THE IMPACT OF CHANGING WRITING TASK WORD LIMITS ON WRITTEN PERFORMANCE

Gábor Szabó
Department of English Applied Linguistics
University of Pécs
szabo.gabor2@pte.hu

Abstract: This paper investigates whether changing word limits in writing tasks has any measurable impact on the qualities of candidate performances in the framework of language examination tasks. In order to examine the effects of planned modifications in the length of writing tasks in ECL language examinations, a study was conducted to analyze the properties of exam performances of differing lengths. In order to acquire objective measures of the text characteristics, the Coh-Metrix TERA platform and the Coh-Metrix L2 readability index were utilized. First, texts of differing lengths were analyzed, and next the text characteristics were checked for significant differences. The results of the analyses indicated that the change in word limits did not result in significant changes in any of the text properties. Thus, the modifications in the examination would not result in any undesirable consequences concerning the standards, construct, or validity of the exam.

Keywords: language testing, writing, text difficulty, word limits

1 Introduction

The testing of writing ability has long been an emphatic part of measuring foreign language proficiency. Such tests have traditionally focused on reproducing real-life writing tasks, such as writing letters, postcards, essays, or the like (see e.g. Hughes, 2003). In recent years, however, written communication seems to have changed fundamentally. Instead of letters, emails are sent; writing postcards has essentially been made obsolete by the sending of instant messages and images. Written communication, in general, appears to have shifted to an online environment, where texts tend to be shorter, their structure seems less orderly, and rules are applied less strictly or are downright ignored. Language learners, in turn, are also likely to be involved primarily in online forms of written communication (Chun et al. 2016). It follows from this that when written language is to be assessed, tasks need to be designed with such changes in mind.

New tasks, however, mean new challenges as well. Changing task characteristics may indicate a change in the construct measured, have implications for test level, or even result in a threat to the validity of the new tasks. To avert such threats, it is necessary to study how specific changes in task properties affect candidate performances, and whether the changes pose a genuine threat. This paper does not attempt to provide a comprehensive account of all possible task properties along with the implication of changes to them. It does, however, present a study in which a particular task characteristic – the required length of responses – is investigated.

2 Context

The European Consortium for the Certificate of Attainment in Modern Languages (ECL) is an exam provider offering language examinations in a total of 15 languages. ECL has been an active participant contributing to the construction of both the CEFR (Council of Europe, 2001) and its Companion Volume (Council of Europe, 2018), indicating that ECL is committed to quality and professionalism. It is this commitment that inspired the revision of ECL writing tasks, the purpose of which was to guarantee a more authentic sampling of test takers' proficiency. In the course of the revision, the length of writing tasks was intended to be changed. The rationale for this was that research indicates language learners tend to be more effective in written communication when it happens in informal online settings (Smith et al., 2017), where usually relatively short texts are constructed. Thus, in an attempt to set more authentic and, thus, more valid tasks, it seemed logical to shorten the required length of texts to be constructed by candidates.

The shortening of texts, however, raised three specific concerns regarding the quality of the performances. First, if candidates are required to produce shorter texts, then, potentially, completing the tasks may become easier, which would have implications concerning the level of the tasks, and, in turn, of the exams. This is a concern stemming from the fact that one criterion along the lines of which CEFR and Companion Volume descriptors make a difference across levels of performances is length.

Second, if candidates produce shorter texts, they may also change the structure of their performances, which may lead to more loosely connected sentences, fewer cohesive devices, and less cohesion in general. Clearly, if this occurs, it would imply structural changes in performances implying that the construct measured would no longer be the same.

Third, shorter texts would, technically speaking, indicate a smaller sample taken from the candidates' writing ability, and a smaller sample, in turn, raises the issue of whether this sample is sufficient to be considered representative. As content validity is typically defined in terms of how much test content is representative of the ability measured (Davies et al., 1999), a shorter text could, in principle, mean that the test's content validity is no longer guaranteed.

In order to examine whether these potential problems actually materialize, a study was designed to determine whether features of writing performances produced according to the original and the modified task requirements differ in any measurable way, other than in terms of length. In order to do this, however, a research design was needed in which objective measures of text properties could be compared to guarantee that the comparison was not based on human judgment. The basis for this approach was the assumption that text properties could be measured and analyzed in an objective manner, much like when text properties are examined in order to determine text difficulty or readability.

3 Measuring text properties

Historically, text properties have been approached from the perspective of meaning, though what exactly influences meaning has been a matter of debate. Halliday (1978), for instance, does not acknowledge that texts have meaning, *per se*; rather, he argues texts merely have meaning potential, which, in turn, is realized by different readers in different ways. It has even been argued that a text may be interpreted in a unique way by each reader (Alderson, 2000). While one may agree with the idea of potentially different interpretations of a text, if this idea is taken to its logical conclusion, one would need to believe that effective communication in writing is not possible.

Instead of adopting this rather extreme view, it seems more beneficial to attempt to identify aspects of texts that can be measured, and by means of which it is possible to describe text characteristics, which define how written communication works. This is exactly what a variety of text readability measures have attempted to achieve. Perhaps the two best known readability formulas are the Flesch Reading Ease and the Flesch-Kincaid Grade Level indices. Both of these are based on a supposed relationship between the number of words, the number of sentences and the number of syllables (Klare, 1974-1975). These indices, however, have been criticized by many (see e.g., Alderson, 2000; Brown, 1998), claiming that they are far too simplistic in their approach.

As a result, more sophisticated models have been developed, which approach text properties in a more complex manner, providing a more accurate depiction of texts themselves. An outstanding example of the more recent models is the Coh-Metrix readability formula (Graesser et al., 2011). Coh-Metrix originally described text characteristics through 53 measures, which were later extended to 108 measures (Graesser, McNamara, Louwerse & Cai, 2004; McNamara, Graesser, McCarthy & Cai, 2014). Clearly, such a high number of text characteristics would be impractical to use for interpretation. Accordingly, principal component analysis was used in order to identify eight principal components, under which all measures could be grouped. These principal components were narrativity, referential cohesion, syntactic simplicity, word concreteness, causal cohesion, verb cohesion, logical cohesion, and temporal cohesion. These components were then mapped to a 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). As a result, Coh-Metrix figures could be expressed along these five dimensions, providing results far easier to interpret.

On the basis of this model, a practical online tool known as TERA (*Coh-Metrix Common Core Text Ease and Readability Assessor)* was developed (Jackson, Allen & McNamara, 2017), the purpose of which was to provide an opportunity for text analysis. TERA reports text characteristics along five dimensions: *Narrativity, Syntactic simplicity, Word concreteness, Referential cohesion,* and *Deep cohesion*. Next, the definition of these five dimensions is presented.

Narrativity is identified as a continuum ranging between texts that are highly narrative in nature, and which, in turn, are assumed to be easier to process and informational texts, which are assumed to present a greater challenge in terms of comprehension. Narrative texts are characterized by a high proportion of frequent words, easy-to-understand verbs, and pronouns that make texts more engaging for readers (Jackson et al., 2017).

Syntactic simplicity is expressed as a function of the complexity of sentences in the text. The actual measure is calculated on the basis of 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 potential occurrence of similarities in sentence construction across paragraphs is also taken into consideration (Jackson et al., 2017).

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

Referential cohesion is defined with respect to overlap between words, word stems and concepts from sentence to sentence. If a high proportion of overlaps is detected in the text, this feature is considered to make comprehension easier (Jackson et al., 2017).

Deep cohesion is expressed as a function of the number of connectives in the text, representing to what extent events or various pieces of information in the text are tied together. If a high number of connectives is found, stronger links are present, making the comprehension of the text easier (Jackson et al., 2017).

In the course of the practical application of TERA, a visual representation of the above five measures is provided, in which the results are expressed in percentile figures. A sample TERA output is presented in Figure 1. below.

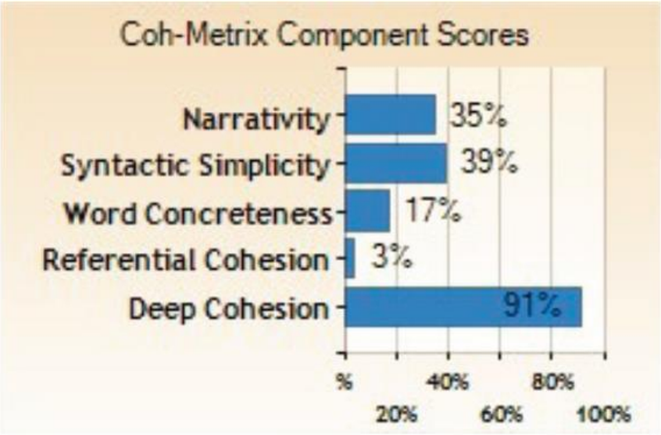

Figure 1. Sample TERA output

In addition to the development of TERA measures, Coh-Metrix has also inspired the development of a specific L2 readability measure, the basis of which is a lexical, a syntactic and a meaning construction index (Crossley et al. 2008). In the framework of a comparative study, the Coh-Metrix-based L2 readability index was contrasted with traditional measures of readability and was found to be superior to them (Crossley et al., 2011). Accordingly, it may well be considered to be of great value when L2 text properties are investigated further.

In light of the above, it seems reasonable to assume that the application of Coh-Metrix – and especially the application of TERA measures along with the Coh-Metrix-based L2 readability index – can provide a highly detailed and multi-faceted account of text properties in an L2 context on the basis of objectively measurable indices. Indeed, Coh-Metrix is a model of text analysis frequently referred in the literature (see e.g.: Aryadoust, & Liu, 2015; Crossley & McNamara, 2012; Jarvis et al., 2012). When used for comparing texts, such an analysis appears to be capable of revealing similarities and differences between various text characteristics, all of which can be considered to feed into the same underlying construct of text readability. Accordingly, as will be discussed, these measures were utilized in the study designed to compare test takers' writing performances.

4 The study

As mentioned above, in order to determine whether test taker performances are different in the case of traditional and modified ECL writing tasks, a study was designed, which was to apply Coh-Metrix- based objective indices of text properties. In the following this study will be discussed in detail.

4.1 Research design

In order to determine whether the differences mentioned actually occur, the study was designed in the following manner. First, writing tasks that had been constructed for three different levels of the exam (B1, B2, C1) were modified by changing the original required length of the responses. At level B1, the original length was 125 words, which was modified to 100 words; at level B2, the original length of 200 words was changed to 150 words; and at level C1, the required length of the response was changed from 300 words to 200 words.

Next, both the original and the modified tasks were completed by candidates with characteristics similar to those of live exam candidates, selected by using the standard procedures employed for identifying pretest candidates. The test takers produced handwritten performances, which were then to be typed by administrative personnel in order to make the performances accessible to computer-based analysis. The accuracy of the transfer to typewritten format was also verified.

Finally, the performances were analyzed using the Coh-Metrix webtool (McNamara et al., 2013), relying on TERA (yielding five indices) and the L2 readability index. Once these indices had been acquired for all texts produced by the candidates, statistical analyses were run in order to determine whether any statistically significant differences could be observed in any of the six Coh-Metrix-based text property indices. As the data were not on an interval scale and a normal distribution could not be assumed, a nonparametric test (Mann-Whitney's U-test) was applied for this purpose.

4.2 Data collection

In the course of data collection, mock exam candidates took both the original and the modified writing tasks. At level B1, candidates completed one writing task, while at levels B2 and C1 they produced responses to two writing tasks. The number of candidates available varied from level to level. A total of 40 candidates' responses were collected at level B1, 44 responses were collected at level B2, and 35 responses were collected at level C1.

4.3 Results and discussion

In accordance with the procedures described in section 4.1, the data collected were analyzed using the Coh-Metrix web tool's TERA platform, and the Coh-Metrix L2 readability indices were also calculated for all performances. Next, these results were investigated for any statistically significant differences between performances on the original and the modified tasks by applying Mann-Whitney's U-test. In the following, a discussion of the findings will be presented.

Figure 2. presents the graphical rendering of the results for *Narrativity* in the B1 task. The blue bars represent results on the modified task, while the green bars depict results on the original tasks.

