HUMAN, TECHNOLOGIES AND QUALITY OF EDUCATION, 2024
https://doi.org/10.22364/htqe.2024.27

COMPARING THE DIFFICULTY OF TEXTS USED IN LATVIAN AND HUNGARIAN SCHOOL-LEAVING EXAMINATIONS IN ENGLISH AS A FOREIGN LANGUAGE

Gábor Szabó[1]

[1] University of Pécs, Hungary

ABSTRACT

Examinations in public education are supposed to guarantee that students' performances are measured using the same standards from exam period to exam period. One important element of guaranteeing the same standards is the use of materials that present the same degree of challenge to candidates in different examinations. In the assessment of foreign language proficiency, such materials primarily include target language texts.

This study offers a comparative analysis of the difficulty of texts used in the reading comprehension section of the school-leaving examinations of the past five years in English as a foreign language in Latvia, along with a comparative analysis of texts used for a similar purpose in Hungary in the same time period. The study examines how stable the difficulty of texts has been across the years, with special regard to the recent changes in the structure of the Latvian exam, along with a comparison of text difficulties and intended exam levels across the two countries. Text difficulty is examined in light of the Coh-Metrix Common Core Text Ease and Readability Assessor (TERA) measures, and similarities and differences are measured by means of applying statistical tests for significant differences. Findings indicate that text clusters from exams targeting the same level showed no significant differences in text difficulty measures; however, in the majority of cases, text clusters from exams targeting different levels showed no significant differences in text difficulty measures either. Thus, further research is needed in order to establish whether the required differences in difficulty were appropriately present in terms of the tasks to be performed by candidates.

Keywords: *school-leaving exam, language assessment, text difficulty, foreign language education, maintaining standards.*

Background

Assessment of learning outcomes is considered to be a critical element of measuring not only of learner achievement, but also of the effectiveness of educational systems.

Consequently, school-leaving exams have a particularly important role in feeding back to both learners and educational decision makers. Considering the importance attributed to foreign language education, school-leaving examinations testing language proficiency are of particular significance these days. It follows from this that guaranteeing the standards of such examinations is also of paramount importance. This is all the more so, as most such examinations tend to identify the level(s) they measure in relation to the Common European Framework of Reference (Council of Europe, 2001), and, more recently, to its Companion Volume (Council of Europe, 2020). Accordingly, comparability is a significant concern, and the most important issue appears to be whether examinations are successfully aligned with the CEFR levels (cf. Harsch & Hartig, 2015; Martyniuk, 2010). While detailed guidelines concerning the procedures of alignment have been available for some fifteen years in the form of the Manual for Relating Examinations to the CEFR (Council of Europe, 2009), doubts about how much exams supposedly aligned with the CEFR are actually comparable have not been dispelled entirely (e.g., Vinther, 2013). Moreover, it has also been argued that the actual alignment may well be the function of local interpretations of CEFR descriptors (e.g., Brunfaut & Harding, 2019).

In light of the above, it appears to be necessary to identify aspects of examinations which make it possible to compare exams across different exam periods as well as, potentially, across different contexts and educational systems. One such aspect that seems readily available is the difficulty of the target language texts that students are expected to understand in tests of comprehension.

Text difficulty in reading comprehension tests

When constructing tests of reading comprehension, it is essential to select texts which meet criteria established in accordance with the requirements of the specific context of assessment. Among many others, one such criterion is the appropriate level of the text. Indeed, text level may well be among the most important properties of a text used in an assessment task. Yet, when taking a closer look, it becomes clear that defining what exactly the term "level" means in this context can be quite challenging. Generally, the concept of text level tends to be first and foremost interpreted as a measure of how easy or difficult it is to understand what the text means. Meaning, nevertheless, is a concept not as straightforward as one may assume. It could be argued, for instance, that, instead of having meaning as such, texts merely have meaning potential (Halliday, 1978). In other words, any text will offer potentially different interpretations, and the actual meaning is rather a function of the reader than of the text itself. As Alderson points out, readers' knowledge and experiences influence how meaning potential is realized, and as their knowledge and experiences differ, so do the interpretations of a text (Alderson, 2000, p. 6). Accordingly, if interpretations differ, so do the elements of what may make the text easy or difficult to understand. Thus, one may claim text level as such does not exist; texts simply have specific characteristics. Though this may seem quite an extreme conclusion, it amply demonstrates the potential difficulties of determining the level of a text. Apart from reader involvement, however, there is another reason why text level

