The Application of Objective Measures of Text Difficulty to Language Examinations

Gábor Szabó
University of Pécs, Hungary
szabo.gabor2@pte.hu

1. Introduction

The comparability of language examinations has long been an issue all over the world. While not all language examinations target specific levels, most of them claim to test candidates' language proficiency as related to specifically defined level requirements, which are either generally accessible or are specific to the exam. With the publication of the *Common European Framework of Reference* (Council of Europe, 2001), exam providers have increasingly been adding CEFR labels to their exams; thus, comparability has become an ever growing concern, and the main issue seems to have been the question of how successfully exams are aligned with the CEFR levels (cf. Harsch & Hartig, 2015; Martyniuk, 2010). Although a detailed set of guidelines concerning alignment have now been available for quite some time in the form of the *Manual for Relating Examinations to the CEFR* (Council of Europe, 2009), the doubts about the comparability of supposedly CEFR-based examinations continue to prevail (e.g., Vinther, 2013).

In the Hungarian context language examinations are typically offered in the framework of national accreditation, which is a requirement for exam providers to offer nationally acknowledged exam certificates. The accreditation procedure is discussed in the *Accreditation Handbook* (Educational Authority, 2018, Chapter 2), and it includes procedures to verify whether the exams are sufficiently linked to the CEFR levels. While the rigorous procedures are intended to guarantee the comparability of the exams, so far few empirical studies (e.g., Szabó & Kiszely, 2010) have attempted to verify this assumption. In this paper a study is presented in which the texts used in the B2 level reading comprehension papers of four commercial examinations and the advanced level school-leaving examination (also recognized as an examination at level B2) are compared in order to examine whether the texts are of the same level of difficulty.

2. Background

The construction of reading comprehension tests necessitates the selection of appropriate texts, which meet requirements that are determined by the specifics of the context for the test. One such requirement tends to be the level of the text. While the term “level” is frequently used in this context, determining the level of an actual text may be more challenging than one might presume. First, the level of the text is usually interpreted in relation to the comprehensibility of the meaning of the text. Meaning, however, is not a clear-cut concept. Instead of meaning, it may be more appropriate to discuss the concept of meaning potential (Halliday, 1978), i.e. the view that texts have no meaning as such; rather, they have potential for meaning, which then is realized in turn by the reader. Indeed, Alderson even argues that individual readers construct unique understandings of a text (Alderson, 2000, p. 6). Along the same lines, one may even argue that, because of unique interpretations, texts do not even *have* levels, only particular characteristics. While this may appear to be a rather extreme position, it demonstrates how challenging it may be to come to a decision concerning a text’s level. The problematic nature of determining text levels is further supported by the fact that texts may be understood at different levels depending on level of detail and interpretation. This approach to text difficulty is, in fact, supported by the CEFR, as the descriptors frequently differentiate levels based on whether only the main points, or also details or even implied messages of the texts are understood (cf. Council of Europe, 2001).

Text level, then, can be interpreted in a variety of ways; yet, several attempts have been made in order to somehow quantify this property of texts in the form of various indices designed to capture the elusive idea of text difficulty, and thus text level. Most of these measures have been readability indices, probably the best known ones of which are the Flesch Reading Ease and the Flesch-Kincaid Grade Level indices, both of which are based on a hypothetical relationship between the number of words, the number of sentences and the number of syllables (Klare, 1974-1975). Several authors (e.g., Alderson, 2000; Brown, 1998), however, have been quite critical of these indices, especially in an L2 environment, mainly on grounds that they are too simplistic.

Recently, more elaborate attempts have been made to develop measures of text difficulty, an example of which is the Coh-Metrix readability formula (Graesser, McNamara, & Kulikowich, 2011). Coh-Metrix describes text characteristics with the help of 53 measures. As interpreting such a large number of measures would be rather impractical, principle component analysis has been used to reduce the number of measures to eight principal components: narrativity, referential cohesion, syntactic simplicity, word concreteness, causal cohesion, verb cohesion, logical cohesion, and temporal cohesion. These components have then in turn been mapped to the five-level theoretical model proposed by Graesser and McNamara (2011): Genre (narrativity), Situation model (causal cohesion, verb cohesion, logical cohesion, and temporal cohesion), Textbase (referential cohesion), Syntax (syntactic simplicity), and Words (word concreteness). Thus, Coh-Metrix results are expressed along these five dimensions, which makes interpretation significantly easier to follow.

On the basis of Coh-Metrix results, a specific L2 readability index has also been constructed, which relies on a lexical, a syntactic and a meaning construction index (Crossley, Greenfield, & McNamara, 2008). The formula has also been compared to traditional formulas developed to measure readability and was found to be superior to them (Crossley, Allen, & McNamara, 2011).

Since Coh-Metrix combines several different facets of text difficulty, it seems like an ideal instrument to be applied for determining text levels in a variety of contexts. Accordingly, in the following Coh-Metrix indices will be utilized to perform a comparative analysis of texts.