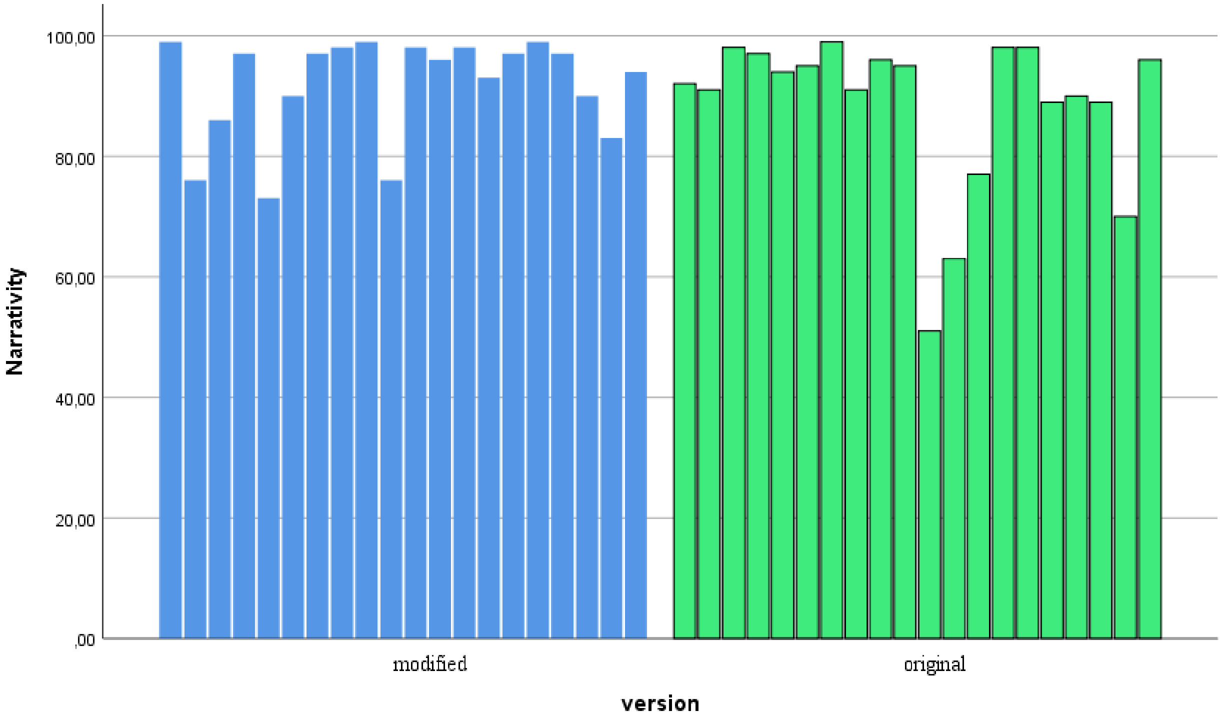

Figure 2. B1 task - TERA results for *Narrativity*

The two groups of performances show a considerable degree of similarity, which may be the consequence of the apparent homogeneity of the results. While there are a few exceptions in both the original and the modified tasks, the general pattern is a high degree of *Narrativity* in both groups.

Figure 3. presents the TERA results for *Syntactic simplicity* of the same performances in the two groups. The results appear to be much more heterogeneous concerning this text property than in the case of *Narrativity*. On the other hand, this heterogeneity seems to be detectable in both groups, suggesting that the groups' results may be similar with regard to this text characteristic.

Figure 4. provides a chart depicting the results concerning *Word concreteness*. Once again, the results appear to be heterogeneous in both groups, seemingly suggesting that the results themselves are similar in the two sets of performances.

Figure 5. offers the graphical representation of the results for *Referential cohesion*. The pattern observable is similar to the ones related to the other text characteristics: results are heterogeneous, and they appear to be similar in the two groups examined.

Figure 6. presents the results the analysis yielded about the fifth text property, *Deep cohesion*. Again, the results appear to show a similar picture to that of the previous figures. The results related to this text characteristic are, once again, varied in both groups, showing an apparently similar pattern.

Figure 7. presents the results for the last text property related to the B1 task, *L2 readability*. The impression one gets is, again, similar to the previous charts, although there appears to be less variation in the results in each group. Yet, the overall picture seems to indicate the two groups examined are quite similar.