may be problematic to determine: it is possible to understand texts at different levels depending on how much detail or what degree of interpretation is called for. Indeed, approaching the concept of text difficulty in this way is justified by the fact that CEFR descriptors often make a difference between different levels based on whether it is only the main points or also the supporting details that are understood. Moreover, at higher levels, it is expected that readers should be able to understand implicit messages as well. (cf. Council of Europe, 2001).

In light of the above, it seems safe to claim that the concept of text level is interpreted in different ways. Despite these theoretical discrepancies, however, numerous attempts have been made to quantify text difficulty in a way that results in the production of some measure which could express difficulty and, in turn, text level. Predominantly, these measures have been readability indices. Arguably, the Flesch Reading Ease and the Flesch-Kincaid Grade Level indices are the best known such readability measures. Both of these indices are grounded on a relationship hypothesized to exist between the number of words, the number of sentences and the number of syllables (Klare, 1974–1975). While these measures have been used widely, several scholars (e.g., Alderson, 2000; Brown, 1998) have criticized them for being too simplistic, particularly in a second language context.

In response to the perceived inadequacy of traditional measures, more complex indices have also been developed. One notable example of the more modern and more elaborate measures is the Coh-Metrix readability formula (Graesser, McNamara, & Kulikowich, 2011; McNamara et al., 2014). Coh-Metrix provides a description of text characteristics in light of 53 measures. Of course, such an immense number of characteristics could potentially make it problematic to interpret the figures in a practical manner. In order to solve this problem, principle component analysis has been applied, the purpose of which was 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. Next the eight principal components were mapped to a five-level theoretical model presented 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). As a result, the original Coh-Metrix measures can now be expressed along the five dimensions of the model, making the interpretation of the results simpler and more transparent.

As Coh-Metrix provides a combination of numerous facets of text difficulty, it appears to be a suitable tool for establishing the level of various texts in a multitude of different contexts. Accordingly, in the following Coh-Metrix measures will be used in comparative analyses in order to determine whether there are differences in the levels of a set of texts under scrutiny. The texts came from the reading comprehension sections of the school leaving examinations in English as a foreign language in Latvia and Hungary between 2019 and 2023.

Methodology

The focus of the study was to examine to what extent texts used in examinations targeting the same level were similar in terms of difficulty, but also to study whether examinations geared toward different levels used texts that differed significantly in terms of difficulty. Thus, on the one hand, the purpose was to check whether examination targeting the same level used texts of the same difficulty in a consistent manner and, on the other hand, whether the texts used in the Latvian and the Hungarian contexts showed any significant differences. In order to do so, texts were first collected from publicly available past exam papers from the reading comprehension sections of the school leaving examinations in English as a foreign language in Latvia and Hungary between 2019 and 2023.

It is worth noting here that the two countries' examination systems differ. In Latvia, up to 2021 only one exam was offered per exam period, although these examinations were designed to cover three levels, CEFR B1, B2, and C1 ("General secondary education in Latvia," 2020). From 2022, however, a new system was introduced in which students can either take the "optimal" level exam, targeting CEFR B2, or the "advanced" level exam, geared toward CEFR C1 ("Svešvaloda [angļu, franču, vācu, krievu] augstākais mācību satura apguves līmenis," 2024). In contrast to this, in the Hungarian system, which remained unchanged during the period under scrutiny, students could choose between the "intermediate" level exam, targeting CEFR B1 and the "advanced" level exam geared toward CEFR B2 ("Élő idegen nyelv," 2021). In order to guarantee representativeness, texts from all levels and exam types have been used in the analysis.