3. The study

3.1. Research design

As was mentioned above, the current study intended to compare the difficulties of the texts used in B2 level English reading comprehension examinations. The study was focusing on English, as it is by far the most commonly learned foreign language in Hungary, and on level B2, because this level is the most popular one in nationally accredited examinations. Data were to be collected from sample materials of four nationally accredited commercial examinations, as well as live test materials used in the advanced level school-leaving examination, which is also officially recognized as a B2 level exam. The materials to be analyzed included all texts used in the reading component of the respective examinations. This meant that in the case of tasks where the texts were not presented to candidates in a complete form (e.g., banked gap-filling), texts were reconstructed into their original form. Thus, the texts, regardless of what task types they were linked to in the exams, became comparable. The advantage of this approach was that, unlike in a similar earlier study (Szabó, 2014), the comparative analysis could be performed on all texts used in each exam, thus providing a more comprehensive account of the difficulty of the texts used.

The actual analysis was conducted using two web tools made available on the Coh-Metrix website: the *Coh-Metrix Common Core Text Ease and Readability Assessor (TERA)* web tool and the *Coh-Metrix* web tool (McNamara, Louwerse, Cai, & Graesser, 2013). These two web tools provide a large number of measures related to text difficulty and readability. The current study relied on the following ones:

- narrativity
- syntactic simplicity
- word concreteness

- referential cohesion
- deep cohesion
- Coh-Metrix L2 readability.

The first five of these measures are provided by TERA, where the results are expressed in percentile figures. A sample TERA output is presented in Figure 1.

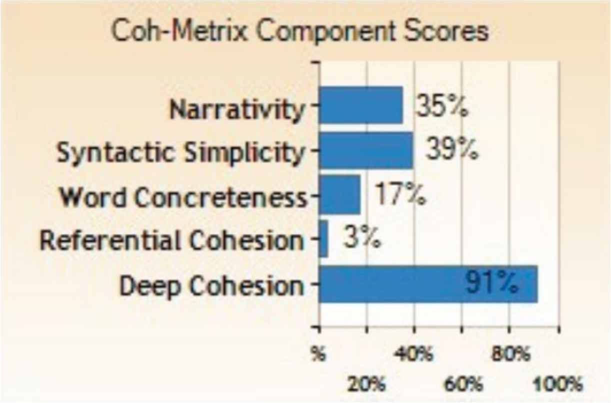


Figure 1. Sample TERA output

Narrativity represents a continuum stretching between texts that are highly narrative in nature, believed to be easier to process, and informational texts, which are more difficult to understand. Narrative texts contain a high proportion of frequent words and easy-to-understand verbs as well as pronouns making texts more engaging for readers (Jackson, Allen, & McNamara, 2017, p. 55.)

Syntactic simplicity is defined in terms of the complexity of sentences in the text. The measure is based on several indices of syntactic complexity, including the number of clauses and the number of words in a sentence, as well as the number of words before the main clause. The similarities in sentence construction across paragraphs are also taken into account (Jackson et al., 2017, p. 56.).

Word concreteness is based on the proportion of abstract and concrete words in the text. Abstract words are believed to make comprehension more difficult; therefore, a text with a large proportion of concrete words is thought to be easier to understand (Jackson et al., 2017, p. 57.).

Referential cohesion is expressed in terms overlap between words, word stems and concepts from sentence to sentence. A high proportion of overlaps is considered to make comprehension easier (Jackson et al., 2017, p. 57.).

Deep cohesion as a measure is based on the number of connectives in the text, representing how well the events or the various bits of information in the text are tied together. A high number of connectives indicates stronger links, making comprehension easier (Jackson et al., 2017, p. 58.).

The fifth measure, *L2 readability*, is accessible with the Coh-Metrix web tool, and it is expressed as a score, which is based on lexical frequency, syntactic similarity, and content word overlap (Crossley, Greenfield, & McNamara, 2008).

Once the indices above were obtained for all texts, statistical checks for significant differences across the texts were performed. On the one hand, it was examined whether the texts used in the same exam showed any differences; on the other hand, and perhaps more importantly, it was checked whether the texts used in different exams showed any significant differences. In order to detect significant differences, an independent samples Kruskal-Wallis test was run. The rationale for this procedure was that, since all the readability indices discussed above feed into the same construct, they may legitimately be considered as different facets of the same property of a text (i.e. difficulty). Hence, all the indices can be treated as scores, even if on an ordinal rather than an interval scale.

It is important to clarify that this study did not intend to map Coh-Metrix scores on the CEFR or vice versa. While references are made in the discussion to how TERA measures may be explained in terms of CEFR descriptors, this can only be done in a tentative manner. The reason for this is that the Coh-Metrix based quantitative measures and the qualitative CEFR descriptors approach difficulty from different perspectives: Coh-Metrix focuses on objectively measurable text properties, while the CEFR descriptors intend to capture what learners at particular levels can do. Thus, the purpose of the study was rather to approach the texts that have most probably been chosen with CEFR levels in mind using a set of objective measures of text readability to see how they compare.