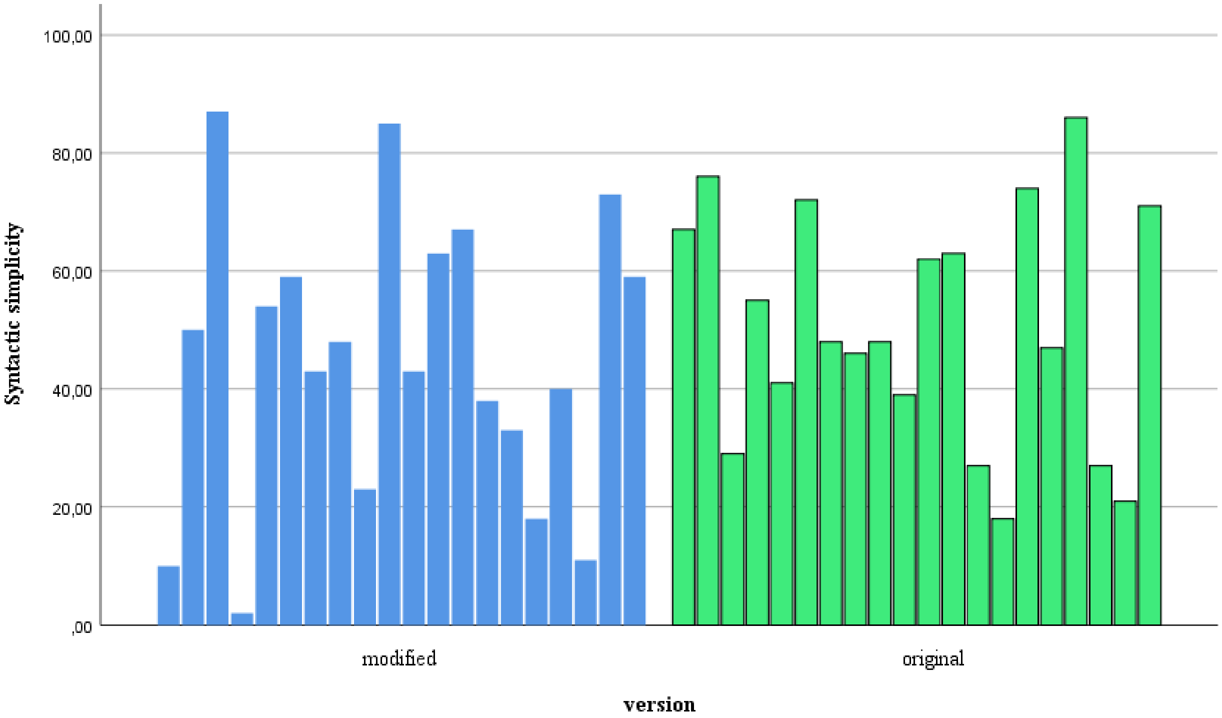

Figure 3. B1 task - TERA results for *Syntactic simplicity*

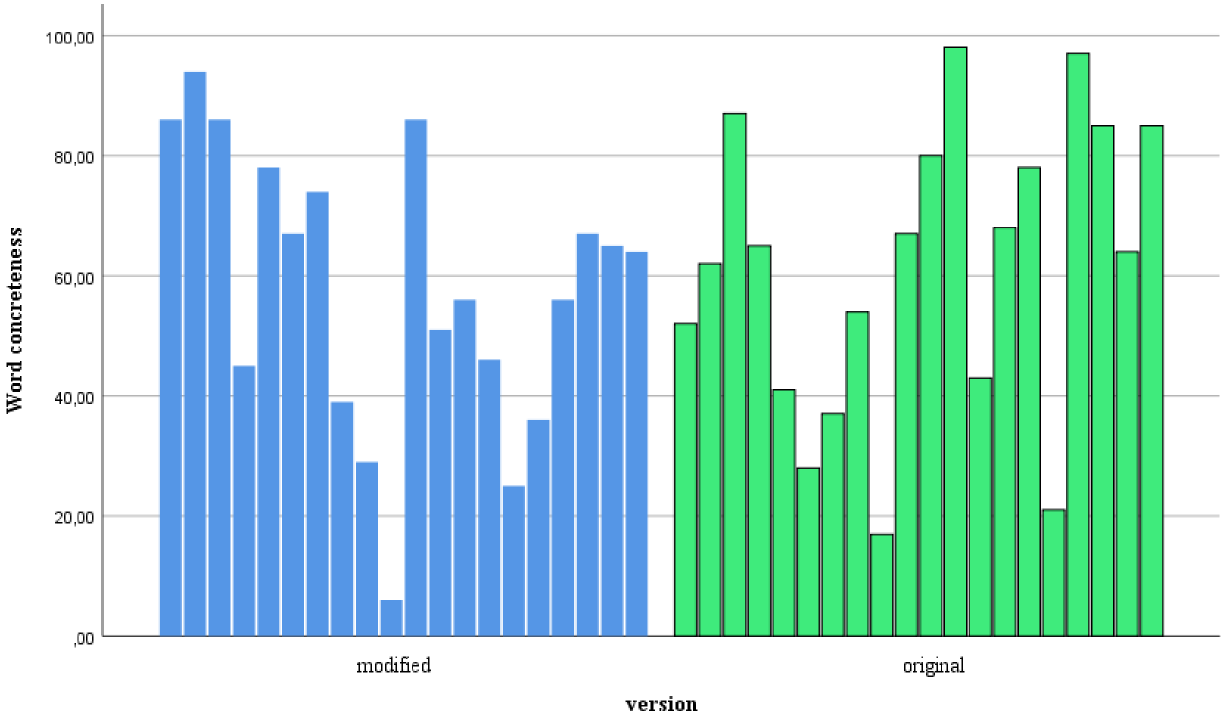

Figure 4. B1 task - TERA results for *Word concreteness*

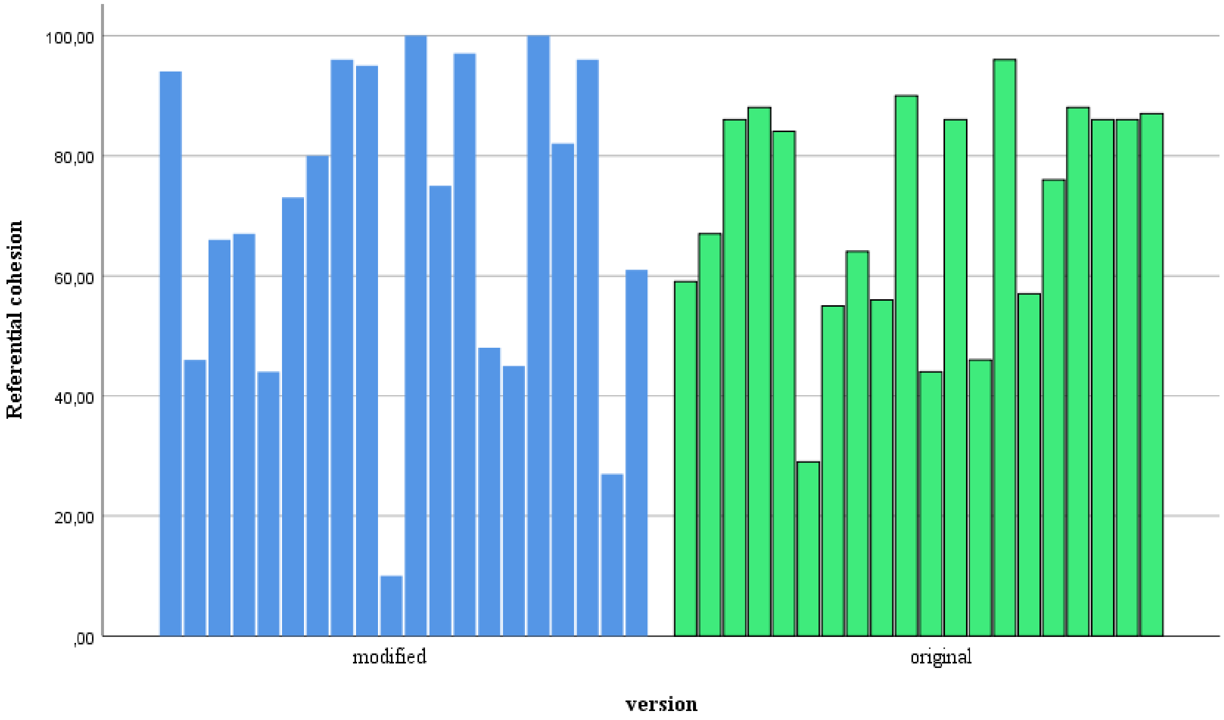

Figure 5. B1 task - TERA results for *Referential cohesion*

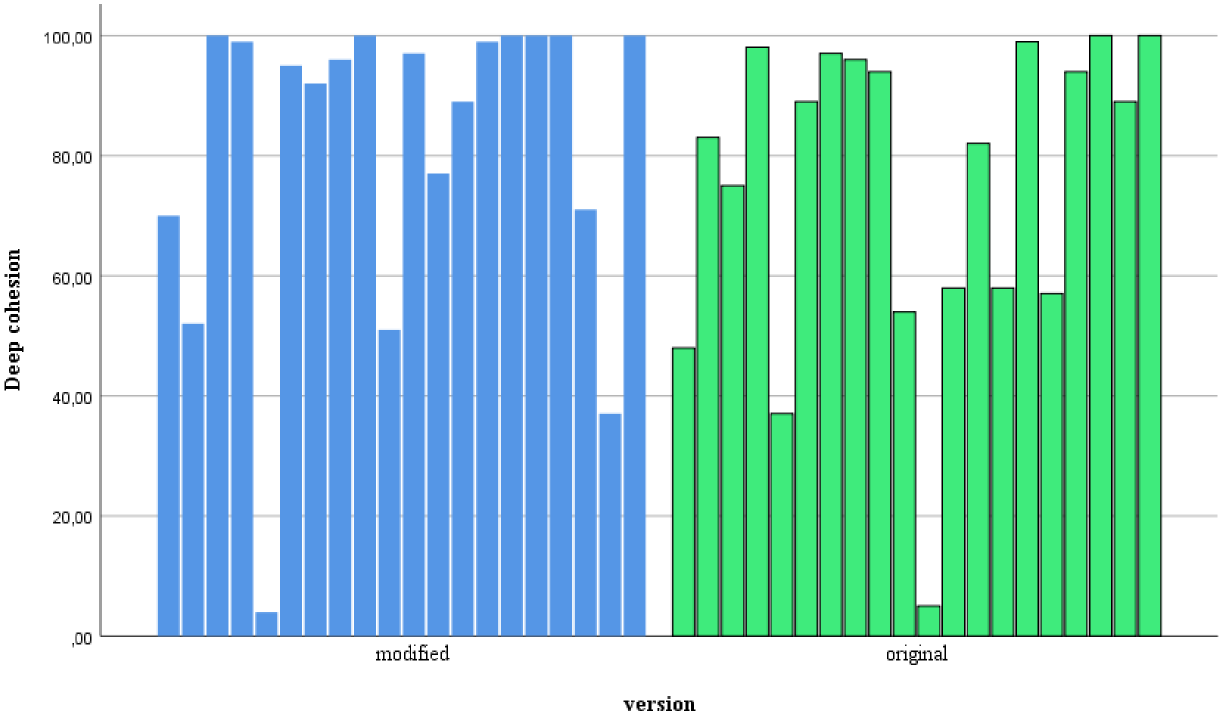

Figure 6. B1 task - TERA results for *Deep cohesion*

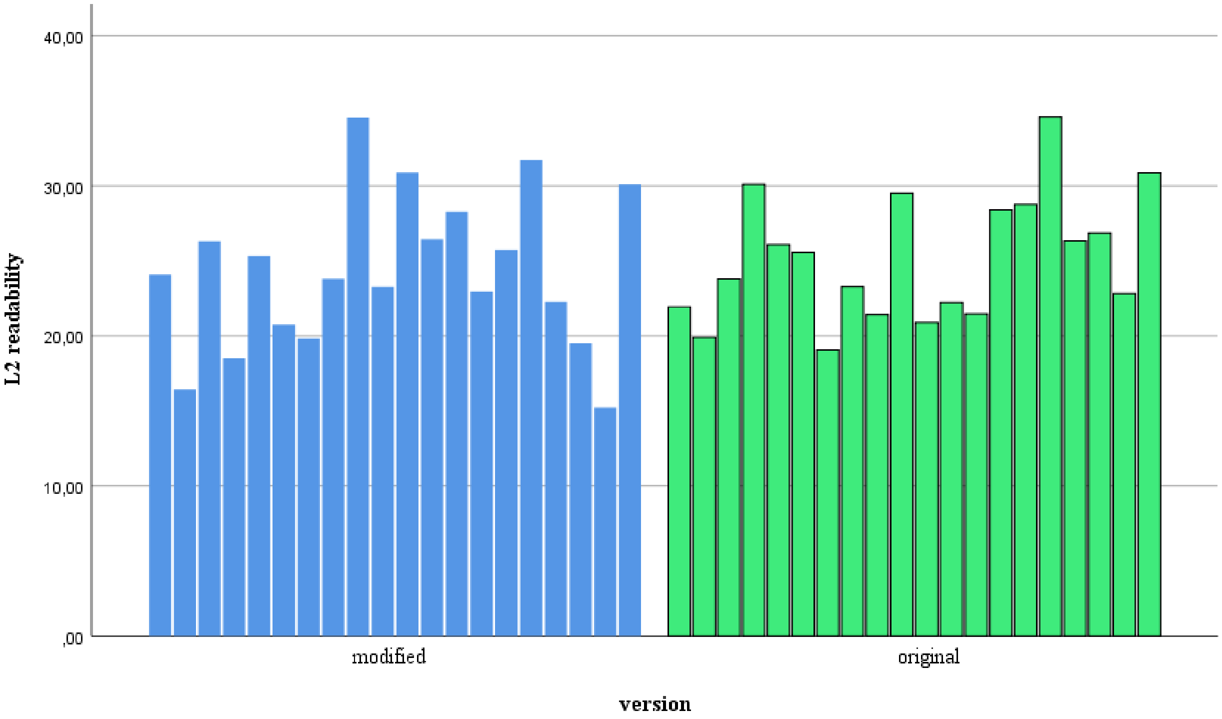

Figure 7. B1 task - Coh-Metrix *L2 Readability* results

All of these apparent similarities, however, are mere impressions, and to form an informed opinion, it is necessary to examine whether the similar patterns observable in the charts actually indicate a genuine similarity. As was discussed in Section 4.1, this was implemented by applying Mann-Whitney's U-tests in order to decide whether any statistically significant differences could be detected in the results on the various text characteristics measures between the two groups. The results of this analysis are presented in Table 1.

	Narrativity	**Syntactic simplicity**	**Word concreteness**	**Referential cohesion**	**Deep cohesion**	**L2 readability**
Mann-Whitney U	162.500	172.500	184.500	192.500	154.500	181.000
Wilcoxon W	372.500	382.500	394.500	402.500	364.500	391.000
Z	-1.018	-.744	-.420	-.203	-1.236	-.514
Asymp. Sig. (2-tailed)	.309	.457	.675	.839	.216	.607
Exact Sig. [2*(1-tailed Sig.)]	.314[b]	.461[b]	.678[b]	.841[b]	.221[b]	.620[b]

b. Not corrected for ties.

Table 1. B1 task – Mann-Whitney's U-test results

As can be observed, there were no significant differences detected in relation to any of the text characteristics examined. This means that the original and the modified tasks produced performances that cannot be distinguished in terms of the text properties examined.

Next, the results of the B2 performances will be examined. For this level, the candidates completed two different tasks. Since the two tasks were independent and the performances were to represent different aspects of the construct, the results of the two tasks were analyzed separately. Figure 8. presents the results for *Narrativity* for the first B2 task.

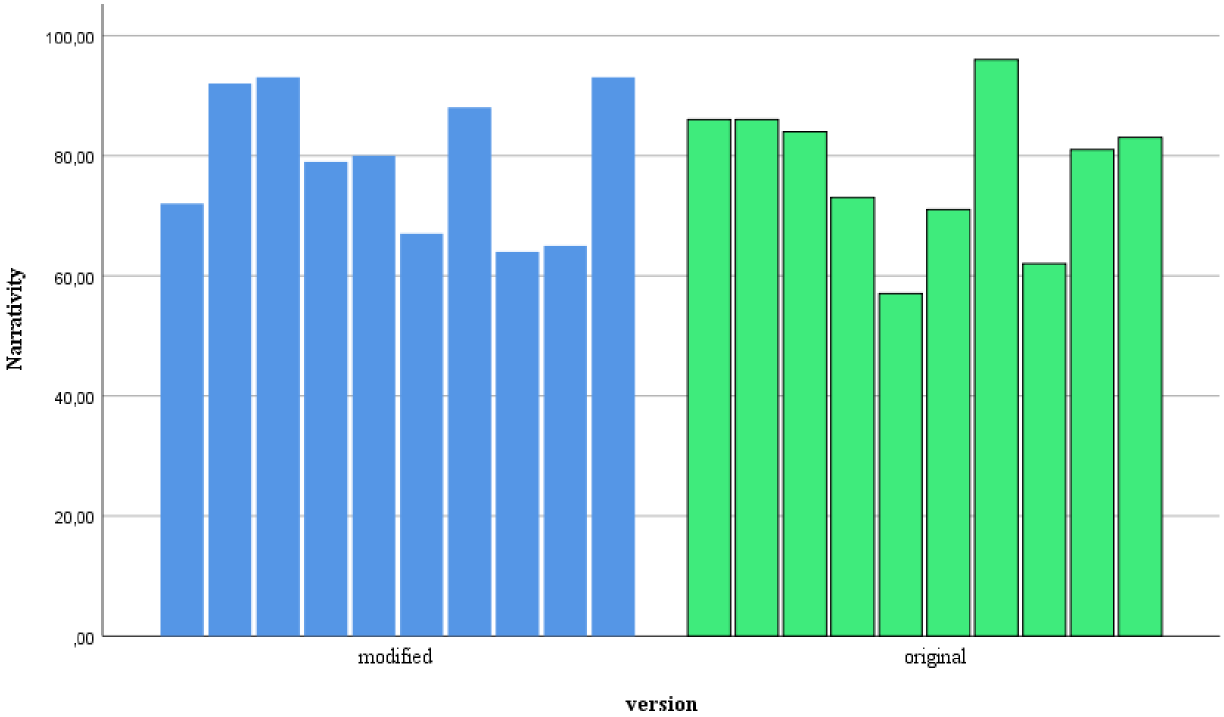

Figure 8. B2 task 1 - TERA results for *Narrativity*

Similarly to the B1 results, scores appear to be reasonably high in both versions, although there seems to be somewhat more variability than in the B1 tasks. Concerning the comparison of the two versions, once again, there appears to be little difference between the modified and the original task performances on this text characteristic.

Figure 9. presents the results for *Syntactic simplicity* in the first B2 task. Unlike in the previous cases, there appears to be a difference between the two versions. The modified task has yielded higher values for this text characteristic. In comparison, the high values in the modified version are higher than in the original, and the low values show a similar tendency. Whether this apparent difference is an actually significant one, however, is to be determined at a later stage of the analysis.

Figure 10. provides the graphic representation of the results for *Word concreteness*. Once again, the two sets of results appear to differ. This time it is the original task that seems to have generated higher results in general because fewer performances show low values. Again, it is yet to be determined whether these differences might be statistically significant or not.