One more issue to be clarified about the texts themselves is related to the task types. In some cases texts were presented in the tasks in their complete form (e.g., in the case of multiple choice tasks), while in other cases they were not. In the latter case (e.g., in gap-filing tasks) before analysis the texts were reconstructed to have their original, complete form. The reason for this was to guarantee that texts are comparable across task types. In some cases tasks included multiple unrelated texts. Whenever this happened, these unrelated texts were analyzed as separate texts. In the course of the analysis a total number of 82 texts were analyzed, 42 of which came from Latvian, while 40 from Hungarian exams.

In conducting the study, texts were analyzed using the *Coh-Metrix Common Core Text Ease and Readability Assessor* (Jackson, Allen, & McNamara, 2017), also known as TERA. It is accessible as a web tool available online (https://soletlab.adaptiveliteracy.com:8443/). This web tool provides the following measures related to text difficulty and readability:

- narrativity
- syntactic simplicity
- word concreteness
- referential cohesion
- deep cohesion.

The results are expressed in percentile figures. A sample TERA output is presented in Figure 1.

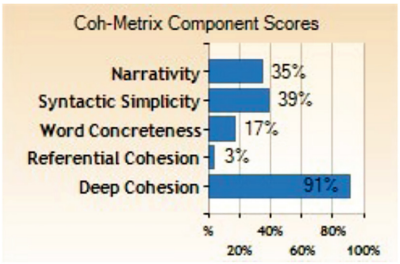

Figure 1 Sample TERA output

Narrativity is to be interpreted as a continuum stretching between texts whose nature is highly narrative, and which are thus thought to be easier to process, and informational texts, which present a higher degree of difficulty in understanding. Narrative texts are made up of a high proportion of high frequency words including easy-to-understand verbs as well as pronouns that make texts more engaging for readers (Jackson et al., 2017, p. 55)

Syntactic simplicity is the reflection of sentence complexity observable in the text. This index is derived from several measures of syntactic complexity, which include the number of clauses and the number of words in a sentence, along with the number of words before the main clause. The measure also accounts for the potentially observable similarities in sentence construction across paragraphs (Jackson et al., 2017, p. 56).

Word concreteness is defined as the proportion of abstract and concrete words included in the text. Abstract words are considered to make comprehension more difficult; consequently, a text in which the proportion of concrete words is large is believed to pose less of a challenge and is thus easier to understand (Jackson et al., 2017, p. 57).

Referential cohesion is characterized by the potential occurrence of overlaps between words, word stems and concepts from one sentence to the next. A high proportion of overlaps is considered to be an indicator of the text being easier to comprehend (Jackson et al., 2017, p. 57).

Deep cohesion is defined in terms of the number of connectives in the text, representing how much the events described or the variety of bits of information presented in the text are tied together. A high number of connectives is interpreted as an indication of stronger links, which make comprehension easier (Jackson et al., 2017, p. 58).

Once the measures 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 significant differences both at the level of the TERA component scores and at the level of the composite of TERA measures. The rationale for the latter procedure was that, as all the component readability measures discussed earlier feed into the same construct, they may legitimately be considered as different facets of the same property of a text (i.e., difficulty). Thus, all the indices can be treated as scores. It needs to be noted, however, that these scores can only be conceptualized as defining an ordinal rather than an interval scale.

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, depending on the number of variables to be compared, independent samples Kruskal-Wallis tests or Mann-Whitney's U tests were run.

At this point it is important to clarify that, although the focus was on the levels of texts, this study did not intend to map Coh-Metrix scores on the CEFR or vice versa. Though references are made in the discussion to how TERA measures may be explained in terms of CEFR descriptors, this is only done in a tentative manner. The reason for this is that Coh-Metrix produces quantitative measures, while CEFR descriptors are qualitative in nature. In other words, they approach the concept of difficulty from different perspectives: Coh-Metrix focuses on objectively measurable text properties, while the CEFR descriptors attempt to capture what learners can do at particular levels. Accordingly, the purpose of the study is rather to examine the texts that have most probably been chosen with CEFR levels in mind using a variety of objective measures of text difficulty to see how they compare.