3.2. Method

As has been discussed earlier, the data for the analysis comprised texts. They were collected from the sample reading tests provided on their homepages by four major commercial exam providers in Hungary: BME, ECL, Euro and Origo. Since the exams have differing structures, the number of texts in each exam is not the same, either. Yet, since the sample materials are to demonstrate the format of the complete reading component in each case, the texts collected may well be considered to be representative of the respective exams' reading components. Three texts were collected from BME, two from ECL, three from Euro and two from Origo. The fifth source of texts was the reading component of the May 2018 version of the advanced level school-leaving exam, which included four texts. It is worth noting that this latter exam was the only one where the analysis could be performed on live test materials.

3.3. Results and discussion

First of all, it seems appropriate to examine the actual figures yielded by the analysis measure by measure. The results for each measure are presented in separate Figures where the texts belonging to the same examination are marked by the same color and have a sequence number. The different examinations are referred to by their names, while the school-leaving exam is abbreviated as SLE. The first measure in line is Narrativity, depicted in Figure 2.

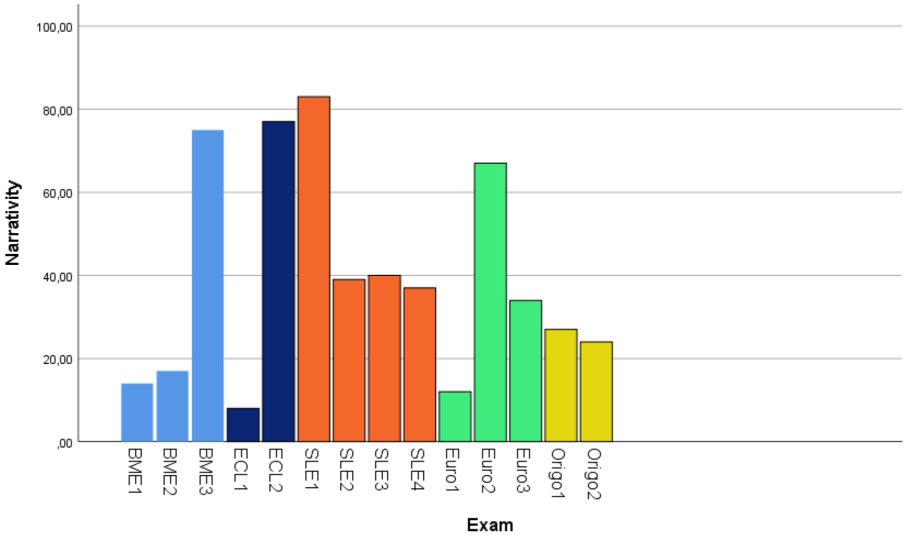


Figure 2. TERA results for "Narrativity"

As can be observed, results for Narrativity indicate considerable variety across the texts. The actual figures range from 8% (ECL1) to 83% (SLE1). Indeed, according to the TERA narrative descriptors, four texts (BME1, BME2, ECL1 and Euro1) are considered to be low in narrativity and thus more difficult to comprehend, another four texts (BME3, ECL2, SLE1 and Euro2) are high in narrativity and, accordingly, are easier to understand, while the rest of the texts are considered to be average in narrativity. Based on the above it seems reasonable to presume that the texts and thus the exams differ in difficulty. A note of caution is appropriate here, however. While narrativity clearly contributes to text difficulty, it would likely be unjustified to claim it is the main feature of text difficulty. Indeed, considering the characteristics of B2 level reading ability, it seems likely that texts at different levels of narrativity may still be considered to qualify for B2 level reading. This assumption is confirmed by a B2 descriptor from the "Overall reading comprehension" scale of the CEFR that says: "Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively" (Council of Europe, 2001, p. 69). The adaptation as well as the difference in texts and reading purposes point to a variety that could clearly include texts not characterized by a high degree of narrativity. Another observation worth noting is the fact that, with the exception of Origo, texts show considerable variation in terms of narrativity within examinations. Euro is a particularly good example of this, where each of the three texts used shows markedly different degrees of narrativity. Considering the fact that validity is significantly supported by the broadest possible sampling of the content domain, a wide variety in terms of narrativity may, in fact, mark a greater degree of content validity.

Next, let us take a closer look at the results on the second measure provided by the analysis, Syntactic simplicity, in Figure 3.