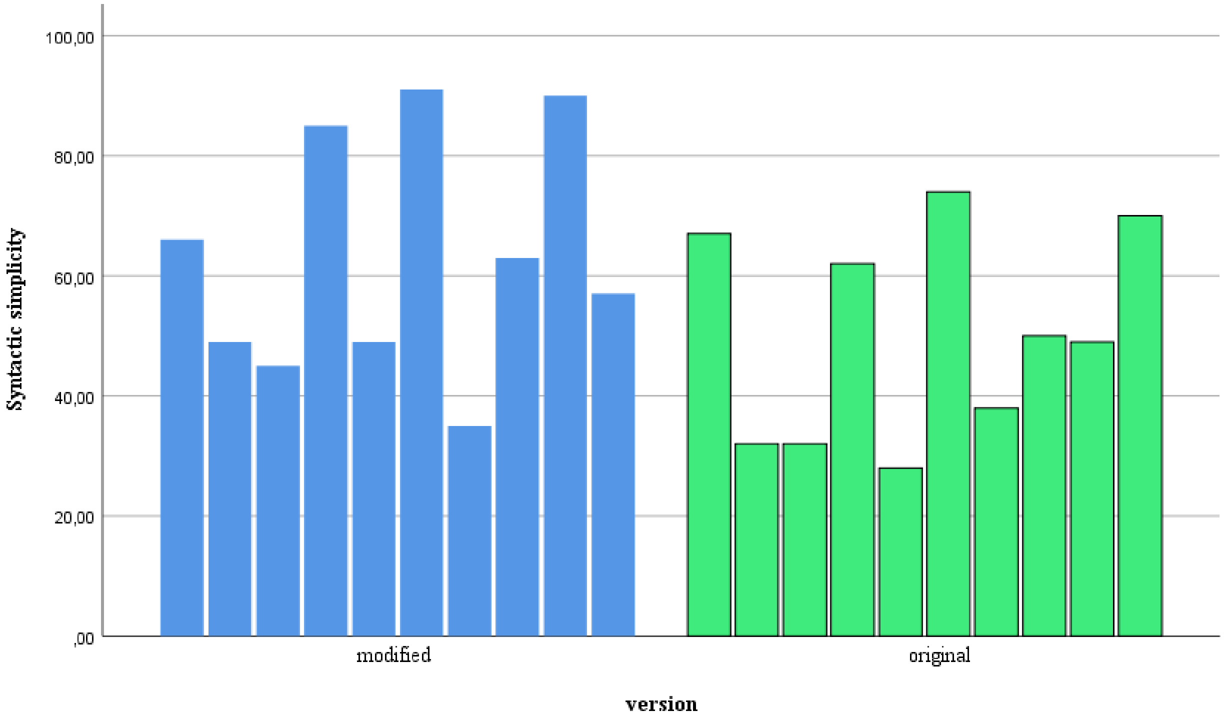

Figure 9. B2 task 1 - TERA results for *Syntactic simplicity*

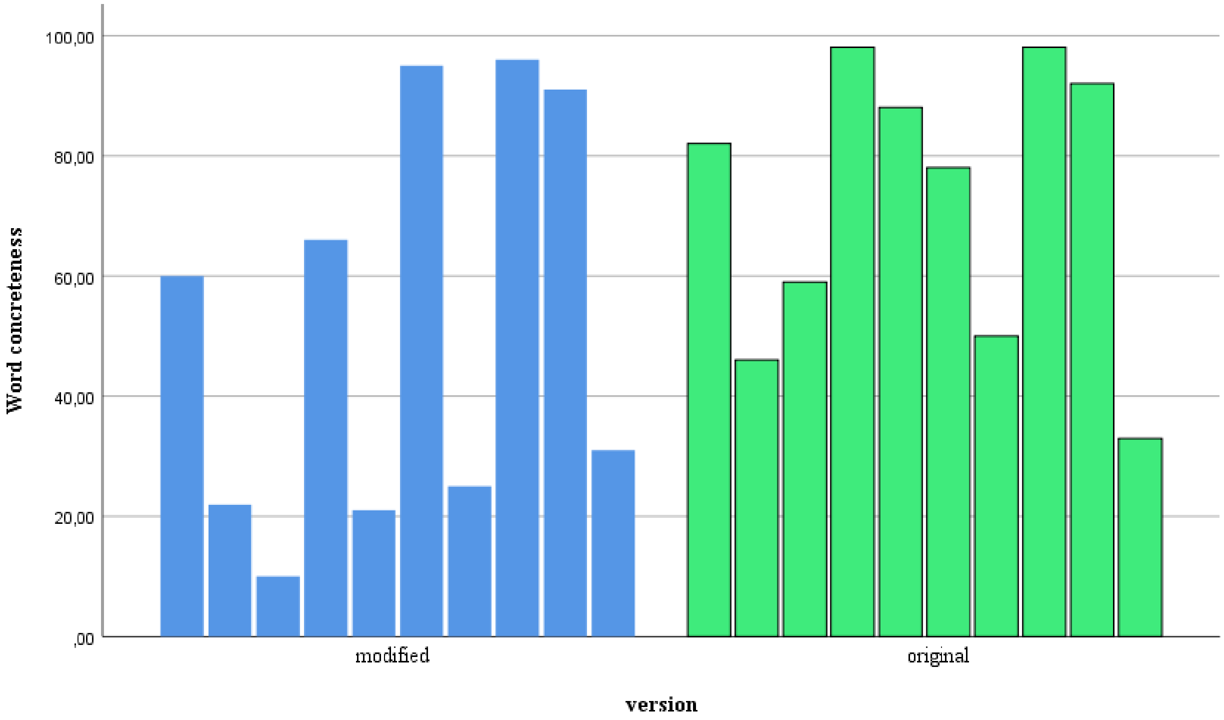

Figure 10. B2 task 1 - TERA results for *Word concreteness*

Figure 11. portrays results for *Referential cohesion* for the two versions of the first B2 task. It is worth noting that, as can be seen, the overwhelming majority of all performances show low values on this text property. While there are a couple of performances in the original task that seem to have reached noticeably higher values, they are not necessarily indicative of a significant difference between the two versions. This issue, again, will be investigated later.

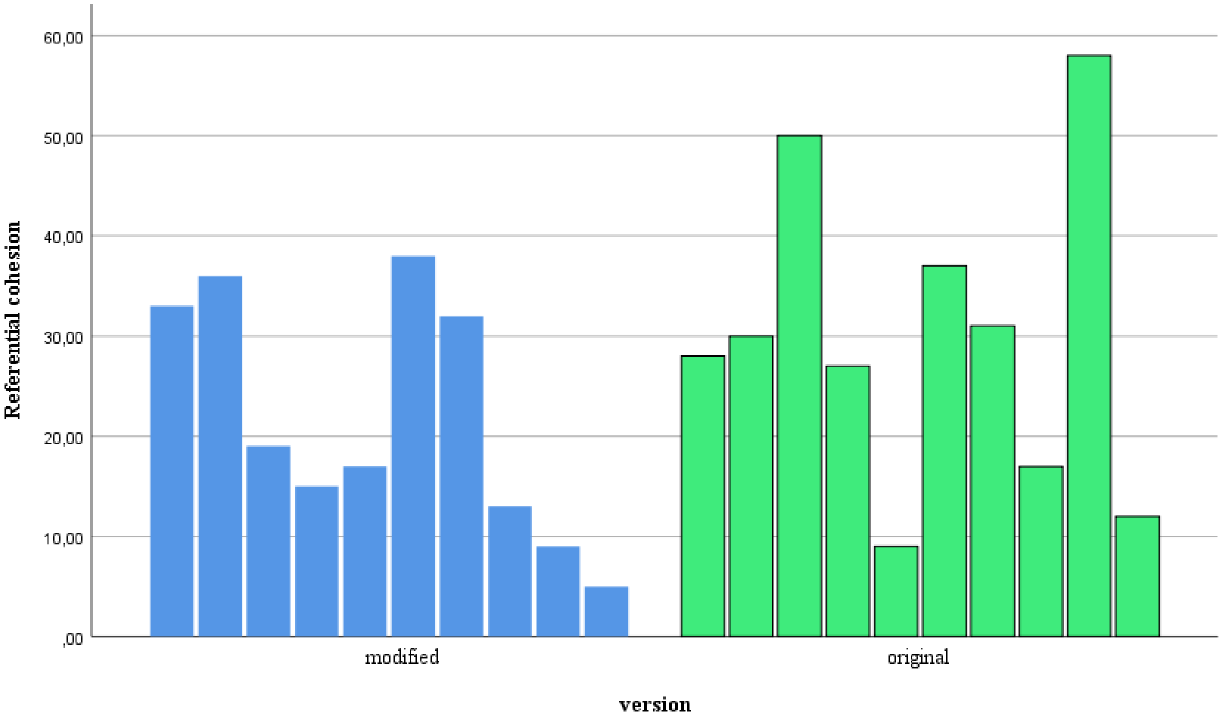

Figure 11. B2 task 1 - TERA results for *Referential cohesion*

Figure 12. presents results for *Deep cohesion* in the two versions. While some differences appear in this case as well, the general tendencies in the two versions do not appear to be very different in the case of this text characteristic, even though the results appear to be somewhat higher for the original version of the task.

Figure 13. offers a graph depicting the results for the last text property under scrutiny, *L2 readability*. As can be observed, no major differences emerge between the two versions, although the values appear to be somewhat higher in the case of the modified task.

In general, it can be stated that the graphs depicting the results concerning the various text properties in the first B2 task seem to suggest that, at least in some cases, there may be actual differences between the two sets of performances. In order to decide whether this, indeed, is the case, results for the two groups were checked for statistically significant differences. The findings are presented in Table 2. As can be inferred from the table, once again, no significant differences were detected between the performances yielded by the two versions of the task. This, yet again, is an indication that the modified task did not generate performances measurably different from the performances on the original task.

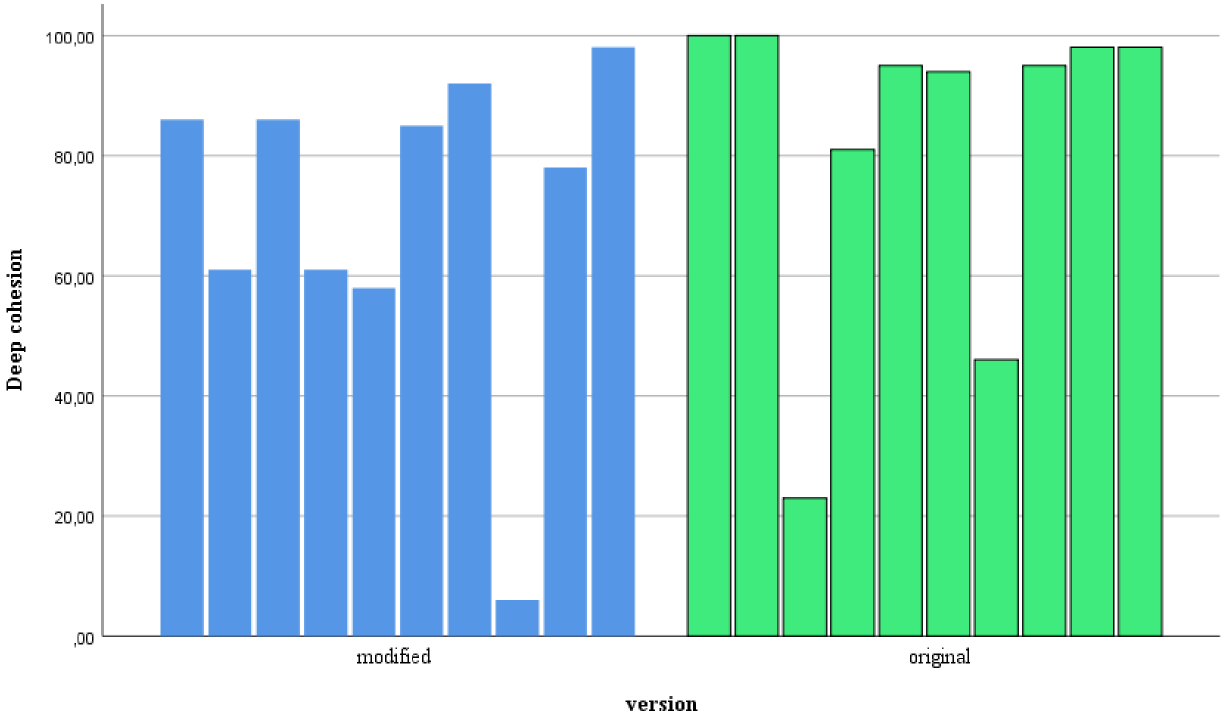

Figure 12. B2 task 1 - TERA results for *Deep cohesion*

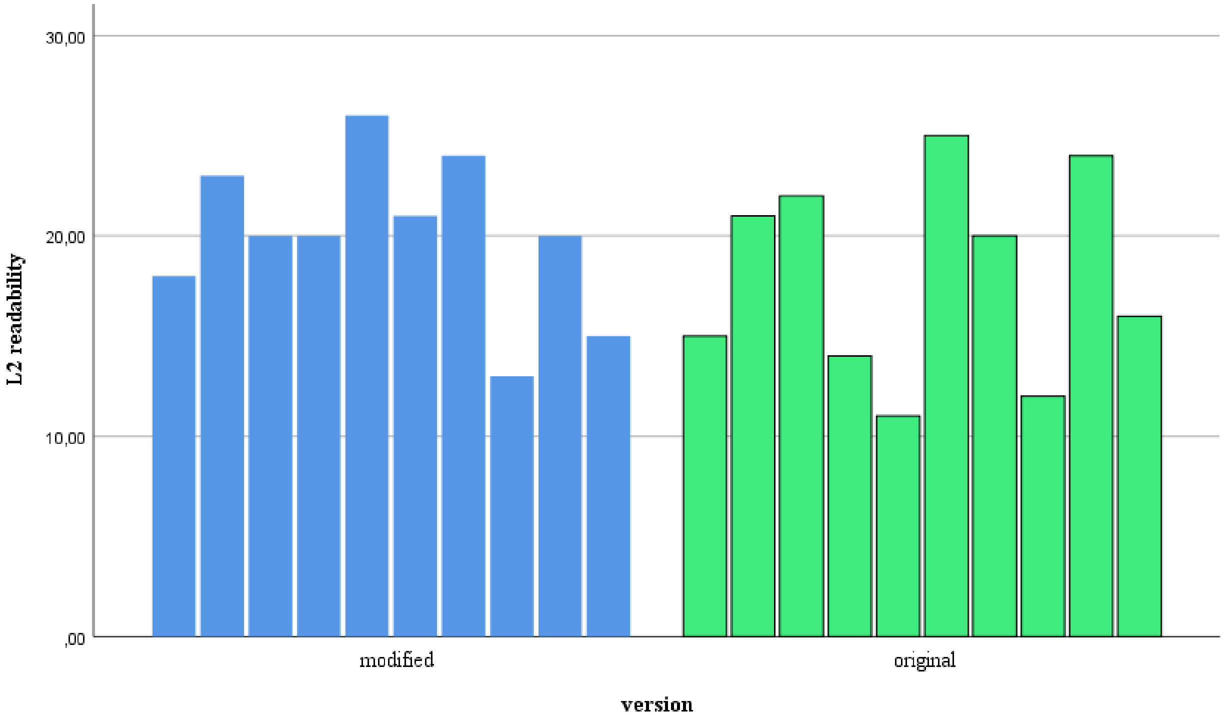

Figure 13. B2 task 1 - Coh-Metrix *L2 Readability* results

	Narrativity	Syntactic simplicity	Word concreteness	Referential cohesion	Deep cohesion	L2 readability
Mann-Whitney U	47.000	34.000	31.000	39.000	27.000	40.000
Wilcoxon W	102.000	89.000	86.000	94.000	82.000	95.000
Z	-.227	-1.212	-1.437	-.832	-1.744	-.760
Asymp. Sig. (2-tailed)	.820	.226	.151	.405	.081	.447
Exact Sig. [2*(1-tailed Sig.)]	.853[b]	.247[b]	.165[b]	.436[b]	.089[b]	.481[b]

b. Not corrected for ties.