Results

In discussing the results of the study, first the comparison of TERA component scores within exams will be presented.

Figure 2 presents the measures for *Narrativity* in the Latvian exams. The size of the bars indicates the values of *Narrativity* in the individual texts. Exam tasks are marked with a number from 1 to 3. Letters refer to different texts found within the same exam task. The color coding identifies the different exam versions. Exam versions between 2019 and 2021 are marked with the indication of the years, while exams in 2022 and 2023 are also marked A ("advanced") and O ("optimal") according to level. As can be observed, actual measures varied greatly across the texts examined both within and across exam versions.

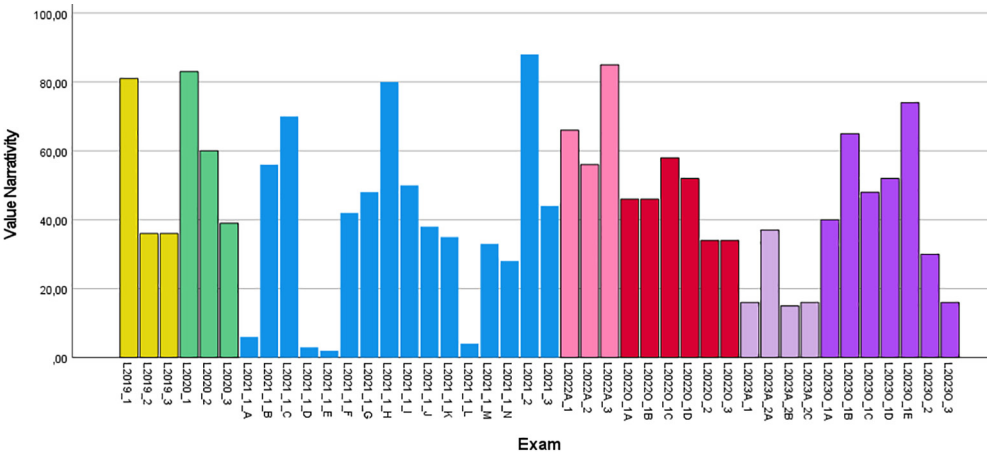

Figure 2 Narrativity measures in Latvian exams

This should come as no surprise, on the one hand, as texts in an exam are meant to be varied in terms of how story-like they are. The analysis also revealed, however, that, when different exam versions were compared, there were no significant differences across them, regardless of what level they were targeting.

Interestingly, a similar pattern can be observed in terms of the other TERA components as well. Individual texts varied a great deal in terms of the values of TERA measures, but these differences were not statistically significant. Next it was examined whether any significant differences can be observed when comparing the three different exam types, i.e., the old exam, targeting three levels, and the two new exams focusing on one specified level each.