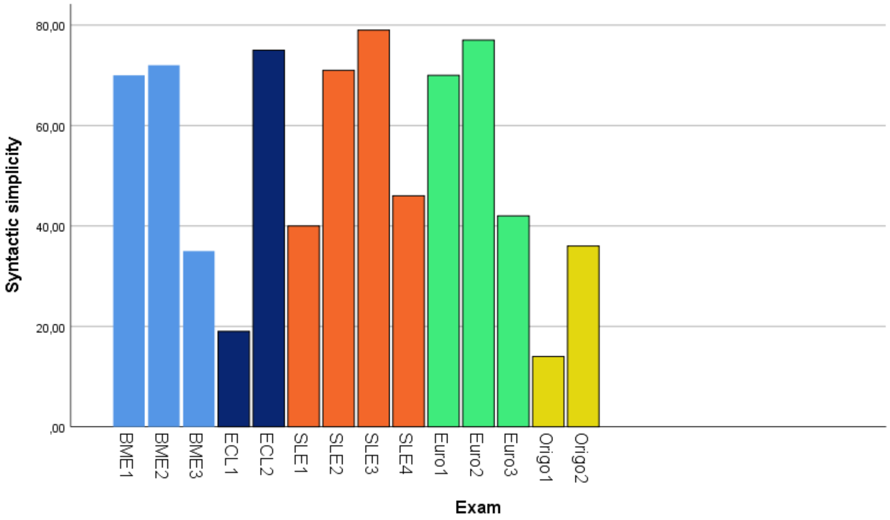


Figure 3. TERA results for “Syntactic simplicity”

Concerning this measure, the picture is slightly different, even though the spread is considerable again ranging from 14% (Origo1) to 79% (SLE3). There are only two texts (ECL1 and Oriogo1) that can be classified as low on syntactic simplicity and thus relatively difficult to read. Seven of the texts (BME1, BME2, ECL2, SLE2, SLE3, Euro1 and Euro2) yielded high scores on this measure, indicating that these texts are relatively easy, while the rest of the texts can be considered average on syntactic simplicity. Once again, it needs to be noted that while syntactic simplicity is a facet of difficulty, it is not necessarily a crucial factor. When attempting to evaluate these figures in light of the CEFR, we need to face a challenge. Though the CEFR discusses general linguistic range and grammatical accuracy in the form of descriptor scales, it does so almost exclusive with production in the focus. This is not fundamentally different in the recently published Companion Volume (CV) to the CEFR either; yet, the Grammatical Accuracy scale of the CV contains a new descriptor that may provide some guidance: “Has a good command of simple language structures and some complex grammatical forms, although he/she tends to use complex structures rigidly with some inaccuracy” (Council of Europe, 2017. p. 132). While the focus is clearly on production, having “a good command” of structures may well be interpreted as referring to reception as well. If the descriptor is interpreted with this in mind, the few examples of texts with low levels of syntactic simplicity seem justified. All the more so, as, once again, variety can be observed within all examinations under scrutiny.

The third measure applied was Word concreteness. The results are presented in Figure 4. As is apparent, the texts examined show, once again, a considerable spread with respect to this measure, ranging from 13% (Euro3) to 95% (ECL1). There are three texts (BME3, ECL2 and Euro3) that are considered to be low on word concreteness, indicating that these texts are relatively difficult to understand, while six of the texts (ECL1, SLE1, SLE2, SLE4, Euro1 and Origo1) are high on word concreteness and are, in turn, relatively easy to comprehend. The remaining texts are considered average in terms of word concreteness. Just like in the case of the previous measures, we need to add here though that word concreteness alone is not to be interpreted as the ultimate measure of difficulty. In an attempt to evaluate the significance of these results, it seems reasonable to consult the scales in the CEFR that relate to vocabulary. A B2 level descriptor from the Overall Reading Comprehension scale says, for instance, that a reader at this level "has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms" (Council of Europe, 2001, p. 69). Another descriptor from the Vocabulary Range scale says that a reader at this level "has a good range of vocabulary for matters connected to his/her field and most general topics" (Council of Europe, 2001, p. 112). A new descriptor from the same scale, as presented in the CV says a B2 learner "can understand and use much of the specialist vocabulary of his/her field but has problems with specialist terminology outside of it" (Council of Europe, 2017, p. 131).

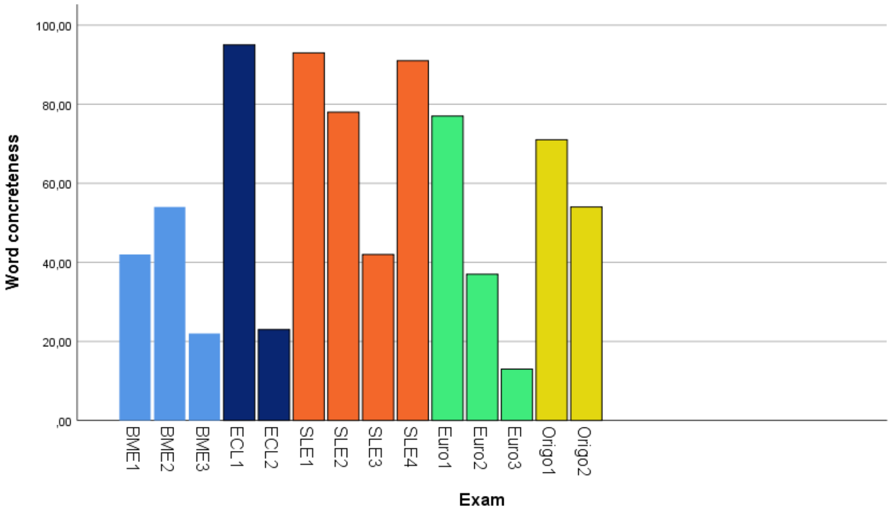


Figure 4. TERA results for "Word concreteness"

