make a methodological contribution in that, instead of examining a limited number of text variables independently, it made use of advanced text analysis software applications and investigated the impact of 135 text variables on readability. On the other hand, as with all studies, the implementation of the present one presented a number of challenges and limitations that we hope will be overcome in future research. For instance, due to the fact that the KPG English language exam battery is still in its infancy, the number of available texts was inevitably constrained to a total of sixty-three; should the number of exam texts increase, the generalizability of present results might be further strengthened. It would also be useful to extend the present analysis to texts at both lower and higher levels of a range of English language exams following a comparative corpus-based approach for evidence-based conclusions to be drawn from a much more extensive dataset.

Acknowledgements

Sincere thanks are due to Prof. Bessie Dendrinos, Director of the RCeL and President of the KPG Central Examination Committee, for her valuable suggestions in the present research. This study was supported in part by the RCeL, which has generously granted me official access to available data. However, the views expressed in this paper do not reflect the official policy of the Center and the responsibility for the way data has been presented and interpreted relies entirely on the researcher.

References

Alderson, C. (2000). Assessing Reading. Cambridge: Cambridge University Press.

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

Bachman, L., Kunnan, A., Vanniaraian, S. & Lynch, B. (1988).Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. Language Testing, 5, 2, 128-159.

Bailin, A. & Grafstein, A. (2001).The linguistic assumptions underlying readability formulas: a critique. Language & Communication, 21, 3, 285-301.

Brown, C. (1993). Factors affecting the acquisition of vocabulary: frequency and saliency of words. In T. Huckin, M. Haynes & J. Coady (Eds.), Second language reading and vocabulary learning (pp. 263-286). Norwood, New Jersey: Ablex.

Britton, B., Glynn, S., Meyer, B. &Penland, M. (1982). Effects of text structure on the use of cognitive capacity during reading. Journal of Educational Psychology, 74, 1, 51-61.

Cacciari, C. &Glucksberg, S. (1995). Understanding idioms: do visual images reflect figurative meanings? European Journal of Cognitive Psychology 7, 3, 283-305.

Caron, J., Micko, H. &Thuring, M. (1988).Conjunctions and the recall of composite sentences.Journal of Memory and Language, 27, 3, 309-323.

Carr, N. (2006). The factor structure of test task characteristics and examinee performance.Language Testing, 23, 3, 269-289.

Chalhoub-Deville, M. & Turner, C. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS and TOEFL. System, 28, 4, 523-539.

Charrow, V. (1988). Readability vs. comprehensibility: a case study in improving a real document. In A. Davison & G. Green (Eds.), Linguistic complexity and text comprehension (pp. 85-114). Hillsdale, New Jersey: Lawrence Erlbaum.

Crossley, S. & McNamara, D. (2009).Computational assessment of lexical differences in L1 and L2 writing. Journal of Second Language Writing, 18, 2, 119-135.

Crossley, S., Greenfield, J. & McNamara, D. (2008). Assessing Text Readability Using Cognitively Based Indices. TESOL Quarterly, 42, 3, 475-492.

Crossley, S., Salsbury, T. & McNamara, D. (2009).Measuring L2 lexical growth using hypernymic relationships.Language Learning, 59, 2, 307-334.

Davies, A. & Irvine, A. (1996). Comparing test difficulty and text readability in the evaluation of an extensive reading programme. In M. Milanovic& N. Saville (Eds.), Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem (pp. 165-183). Cambridge: Cambridge University Press.

Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.), Learning to read in different languages (pp. 93-102). Washington: Center for Applied Linguistics.

Dufty, D., Graesser, A., Louwerse, M. & McNamara, D. (2006). Assigning Grade Levels to Textbooks: Is it just Readability?In R. Son (Ed.), Proceedings of the 28th Annual Meeting of the Cognitive Science Society (pp. 1251-1256). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Field, J. (2004). Psycholinguistics: The key concepts. New York: Routledge.

Frazier, L. (1988). The study of Linguistic Complexity.In A. Davison & G. Green (Eds.), Linguistic Complexity and Text Comprehension (pp. 193-221). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Freedle, R. &Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks.Language Testing, 16, 1, 2-32.

Fulcher, G. (1997). Text difficulty and accessibility: Reading Formulas and expert judgment. System, 25, 4, 497-513.

Fulcher, G. (2000). The "communicative" legacy in language testing.System, 28, 4, 483-497.

Geva, E. (1992). The role of conjunctions in L2 text comprehension.TESOL Quarterly, 26, 4, 731-745.