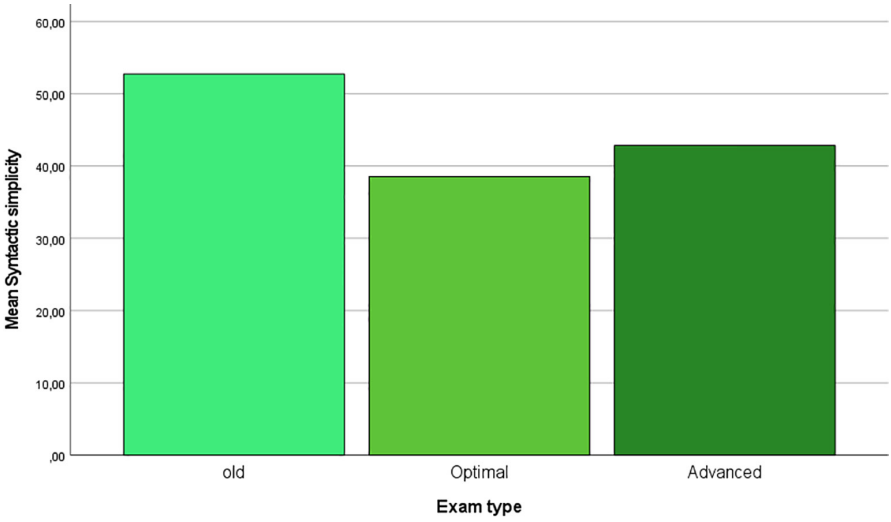

Figure 3 Mean Syntactic simplicity measures for Latvian exams by exam type

Figure 3 provides an illustration to examining TERA component measures according to exam types. In Figure 3 mean measures for *Syntactic simplicity* are presented across the three Latvian exam types. As can be observed, the three exam types appear to be different, although these differences do not seem to mirror the intended levels of difficulty. More importantly, however, these apparent differences were not statistically significant. Once again, a similar phenomenon could be observed in the case of the rest of the TERA components, too. While measures varied somewhat across the exam types, they were not found to be significantly different.

Concerning the Hungarian examinations and the texts therein, a similar tendency could be observed. Texts tended to be varied with respect to the TERA component measures, but they were not significantly different. One noteworthy exception is presented in Figure 4.

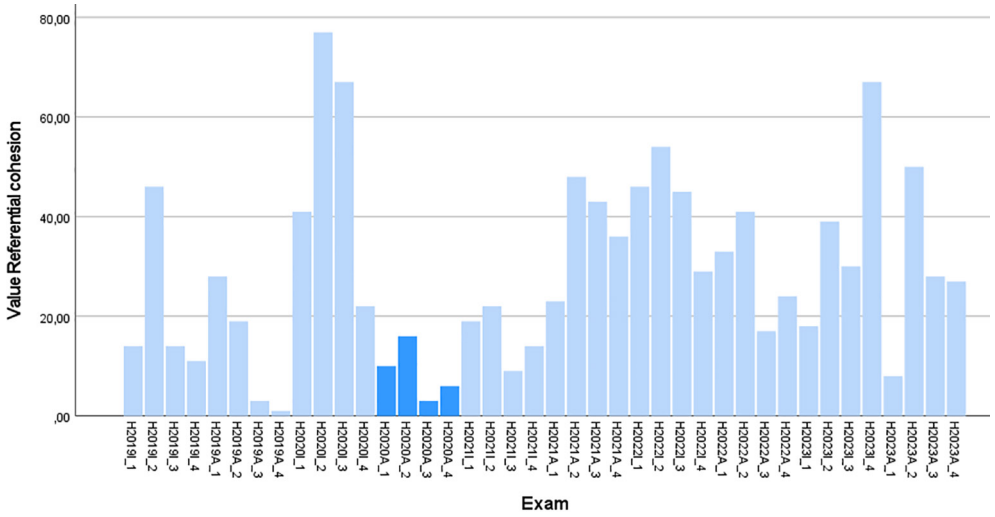

Figure 4 Referential cohesion measures in Hungarian exams

In the case of this measure texts, again, show considerable variation, and the differences in this case are close to being significant ($p = 0.011$), mainly owing to the texts used in the 2020 advanced level exam, highlighted in Figure 4, showing an apparent consistency in a low level of referential cohesion. Yet, at the level of exam types, the difference, again, is not significant.

As to the comparison of exam types in general, similarly to the Latvian exams, TERA measures tended to show no significant differences between the two types of exams. Again, there is a case of a nearly significant difference as presented in Figure 5.