As can be seen, the scales are more informative in this case than concerning syntactic simplicity, even though word *concreteness* is not specifically addressed. Yet, the reference to "low frequency idioms" as well as "specialist terminology" seem to relate to word concreteness (or the lack of it), indicating that at this level a higher degree of word concreteness is most probably acceptable. Just like in the case of the previous measures, variation within examinations is again to be emphasized, with Euro, once again, providing three texts with three markedly different levels of word concreteness, but other examinations, apparently, also sampling the construct with the help of texts differing in this respect.

The fourth measure to be examined is Referential cohesion. The results for this measure are presented in Figure 5. Compared to the previous measures, the figures here appear to be quite different. First, they are lower than in previous measures, and they also show a much narrower spread. The actual range is from 1% (BME2) to 31% (BME3 and SLE1). What this means is that all but the highest two texts fall into the low category on this measure, and even 31% means an average level of referential cohesion. To understand what may lie behind this phenomenon, it is worth coming back to the notion of referential cohesion. As has already been discussed, this property refers to the overlap that may exist between words, especially those referring to content, with the help of similar or identical words as well as the ideas they convey. A low level of referential cohesion is interpreted as a source of difficulty (McNamara, Graesser, Cai, & Kulikowich, 2011, p. 8). As the TERA analysis puts it, "… low referential cohesion indicates there is less overlap in explicit words and ideas between sentences. These conceptual gaps require the reader to make more inferences" (TERA text comparison, 2018).

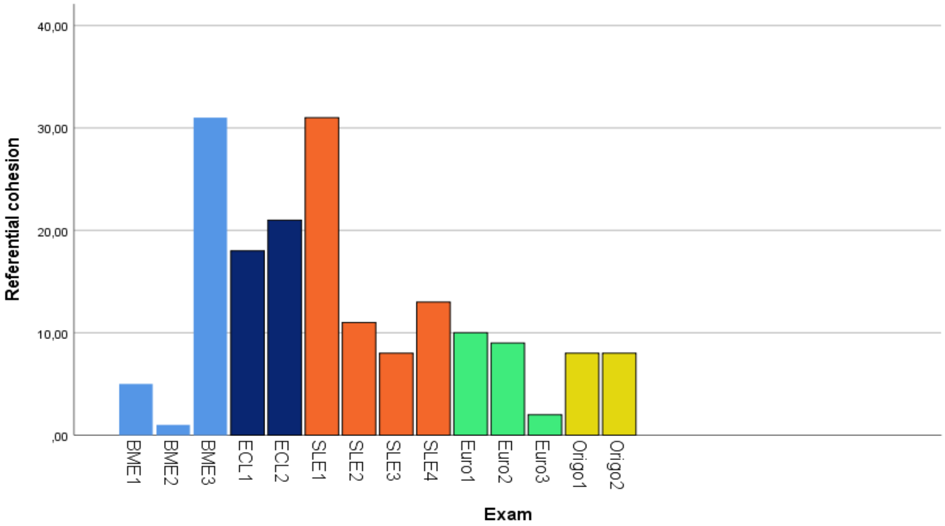


Figure 5. TERA results for "Referential cohesion"

The indication of the above is that the exam providers seem to effectively all agree that the lack of such overlaps and the resulting inferencing are hallmarks of comprehension at this level. Can this be justified in light of CEFR descriptors? Interestingly, though cohesion is clearly an important element contributing to comprehension, the CEFR scales related to reading make no mention of this property. While there is a "Cohesion and Coherence" scale in the CEFR, it refers exclusively to productive skills. In the CEFR's "Identifying Cues and Inferring" scale, the B2 descriptor says a learner at this level "can use a variety of strategies to achieve comprehension, including listening for main points; checking comprehension by using contextual clues" (Council of Europe, 2001, p. 72). Especially the use of "contextual clues" may be related to the kind of inferencing the TERA analysis refers to. Also, a new descriptor in the "Reading for Information and Argument" scale in the CV states that a B2 reader "can recognise different structures in discursive text: contrasting arguments, problem-solution presentation and cause-effect relationships" (Council of Europe, 2017, p. 62). The recognition of "different structures in discursive text" may also be interpreted as an element linked to referential cohesion. Despite the above, there appears to be no clear CEFR-based argument to justify such low levels of referential cohesion in the texts examined. Yet, somewhat paradoxically, it is this measure on which the texts appear to show most homogeneity.

The fifth measure applied was Deep cohesion. The results of the analysis are presented in Figure 6.