Gorin, J. (2005). Manipulating processing difficulty of reading comprehension questions: the feasibility of verbal item generation. Journal of Educational Measurement, 42, 4, 351-373.

Graesser, A., McNamara, D., Louwerse, M. &Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments & Computers, 36, 2, 193-202.

Graesser, A., Millis, K. &Zwaan, R. (1997). Discourse comprehension. Annual Review of Psychology, 48, 1, 163-189.

Horning, A. (1987).Propositional Analysis and the Teaching of Reading with Writing.Journal of Advanced Composition, 6, 49-64.

Kaup, B. (2001). Negation and its impact on the accessibility of text information.Memory & Cognition, 29, 7, 960-967.

Kemper, S. (1987). Life-span changes in syntactic complexity. Journal of Gerontology, 42, 3, 323-328.

Khalifa, H. & Weir, C. (2009). Examining Reading: Research and practice in assessing second language reading. Cambridge: Cambridge University Press.

Kintsch, W., Kozminsky, E., Streby, W., McKoon, G. & Keenan, J. (1975).Comprehension and recall of text as a function of content variables.Journal of Verbal Learning and VerbalBehavior, 14, 2, 196-214.

Kintsch, W. &Vipond, D. (1979).Reading comprehension and readability in educational practice and psychological theory. In L. Nilsson (Ed.), Perspectives on Memory Research (pp. 329-366). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Kirschner, M., Wexler, C. & Spector-Cohen, E. (1992).Avoiding Obstacles to Student Comprehension of Test Questions. TESOL Quarterly, 26, 3, 537-556.

Laufer, B. (1997). Beyond 2000: A measure of productive lexicon in a second language. In L. Eubank, L. Selinker& M. Sharwood-Smith (Eds.), The current state of Interlanguage: Studies in honor of William E. Rutherford (pp. 265-272). Amsterdam: John Benjamins.

Lee, J. &Musumeci, D. (1988).On Hierarchies of Reading Skills and Text Types. The Modern Language Journal, 72, 2, 173-187.

Long, D. &Baynes, K. (2002). Discourse representation in the two cerebral hemispheres. Journal of Cognitive Neuroscience, 14, 2, 228-242

McDonald, S. &Shillcock, R. (2001).Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing.Language and Speech, 44, 3, 295-323.

McNamara, D., Louwerse, M., McCarthy, P. & Graesser, A. (2010).Coh-Metrix: Capturing Linguistic Features of Cohesion. Discourse Processes, 47, 4,292-330.

Nation, P. (1993). Vocabulary size, growth and use. In R. Schreuder& B. Weltens (Eds.), TheBilingual Lexicon (pp. 115-134). Amsterdam: John Benjamins.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? The Canadian Modern Language Review, 63, 1, 59-82.

Nagabhand, S., Nation, P. & Franken, M. (1993). Can Text be too Friendly? Reading in a Foreign Language, 9, 2, 895-907.

Rashotte, C. &Torgesen, J. (1985). Repeated reading and reading fluency in learning disabled children. Reading Research Quarterly, 20, 2, 180-188.

Rayner, K., Pollatsek, A., Ashby, J. & Clifton, C. (2011).The psychology of reading (2nd edition). New York: Taylor & Francis Ltd.

Oakland, T. & Lane, H. (2004). Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests. International Journal of Testing, 4, 3, 239-252.

Salsbury, T., Crossley, S. & McNamara, D. (2011).Psycholinguistic word information in second language oral discourse.Second Language Research, 27, 3, 343-360.

Schwanenflugel, P., Stahl, S. &McFalls, E. (1997).Partial Word Knowledge and Vocabulary Growth during Reading Comprehension. Journal of Literacy Research, 29, 4, 531-553.

Shokrpour, N. (2004). Systemic Functional Grammar as a Basis for Assessing Text Difficulty.Indian Journal of Applied Linguistics, 30, 2, 5-26.

Spadorcia, S. (2005). Examining the text demands of high-interest, low-level books. Reading & Writing Quarterly, 21, 1, 33-59.

Wallace, C. (1992). Reading. Oxford: Oxford University Press.

Weir, C. (2005).Limitations of the Common European Framework for developing comparable examinations and tests. Language Testing, 22, 3, 281-300.

White, S. (2011). Understanding adult functional literacy: connecting text features, task demands and respondent skills. New York: Routledge.

Zadeh, E. (2006). The Role of Textual Signals in L2 Text Comprehension.ESP Malaysia, 12,1-18.