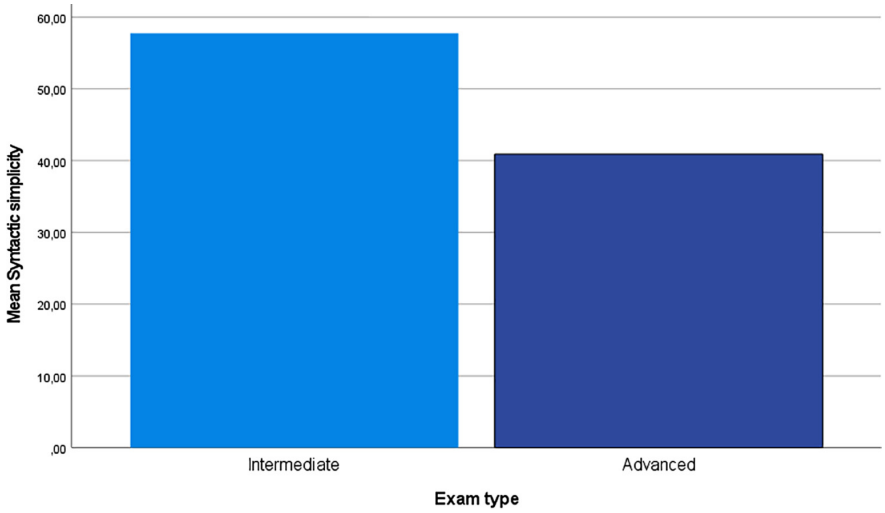

Figure 5 Mean Syntactic simplicity measures for Hungarian exams by exam type

In this case the difference in Syntactic simplicity measures is apparent, and is, indeed, close to being statistically significant ($p = 0.014$).

Finally, it was examined whether the TERA measures used as facets of a composite readability measure can be shown to be different across examinations. Initially, it was checked whether a difference can be detected between all the texts used in the examinations of two countries. Figure 6 presents the comparison in a graphical way.

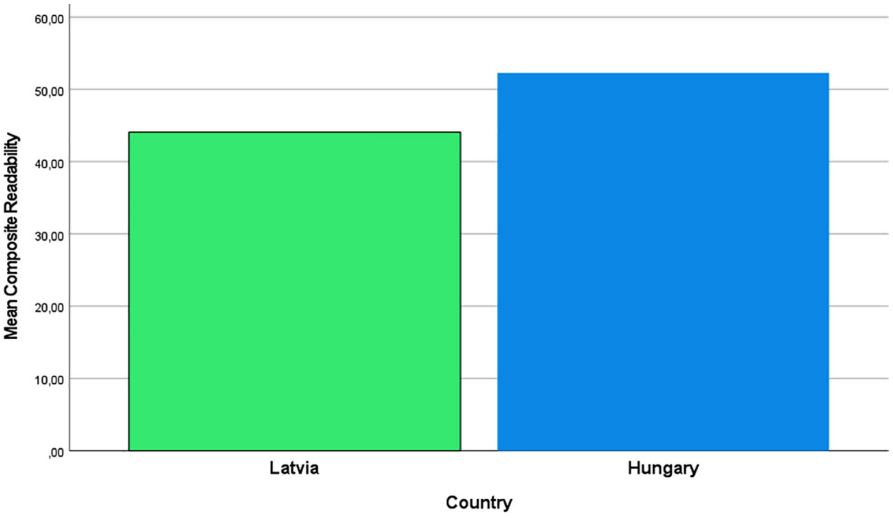

Figure 6 Mean composite readability figures of all texts in Latvia and Hungary

As can be observed, the composite readability of the texts used in Hungarian exams appears to be notably higher. Indeed, this difference was found to be statistically significant ($p = 0.002$). While this is an interesting finding, it should not be surprising, considering the fact that some of the texts in the Latvian exams were meant to be used in tasks targeting level C1, while no tasks meant to target this level in the Hungarian exams. Similarly, while half of all texts in the Hungarian exams were meant to be used in tasks targeting level B1, there was only a small minority of such texts in the Latvian exams. Thus, the overall difference being significant should hardly be unexpected. But such overall comparisons may mask all sorts of differences. It is perhaps more intriguing to compare the level-specific exams in both countries in terms of the overall readability measures. This comparison is presented in a graphical format in Figure 7.