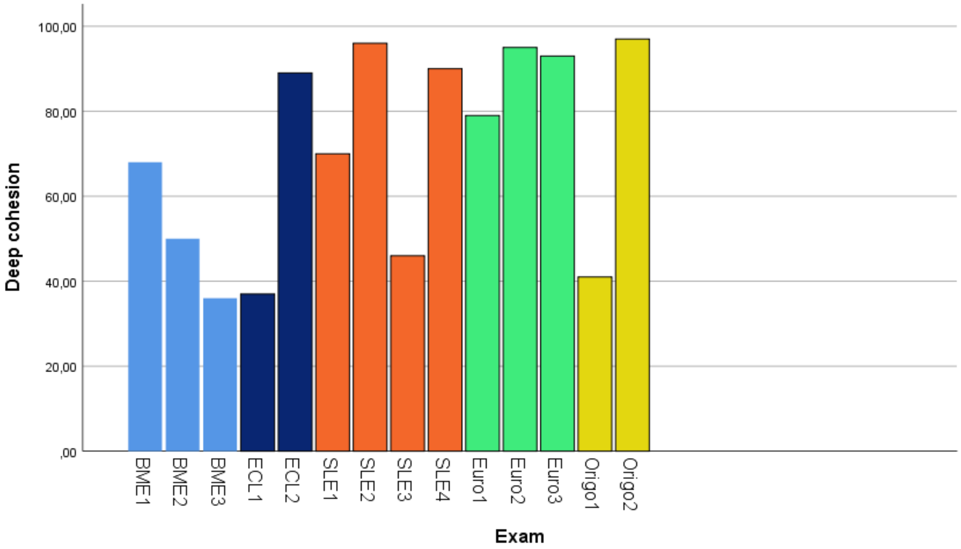


Figure 6. TERA results for "Deep cohesion"

The pattern that can be observed here is markedly different from that of the previous measure. Figures range from 36% (BME3) to 97% (Origo2). What this means is that seven texts (ECL2, SLE2, SLE4, Euro1, Euro2, Euro3 and Origo2) are considered to score high on deep cohesion, and the rest of the texts are in the average range. Thus, the texts seem to indicate that at this level the expectation is that learners will be able to handle texts where the logical connectedness of sentences is clearly marked. Finding justification for this assumption on the basis of the CEFR is, once again, challenging. For reasons discussed earlier, it is difficult to find descriptors that match the construct of deep cohesion. A descriptor that has already been quoted in relation to referential cohesion may be of some assistance though. According to this descriptor, found in the modified "Reading for Information and Argument" scale in the CV, a B2 reader "can recognise different structures in discursive text: contrasting arguments, problem-solution presentation and cause-effect relationships" (Council of Europe, 2017, p. 62). Concerning deep cohesion, it is probably the end of the descriptor than may be worth considering. "Cause-effect relationships" are clearly within the realm of deep cohesion, even if the rest of the descriptor presents no one-to-one match with it. Thus, similarly to referential cohesion, we find relative homogeneity across the texts without a solid CEFR-based argument to explain it. It is worth noting that this similarity is of particular significance, as it seems to suggest the relatedness of the two constructs as well. A similar tendency has been observed in earlier research of comparable focus (Szabó, 2014), which suggests that the relationship is an actually existing one.

The last measure employed was a specific L2 readability index, which is not part of TERA, but which a Coh-Metrix analysis still provides. The results are presented in Figure 7.

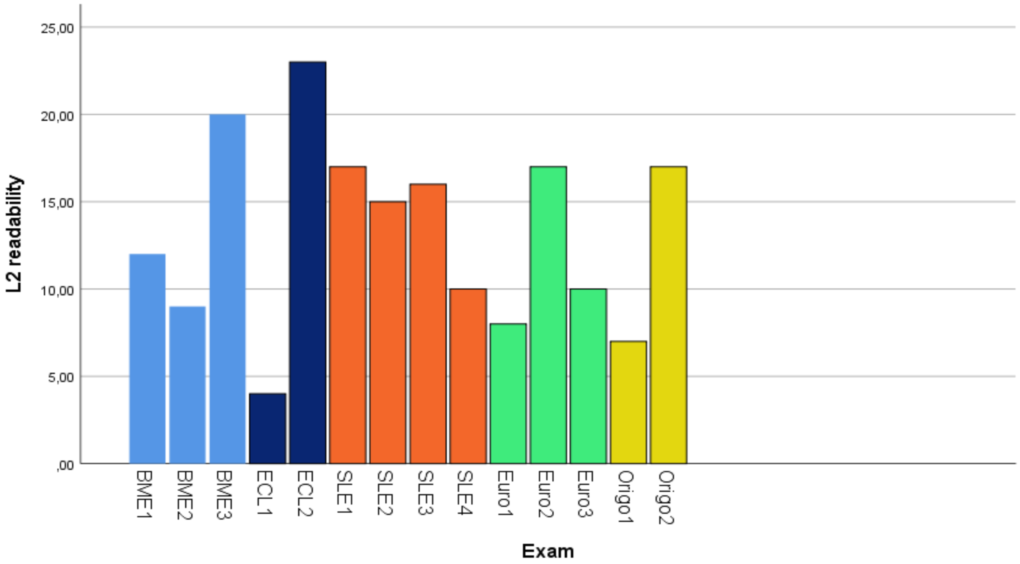
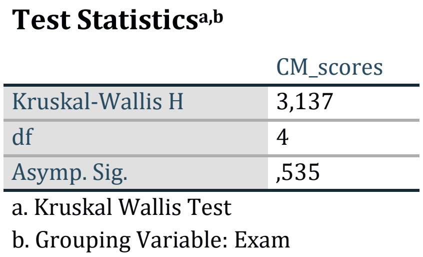