Table 2. B2 task 1 – Mann-Whitney's U-test results

Next, the results for the second B2 task will be examined. Figure 14. presents the results for *Narrativity* in the two groups.

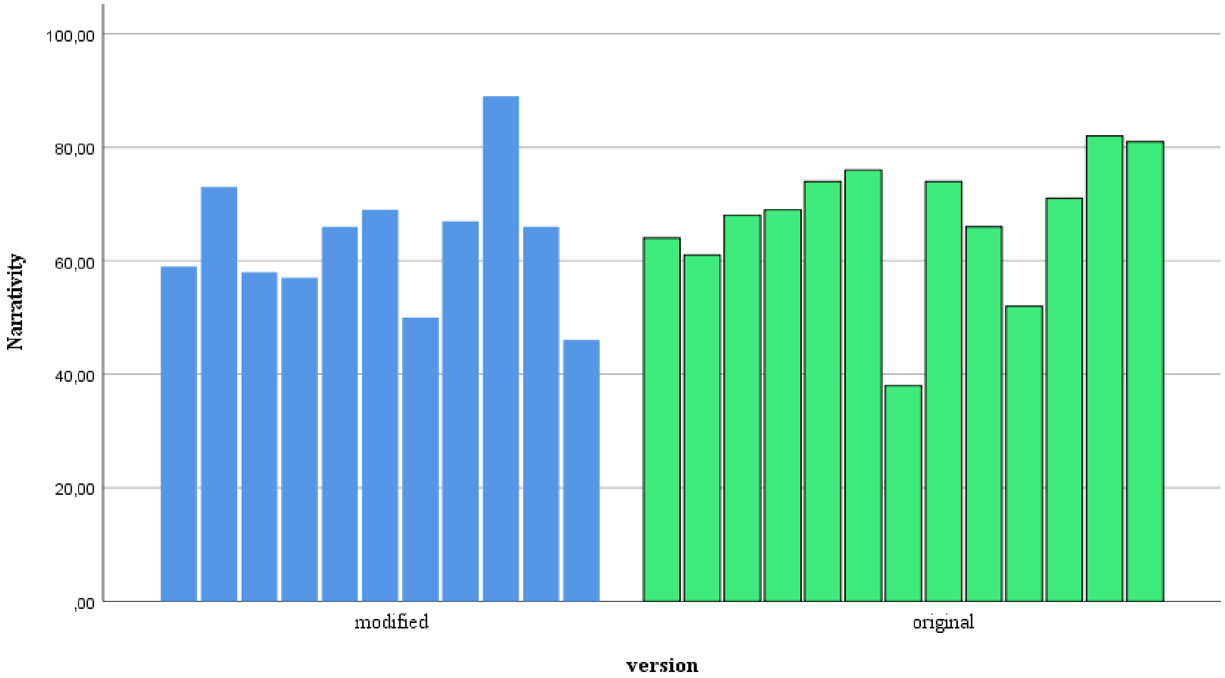

Figure 14. B2 task 2 - TERA results for *Narrativity*

The results appear to be relatively homogeneous within the groups. Since the range of scores in the two groups seems similar, this, again, gives the impression that there is little difference between the two groups of performances in terms of this text characteristic.

Figure 15. depicts the results for *Syntactic simplicity*. As is apparent, the range seems similar in both groups, although performances with high scores appear to be more numerous in the group that completed the original task.

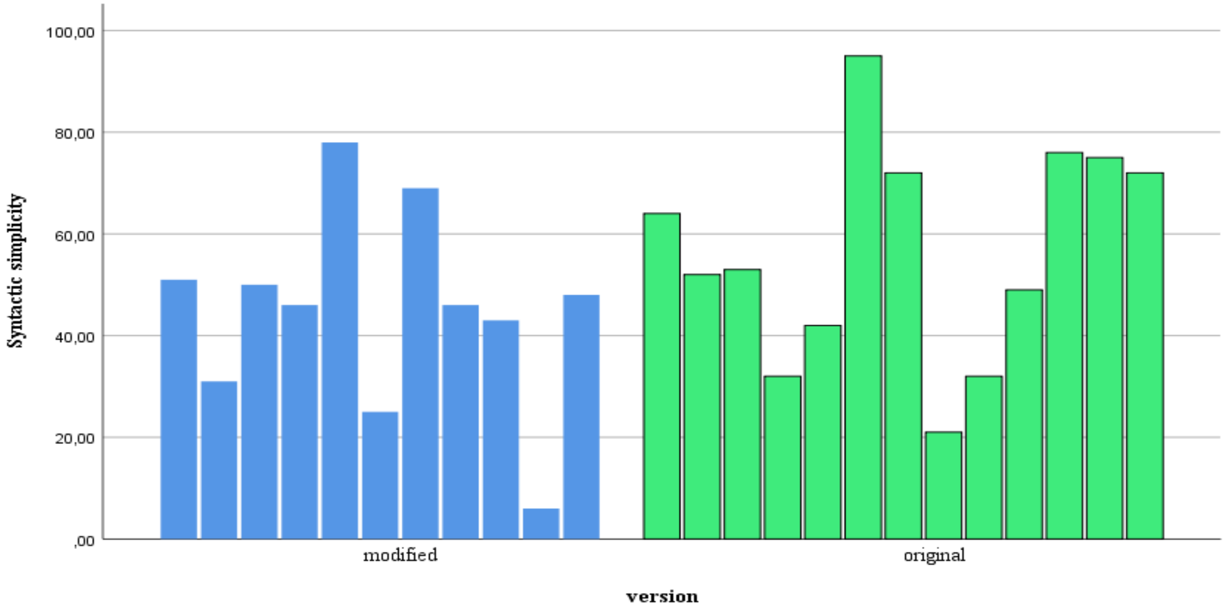

Figure 15. B2 task 2 - TERA results for *Syntactic simplicity*

Let us now examine the results for *Word concreteness*. A graphic representation is presented in Figure 16.

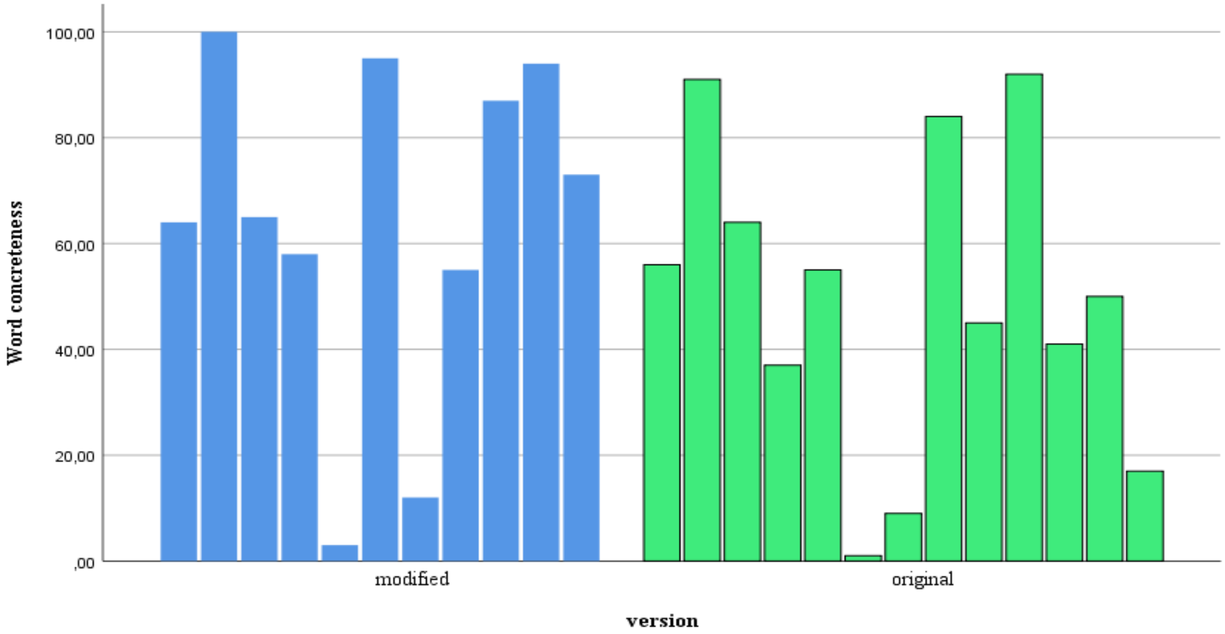

Figure 16. B2 task 2 - TERA results for *Word concreteness*

Apparently, the two groups' performances show similar results on this text property. It is also worth noting that the range of scores seems quite extreme in both groups, indicating noticeable variety.

Figure 17. presents the results for *Referential cohesion*. As can be observed, there appears to be a difference between the two groups of performances here in that a number of high scores can be detected in performances on the modified task. While these results suggest a genuine difference between the two groups, it needs to be checked whether the differences observed are statistically significant. As in the case of the former tasks, this will be done along with the other text characteristics.

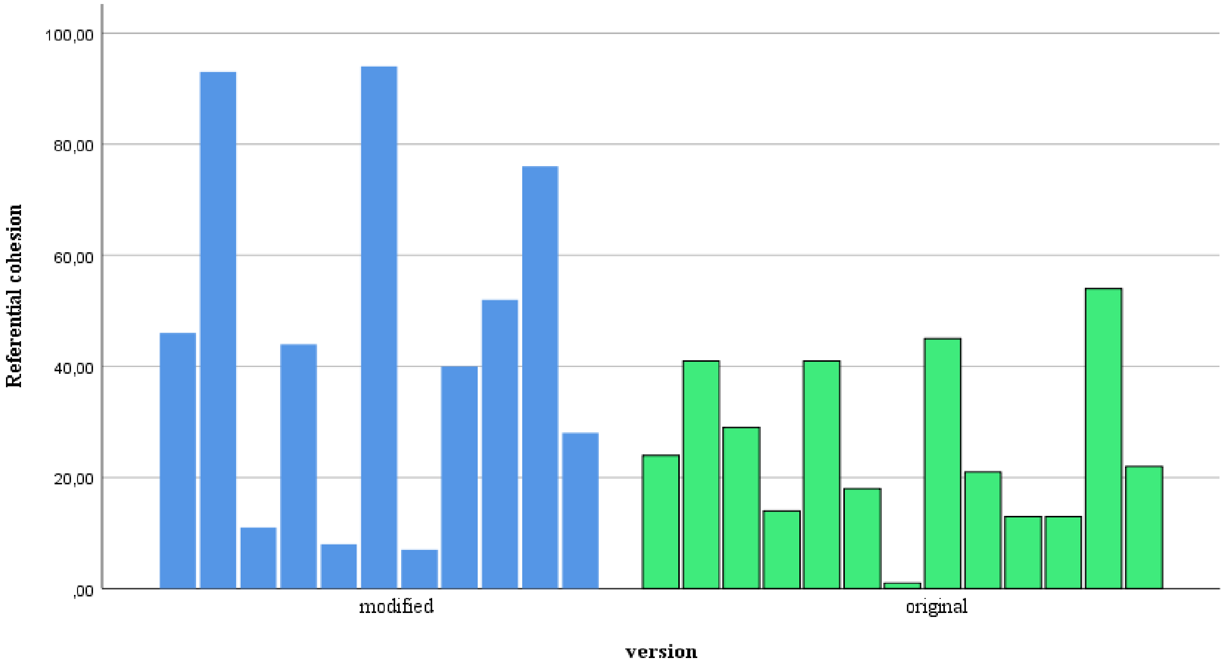

Figure 17. B2 task 2 - TERA results for *Referential cohesion*

Figure 18. offers a graphic representation of the results for *Deep cohesion*. Unlike in the previous figure, there appears to be little difference between the two groups here. While results appear to be somewhat less homogeneous in the case of the modified task, the general tendency in both groups seems to be the presence of relatively high scores on this text characteristic.