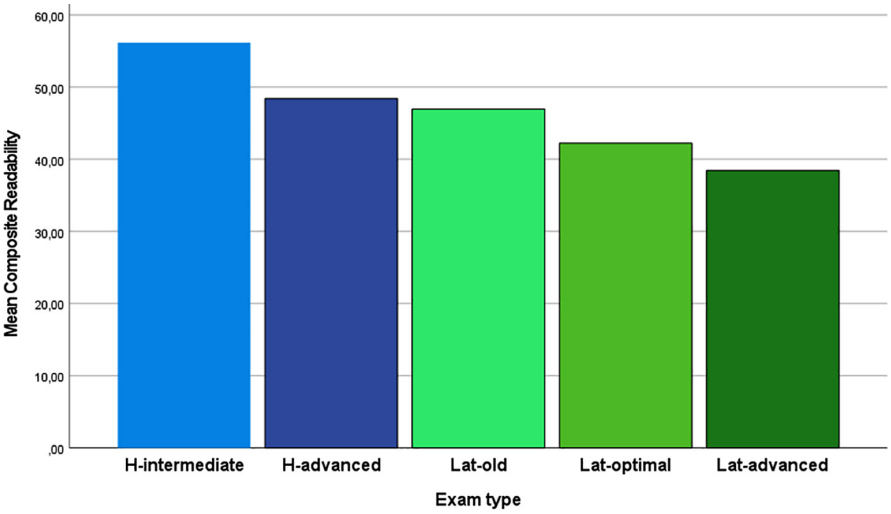

Figure 7 Mean composite readability figures of all texts by exam type

The composite readability figures appear to show a neat progression from the exam declared to target the lowest level (B1) to the one aiming at the highest level (C1). This impression may be somewhat deceptive, however, considering the fact that the "old" Latvian exam was meant to target three levels, and that the Hungarian "advanced" level exam is geared toward the same level (B2) as the Latvian "optimal" level exam. Nevertheless, more important than impression are the checks for statical significance of differences. As the results reveal, the composite readability of texts used in the Hungarian "intermediate" exams is significantly different from those in the Latvian "optimal" and "advanced" level exams ($p < 0.001$ in both cases), and the difference is nearly significant in the case of the "old" Latvian exam ($p = 0.016$). On the other hand, no other differences are significant, which raises interesting questions about the level of texts used in exams meant to target different levels.

Discussion

In light of the results, it is safe to make a few observations concerning the level of the texts used. First, it seems clear that the different versions of the exams have proven to be quite similar in terms of the various TERA measures, and thus in terms of the difficulty of the texts used. This is so even if the measures of the actual TERA components fluctuated considerably. Indeed, it seems this fluctuation should rather be interpreted as an indication of the variety found in the texts, which is often seen as a way to provide a sufficiently varied sampling of the targeted construct. The lack of significant differences is certainly a positive feature here, as it demonstrates that at least one element of the ways in which the consistency of exam levels can be guaranteed is, indeed, unproblematic.

Similarly, the second observation, namely that there are no significant differences between texts used in exams targeting the same level is also in line with expectations. Indeed, this appears to be an indication that exams in the two different countries seem to employ similar texts for the same purpose.

More puzzling, however, is the third observation concerning the lack of significant differences across exams targeting different levels. Indeed, as the results reveal, such differences do not appear to exist either in terms of individual TERA measures, whether examined by exam versions or by exam types, or in terms of composite readability figures, at least in the majority of the cases. While the texts used in the Hungarian "intermediate" exam seem to be significantly different from two of the Latvian exams with regard to composite readability, no such differences can be observed between the two Hungarian exams targeting different levels or across the three Latvian exams, also geared toward different levels. The case of the "old" Latvian exam is the least enigmatic, of course, as this is the only exam declared to have targeted all three levels in question, so the lack of significant differences may be explained on grounds that parts of the exam must necessarily have included texts to cater for the needs of candidates at various levels. In the case of the other exams, however, the apparent lack of significant differences seems to suggest that for testing different levels of reading comprehension texts of very similar properties have been used.