Figure 7. Coh-Metrix "L2 Readability" results

Once again, the scores for the different texts appear to be spread out noticeably, ranging from a score of 4 (ECL1) to 23 (ECL2). It is also worth noting how the three texts appearing to be the most similar (SLE1, Euro2, Origo2, all scoring 17), are so markedly different in light of the TERA measures. Indeed, the L2 readability index is reported to be a composite measure of a lexical, a syntactic and a meaning construction index (Crossley, Greenfield, & McNamara, 2008). Thus, it is likely that there may be different subscores for different texts, which may balance out at the level of the L2 readability index. What this suggests, in turn, is that the differences across the texts observed with respect to this measure are likely to stem from the combination of a variety of measures. Also, it needs to be emphasized that while the L2 readability index, as the term suggests, was developed with non-native readers in mind, the TERA measures are to be interpreted without reference to the reader's native language.

What has been discussed so far seems to point to the direction that the texts analyzed may well be significantly different from one another in terms of difficulty. In order to check this impression, an independent samples Kruskal-Wallis test was performed where the six measures the Coh-Metrix analysis yielded were interpreted as six facets of a common underlying variable, text difficulty. The analysis was performed both at the level of the exams, i.e. all texts associated with a particular exam were treated together as feeding into a common text difficulty measure for each exam, and at the level of the texts, where individual texts were treated separately, as if they had not been part of clusters of texts for individual exams. Figures 8 and 9 present the results of the Kruskal-Wallis tests.



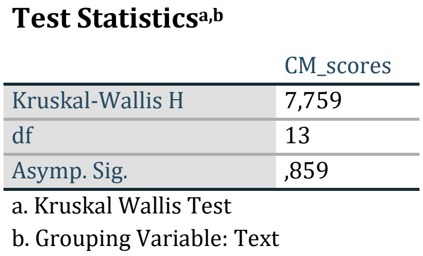
Test Statistics[a,b]

	CM_scores
Kruskal-Wallis H	3,137
df	4
Asymp. Sig.	,535

a. Kruskal Wallis Test

b. Grouping Variable: Exam

Figure 8. Kruskal-Wallis test results at exam level



Test Statistics[a,b]

	CM_scores
Kruskal-Wallis H	7,759
df	13
Asymp. Sig.	,859

a. Kruskal Wallis Test

b. Grouping Variable: Text

Figure 9. Kruskal-Wallis test results at test level

Somewhat surprisingly, the analyses in both cases indicated no significant differences. As can be seen in Figure 8, in the case of the exam-level analysis, it was found that the exams were not significantly different ($p \leq 0.535$), while the text-level analysis, presented in Figure 9, revealed that the individual texts were also not significantly different ($p \leq 0.859$). What could explain this result, seemingly contradictory to the observations made at the level of the individual measures? The answer, most probably, is quite simple. As has been discussed, individual text characteristics varied a great deal across the texts, but it seems they were acting in a compensatory manner. In other words, where certain texts appeared to be more difficult on one measure, they were less difficult on another. For instance, ECL1 scored low on Narrativity and Syntactic simplicity, but extremely high on

Word concreteness, and average on Referential cohesion and Deep cohesion, even if low on L2 readability. What this seems to imply is that the examinations used texts of differing characteristics, as suggested by the variety of results even within text clusters in terms of the individual measures, which, however, still result in the same overall level of text difficulty. As mentioned earlier, this may, in fact be a highly positive feature, as the variety of texts may well indicate a broader content domain sampling, and thus a higher degree of content validity.

4. Conclusion

In this paper we attempted to compare texts used in a variety of B2 level language examinations in Hungary by means of relying on text readability indices generated by the Coh-Metrix platform. As has been pointed out, the texts appeared to differ in terms of the individual measures, but showed no statistically significant differences when the measures were treated as facets of text difficulty, feeding into the same underlying construct. While some preliminary conclusions have already been drawn in the previous section, a few caveats need to be noted at this point. First, though the differences identified at the level of the individual measures may mean that the texts provide a broad content domain sampling, this is not necessarily so. Indeed, as was pointed out in the course of the analysis, the Coh-Metrix based results were frequently difficult to link to the proclaimed CEFR-based content domain of B2 level reading ability. There are indications of the links between the measures used in this analysis and the CEFR-based content domain, but these links would need to be further confirmed through different types of analyses as well.