The last text property to examine in the analysis of B2 tasks is *L2 readability*. Results are presented in Figure 19. Again, there seems to be little difference between the two groups, although the highest scores in the group of performances on the modified task exceed the highest scores in the other group.

As in the case of the previous two tasks, whether the apparent differences in the scores on some text characteristics actually indicate any statistically significant differences will be examined. The results of Mann-Whitney's U-test are presented in Table 3.

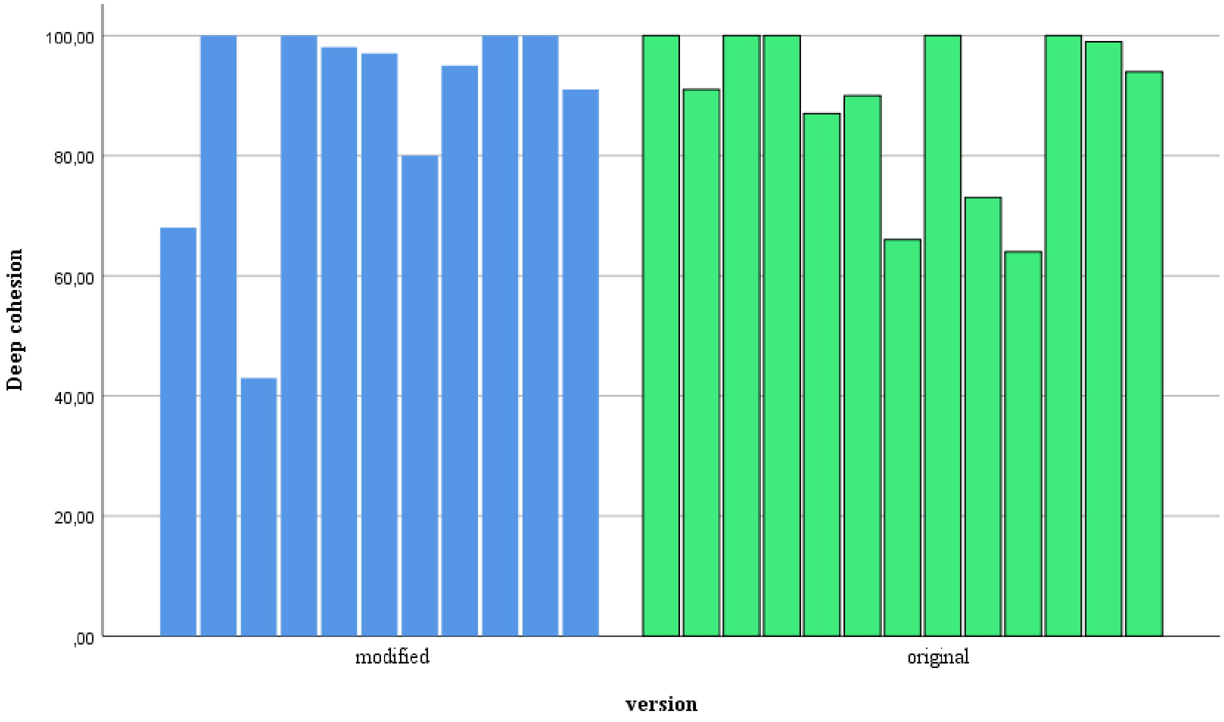

Figure 18. B2 task 2 - TERA results for *Deep cohesion*

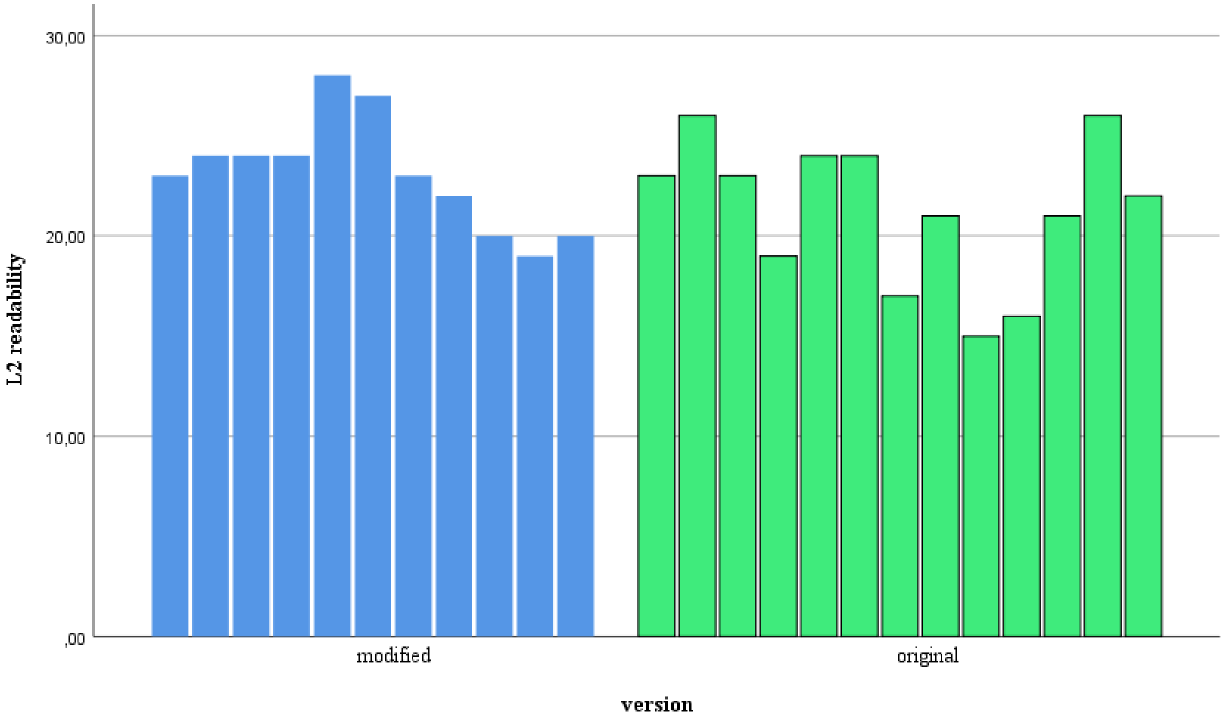

Figure 19. B2 task 2 - Coh-Metrix *L2 Readability* results

	Narrativity	Syntactic simplicity	Word concreteness	Referential cohesion	Deep cohesion	L2 readability
Mann-Whitney U	49.500	48.000	44.000	49.000	69.500	53.000
Wilcoxon W	115.500	114.000	135.000	140.000	160.500	144.000
Z	-1.276	-1.362	-1.594	-1.304	-.119	-1.080
Asymp. Sig. (2-tailed)	.202	.173	.111	.192	.905	.280
Exact Sig. [2*(1-tailed Sig.)]	.207	.186	.119	.207	.910	.303

b. Not corrected for ties.

Table 3. B2 task 2 – Mann-Whitney's U-test results

As is discernible from Table 3., there were, again, no significant differences detected in relation to any of the text characteristics. What we can conclude from this is that the modification in the length of task 2 at level B2 had no significant impact on the properties of the performances.

The results for the C1 tasks will now be examined. Figure 20. presents the results for *Narrativity* in the first task.

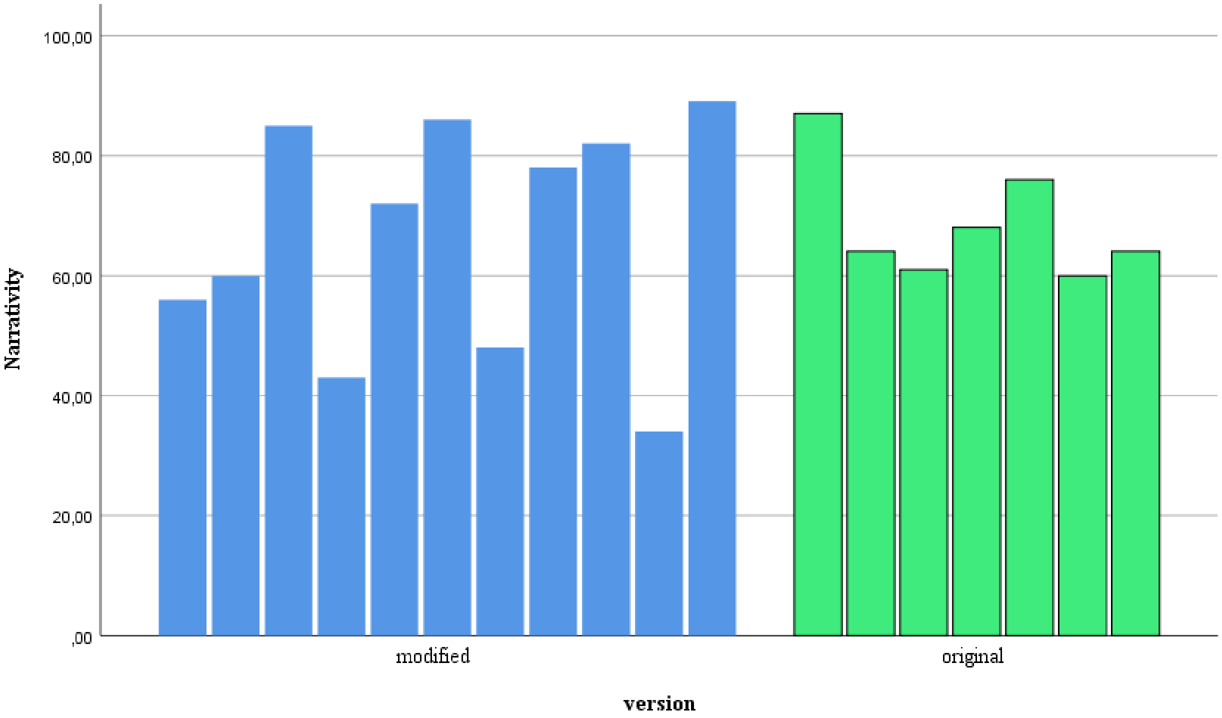

Figure 20. C1 task 1 - TERA results for *Narrativity*

Again, there appears to be a difference between the two groups of performances. The range of scores is greater in the case of the modified tasks, and there are some performances with relatively low scores, while in the other group all scores are relatively high. As was the case with the other levels, whether or not the difference is significant will be determined at a later time.

Figure 21. provides a graphic overview of the results for *Syntactic simplicity*. The patterns in the two groups appear to be different again. The performances on the modified task seem to have a greater variety in terms of range, and the actual scores at both ends of the scale are more extreme than in the other group.