At this point it is important not to jump to false conclusions. While it may be tempting to suggest that the exams targeting different levels are, in fact, at the same level, this would not at all be justified in light of the results. It needs to be remembered that the difficulty of a reading task is only partly determined by the difficulty of the text. Indeed, it is quite possible to potentially use the same text at different levels, if the focus of the items differs in the two tasks. This, of course, is in line with the point made earlier in this paper about how texts do not really have levels, because they may be understood at different levels, depending on the depth of understanding required. Accordingly, the similarity of text characteristics tells us nothing about the nature of the tasks used in the examinations. Thus, without examining and comparing the actual tasks themselves, no opinions can be formed about the level of the exams, either.

It also needs to be remembered that, while the TERA measures provide an impressive account of text characteristics, they offer no information on a text feature that is of outstanding significance when it comes to determining level appropriacy: the topic of the text. Indeed, when it comes to CEFR and CV descriptors, one important way in which reading performances are differentiated according to level is related to the topic of the texts. It is quite possible that the texts from examinations at different levels analyzed in this study differ considerably in terms of topic. It should be noted, however, that classifying the topics would necessarily include some degree of subjective judgement, which is one reason why the topics of the texts analyzed were not examined in this study.

Conclusions

This paper intended to provide an insight into a study comparing the difficulty of texts used in Latvian and Hungarian school-leaving examinations in English as a foreign language. The findings indicate that, in the majority of cases, text properties were not significantly different in the case of examinations targeting the same levels, and the same is true of exams targeting different levels. On the one hand, this indicates a welcome consistency of difficulty in texts used in exams targeting the same level, but it also suggests a potential challenge related to texts in exams focusing on different levels.

As has been discussed, there are different potential explanations to why such results emerged. In order to clarify the matter, it would seem useful to conduct further research into both the properties of texts inaccessible to automatized text analysis, as well as into the qualities of the tasks in which the texts were used.

A further issue that could be raised is how TERA analyses may be utilized in the future for test construction purposes. In most examinations text selection is likely to be conducted on the basis of a subjective evaluation of text properties. As the current study has demonstrated, the process of text selection could also be made more efficient by employing automated analyses of potential texts. It has been shown that in the case of certain TERA measures exam texts displayed remarkable variety, which may be a positive feature in some cases (e.g., *Narrativity*), but which may also raise issues of standardization in other cases (e.g., *Syntactic simplicity*). By using TERA analyses on a regular basis, texts could be monitored for important characteristics influencing text difficulty, which would likely make the levels of the texts, and thus of the examinations themselves, even more stable and justified.

As has been demonstrated, an objective analysis of text properties provides an insight into certain elements of text difficulty. By employing such analyses, the transparency of national examinations could also be increased, which would contribute to a greater acceptance of test results, a goal all test providers seek to achieve.

Author note

The publication of this paper was made possible by the generous support of the Faculty of Humanities and Social Sciences of the University of Pécs.

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.

Brunfaut, T., & Harding, L. (2019). International language proficiency standards in the local context: Interpreting the CEFR in standard setting for exam reform in Luxembourg. *Assessment in Education: Principles, Policy & Practice*, *27*(2), 215–231.

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. (2020). *Common European Framework of Reference for Languages: learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Council of Europe.

Élő idegen nyelv. Részletes érettségi vizsgakövetelmény [Modern languages. Detailed exam requirements for the school-leaving exam]. (2021, July 16). Retrieved April 25, 2024, from https://www.oktatas.hu/pub_bin/dload/kozoktatas/erettsegi/vizsgakovetelmenyek2024/elo_id_nyelv_2024_e.pdf

General secondary education in Latvia (2020, June 26). Retrieved April 25, 2024, from https://www.visc.gov.lv/en/examinations/gse_in_latvia1.pdf

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., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Svešvaloda (angļu, franču, vācu, krievu) augstākais mācību satura apguves līmenis [The highest level of acquisition of study content in foreign languages (English, French, German, Russian)]. (2024, April 16). Retrieved April 25, 2024, from https://www.visc.gov.lv/sites/visc/files/media_file/pr_svesvaloda_al_2024.pdf

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*, 242–247. Antwerp: University of Antwerp.