Second, it needs to be noted that while the findings may be interpreted as reflections of good practice on the part of the exam providers, the data are not sufficient to serve as the basis for such claims. On the one hand, in all cases except for the SLE, the texts examined came from sample materials, not live tests. Thus, they are not necessarily indicative of regular text selection practices. Moreover, even in the case of the school-leaving exam the texts were selected from one particular exam period (May, 2018), which means we have no indication of how stable the measures may be in comparison to other exam periods.

Third, it needs to be emphasized that, as was mentioned earlier, this study did not intend to create a correspondence between Coh-Metrix based scores and CEFR descriptors. It did reveal, however, that this issue would be worthy of further investigation. While some of the criteria along which differentiation is defined in CEFR terms seem to fall outside the scope of Coh-Metrix (e.g. text length), others (e.g. sentence complexity) could be explored further in order to provide support for linking tests to the CEFR.

Most importantly, however, we need to emphasize that text difficulty does not equal task difficulty. As is commonly pointed out (e.g.: Castello, 2008, p. 16), task difficulty in reading is a multi-faceted concept, determined by the combined effect of the text, the task and even the reader. Even if we leave out the reader (as an ever changing parameter) from this equation, we still need to make it clear that in any examination the supposed level of the texts involved is only part of the story. Thus, while the results of the Kruskal-Wallis tests seem to indicate that the texts may not have differed in terms of difficulty, this does not mean that the level of the tasks was the same as well. Since all the examinations discussed in this study claim to test the same level, we may presume they are at the same level. The current study, however, provides partial evidence at best to support this claim. Yet, perhaps this study, along with other similar ones can contribute to a better understanding of how tests of reading comprehension work, and more extensive examinations of texts and tasks can help build a better validity argument resulting in better tests. This, I believe, is a purpose worth working for.

5. References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press

Brown, J. D. (1998). An EFL readability index. *JALT Journal*, 20(2), 7-36

Castello, E. (2008). *Text complexity and reading comprehension tests*. Bern: Peter Lang

Council of Europe. (2001). *Common European Framework of Reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for languages: learning, teaching, assessment (CEFR). A manual*. Strasbourg: Language Policy Division

Council of Europe. (2017). *Common European Framework of Reference for languages: learning, teaching, assessment*. *Companion volume with new descriptors*. Provisional edition. Strasbourg: Council of Europe

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1), 84-102.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3), 475–493.

Educational Authority. (2018). *Accreditation handbook*. Retrieved September 27, 2018, from https://nyak.oh.gov.hu/nyat/ah2018-eng.asp

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3, 371-398.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, 40(5), 223–234.

Halliday, M. A. K. (1978). *Language as social semiotic. The social interpretation of language and meaning*. London: Edward Arnold

Harsch, C., & Hartig, J. (2015) What Are We Aligning Tests to When We Report Test

Alignment to the CEFR?, *Language Assessment Quarterly*, 12(4), 333-362.

Jackson, G.T., Allen, L.K, & McNamara, D.S. (2017). Common Core TERA: Text Ease and Readability Assessor. In Crossley, S.A., & McNamara, D.S. (eds.), *Adaptive Educational Technologies for Literacy Instruction*. (pp. 49-68). New York: Routledge.

Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly*, 10, 62–102.

Martyniuk, W. (ed.) (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual.* Studies in Language Testing 33. Cambridge: Cambridge University Press

McNamara, D.S., Graesser, A.C., Cai, Z., & Kulikowich, J.M. (2011). *Coh-metrix easability components: Aligning text difficulty with theories of text comprehension.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2013). *Coh-Metrix version 3.0*. Retrieved June 6, 2018, from http://cohmetrix.com

Szabó, G. (2014). Applying objective measures of text difficulty in the comparison of texts in reading comprehension tasks. In J. Horváth, & P. Medgyes (Eds.), *Studies in honour of Marianne Nikolov* (pp. 385-398). Pécs: Lingua Franca Csoport.

Szabó, G., & Kiszely Z. (2010). Államilag elismert nyelvvizsgarendszerek illetve az emelt szintű érettségi összevetése próbavizsgázói teljesítmények tükrében német és angol nyelvből. *Modern Nyelvoktatás*. 16(4), 19-38.

TERA text comparison (2018). Retrieved June 6, 2018, from http://129.219.222.66:8084/ Grid/Coh-MetrixCompare.aspx

Vinther, J. (2013). CEFR - in a critical light. In J. Colpaert, Simons, Mathea, Aerts, Ann, and Oberhofer, Margret (Ed.), *Language Testing in Europe: Time for a New Framework?* Proceedings, pp. 242-247. Antwerp: University of Antwerp.