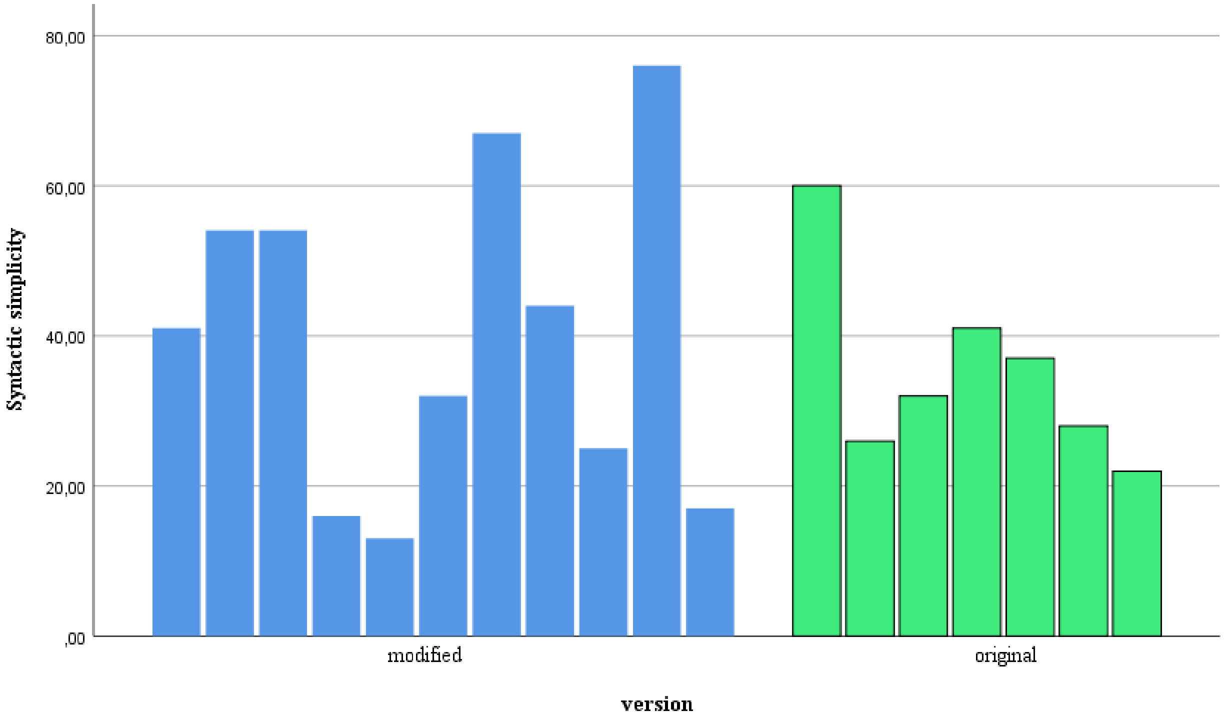

Figure 21. C1 task 1 - TERA results for *Syntactic simplicity*

The next text property to examine is *Word concreteness*. The results are presented in Figure 22. The striking observation is that the overwhelming majority of scores is extremely low in both groups. Considering, however, that this was a task at level C1, the low scores on this text characteristic become less surprising, as this level does involve a greater degree of abstraction in general. Despite this similarity, the two groups appear to differ both in terms of the range and values of the actual cases, with scores in the modified version seeming to be systematically lower.

Figure 23. depicts the results for the next text property, *Referential cohesion*. Here, while the range of scores is clearly different, this may well be the effect of only a small number of cases, and the general tendencies in the two groups appear to be similar.

Figure 24. presents the results for *Deep cohesion*. While the range is greater in the case of the original task, there appears to be the same general tendency in the two groups for scores to be relatively high. Thus, it seems that performances do not appear to differ very much in the two sets.

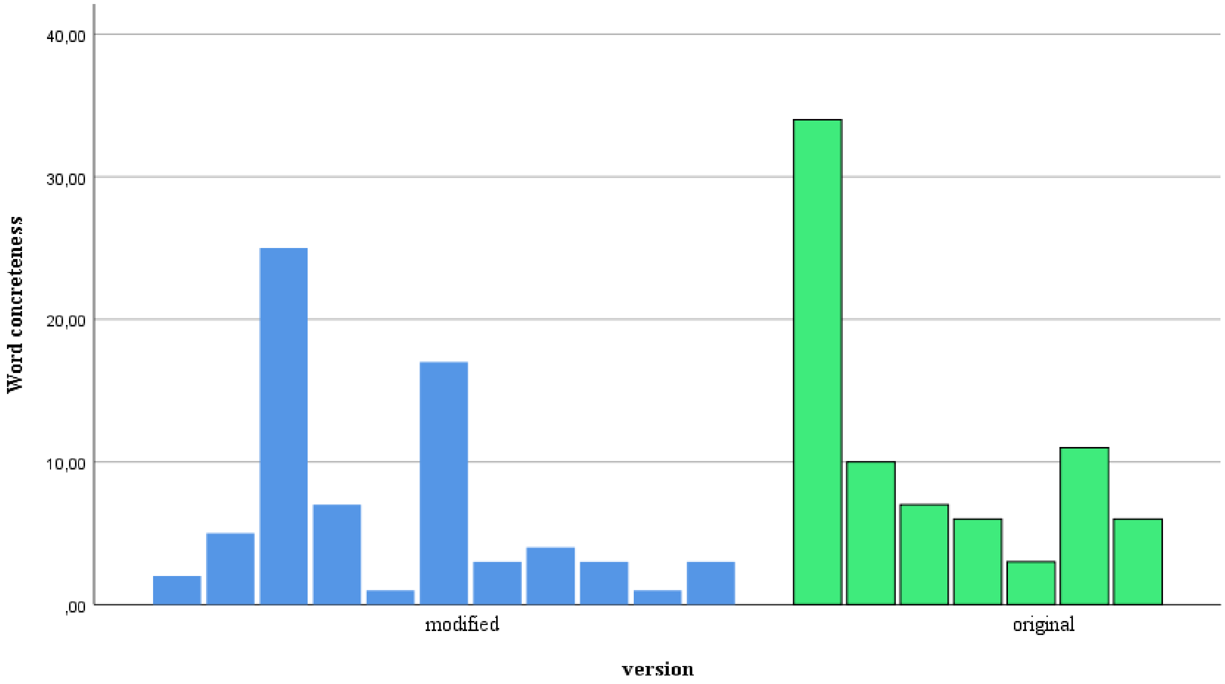

Figure 22. C1 task 1 - TERA results for *Word concreteness*

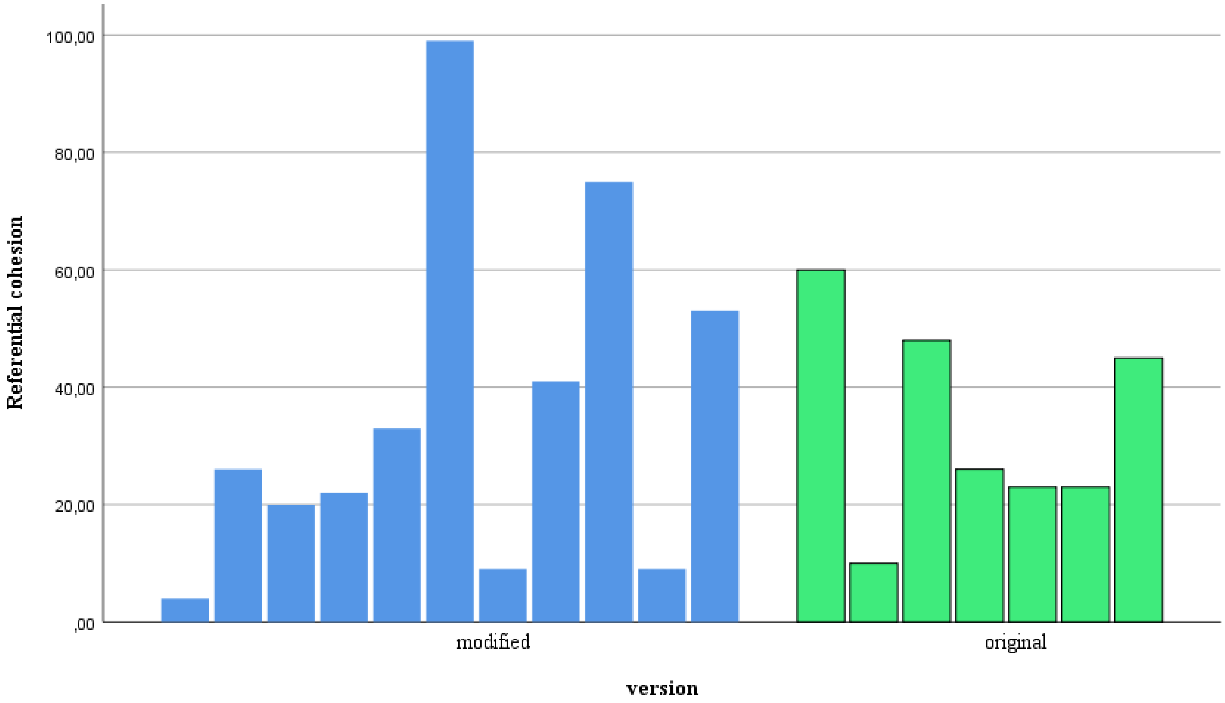

Figure 23. C1 task 1 - TERA results for *Referential cohesion*

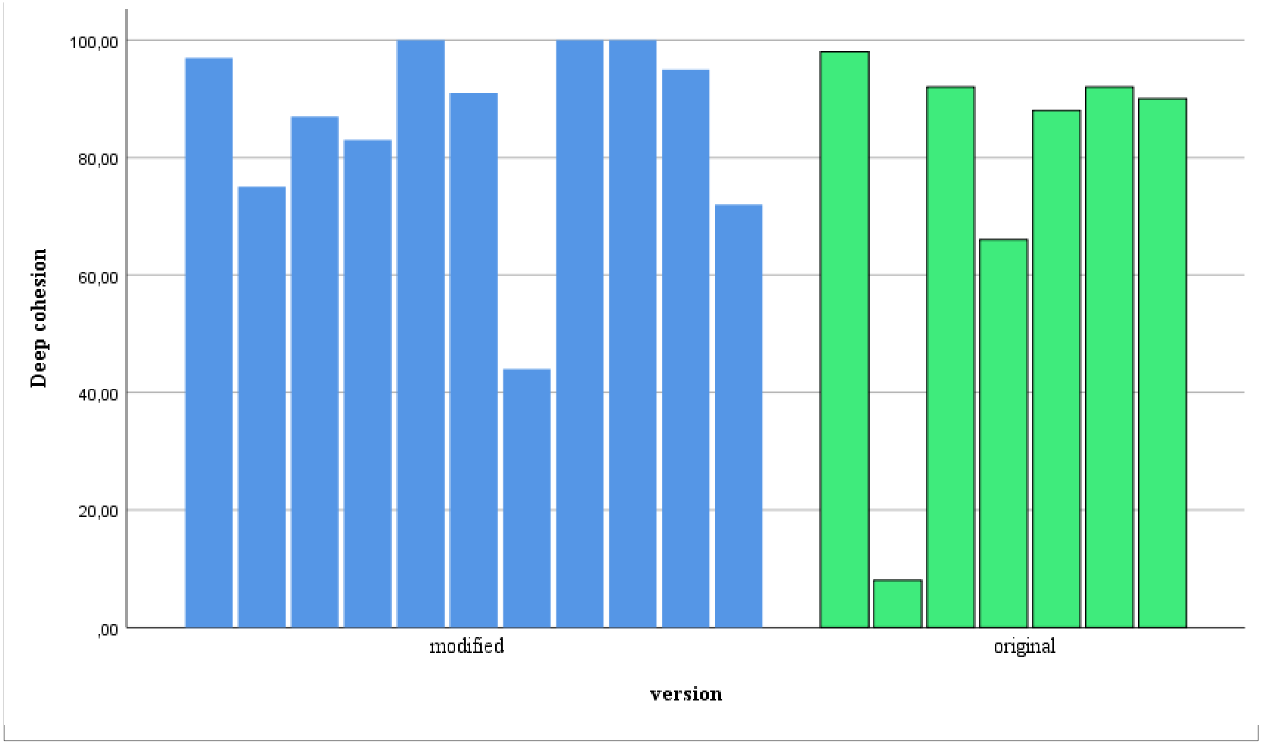

Figure 24. C1 task 1 - TERA results for *Deep cohesion*

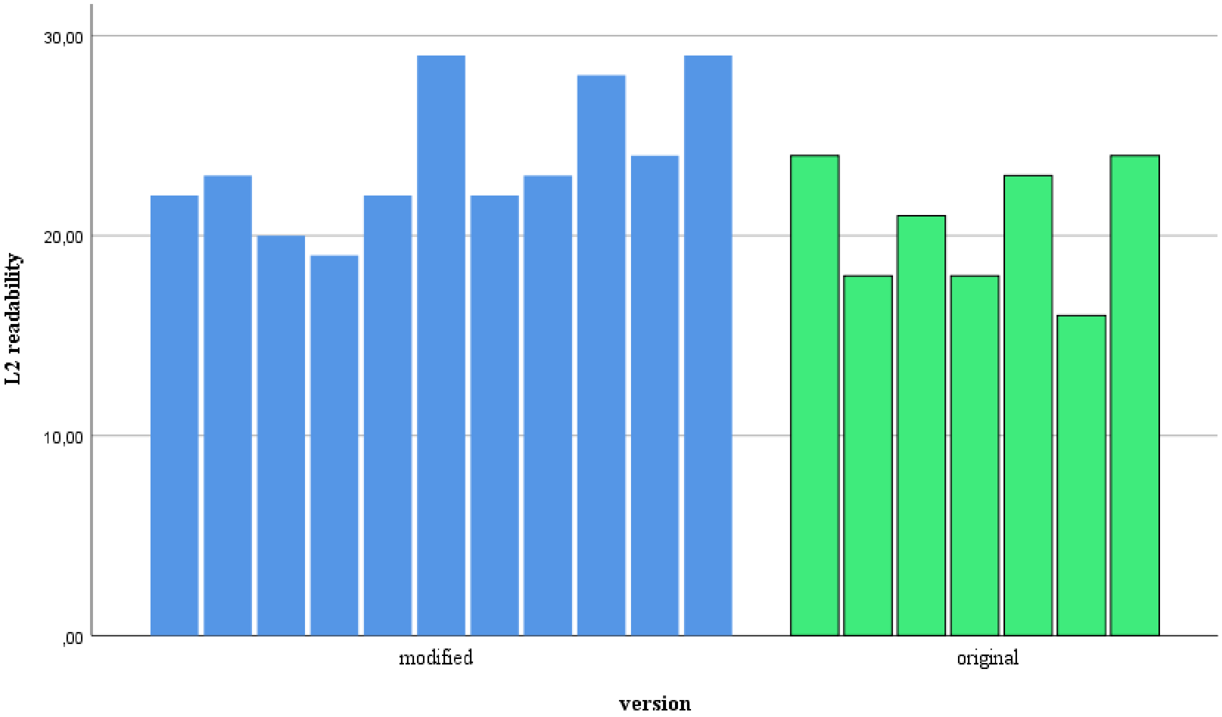

Figure 25. C1 task 1 - Coh-Metrix *L2 Readability* results

The last text characteristic to examine in connection with the first C1 task is *L2 readability*. Results are presented in Figure 25. Once again, while some performances on the modified task received higher scores than any on the original task, the general tendency appears to be quite similar in the two groups.

In order to find out whether the apparent differences or similarities are indicative of reality, it is time now to examine the results of Mann-Whitney's U-test in Table 4.

	Narrativity	**Syntactic simplicity**	**Word concreteness**	**Referential cohesion**	**Deep cohesion**	**L2 readability**
Mann-Whitney U	36.500	35.000	19.000	33.500	31.000	23.000
Wilcoxon W	102.500	63.000	85.000	99.500	59.000	51.000
Z	-.181	-.317	-1.778	-.454	-.681	-1.414
Asymp. Sig. (2-tailed)	.856	.751	.075	.650	.496	.157
Exact Sig. [2*(1-tailed Sig.)]	.860[b]	.791[b]	.085[b]	.659[b]	.536[b]	.179[b]

b. Not corrected for ties.

Table 4. C1 task 1 – Mann-Whitney's U-test results

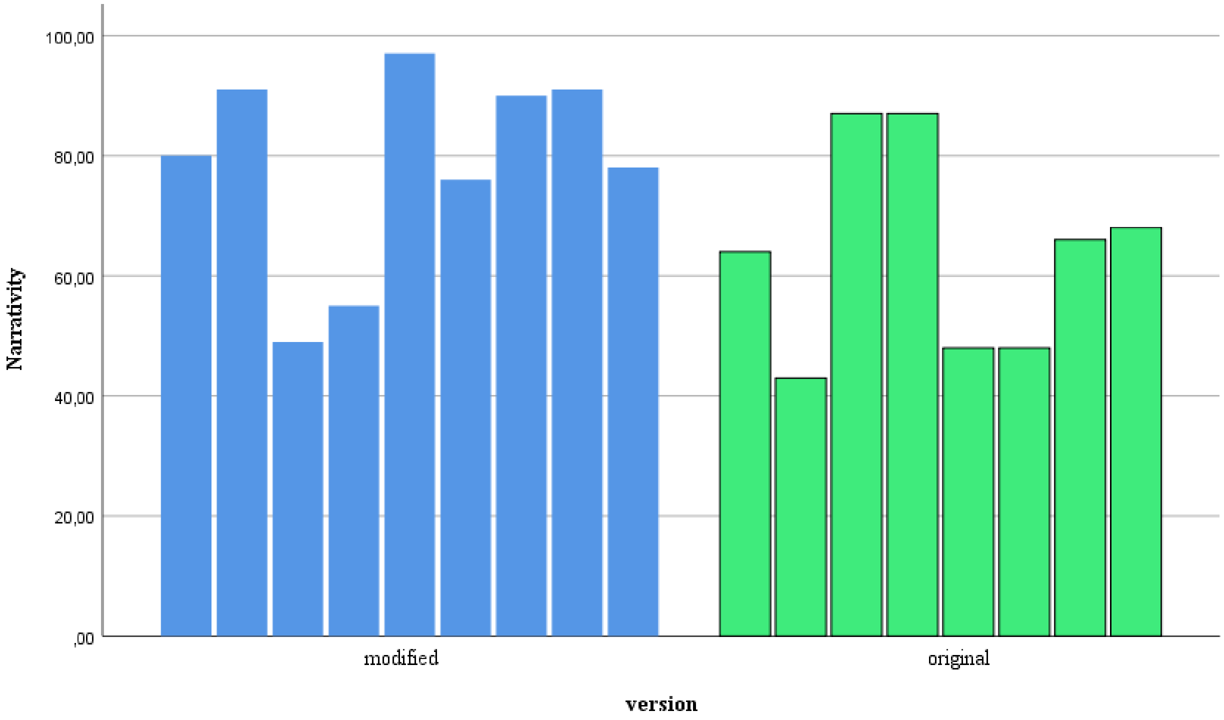

Figure 26. C1 task 2 - TERA results for *Narrativity*

Similarly to the previous tasks, the results indicate once again that there were no statistically significant differences between the performances on the original and the modified tasks for any of the text characteristics examined.

As the final stage of this analysis, the results of the second C1 task will now be discussed. Figure 26 presents the results for *Narrativity*. As can be observed, there appears to be some difference between the two groups in that although the ranges seem quite similar in the two sets, the performances on the modified task received higher scores for this text property. The difference, however, does not appear to be very big.

The next text characteristic is *Syntactic simplicity*, the analysis of which yielded results presented in Figure 27. As has been observed for the other C1 tasks, in these instances the range is greater in the case of performances on the modified task, and the scores assumed more extreme values as well. Thus, the two sets appear to be different again.

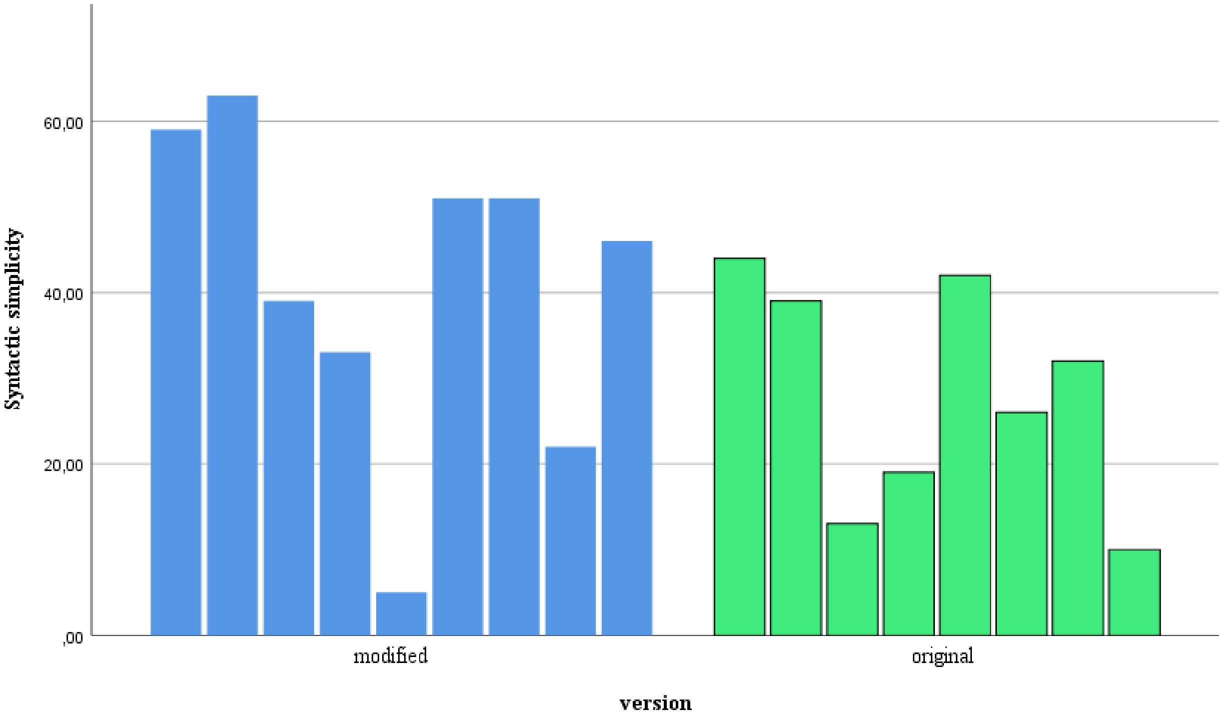

Figure 27. C1 task 2 - TERA results for *Syntactic simplicity*

Figure 28. depicts the results for *Word concreteness*. Again, similarly to the other C1 task, it is also observable in this case that the scores are quite low, indicating a high frequency of abstract vocabulary. The similarity is further reinforced by the pattern of the scores in the two groups. Once again, performances for the modified task tended to receive scores that are visibly lower, and the range is smaller in this set again, indicating a somewhat more homogeneously higher frequency of abstract vocabulary than in the other group. As we have seen in the case of the first C1 task, however, this does not necessarily indicate a significant difference.

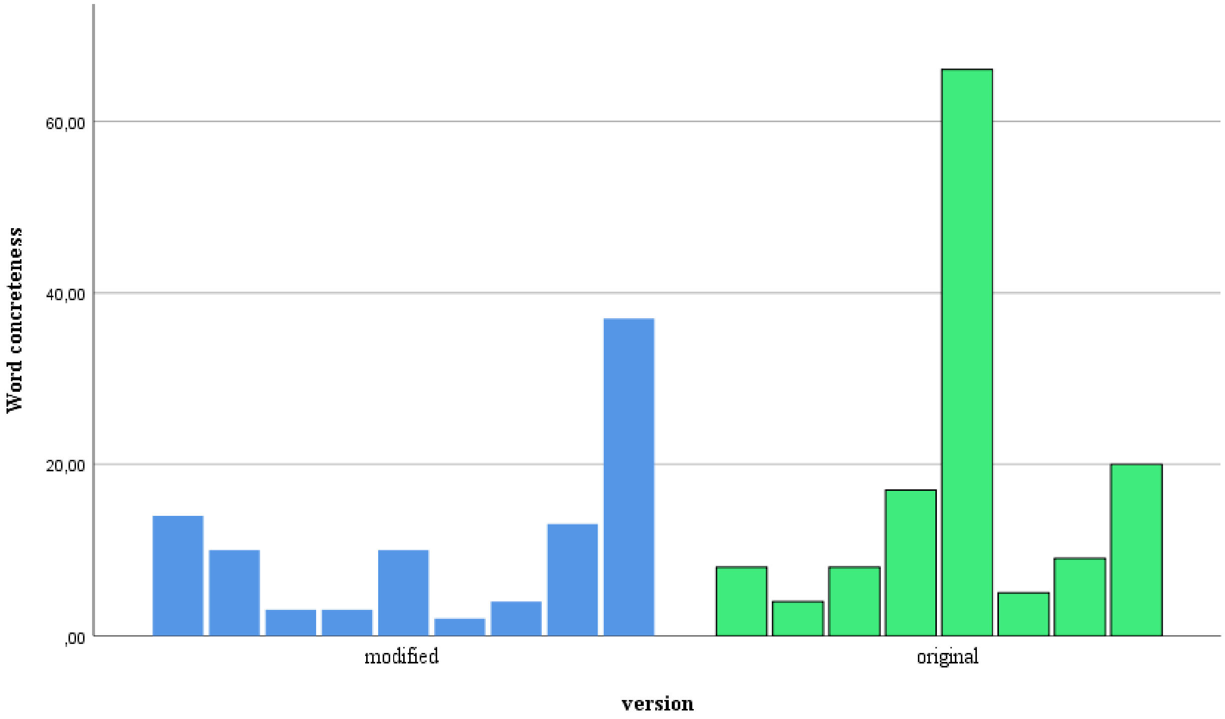

Figure 28. C1 task 2 - TERA results for *Word concreteness*

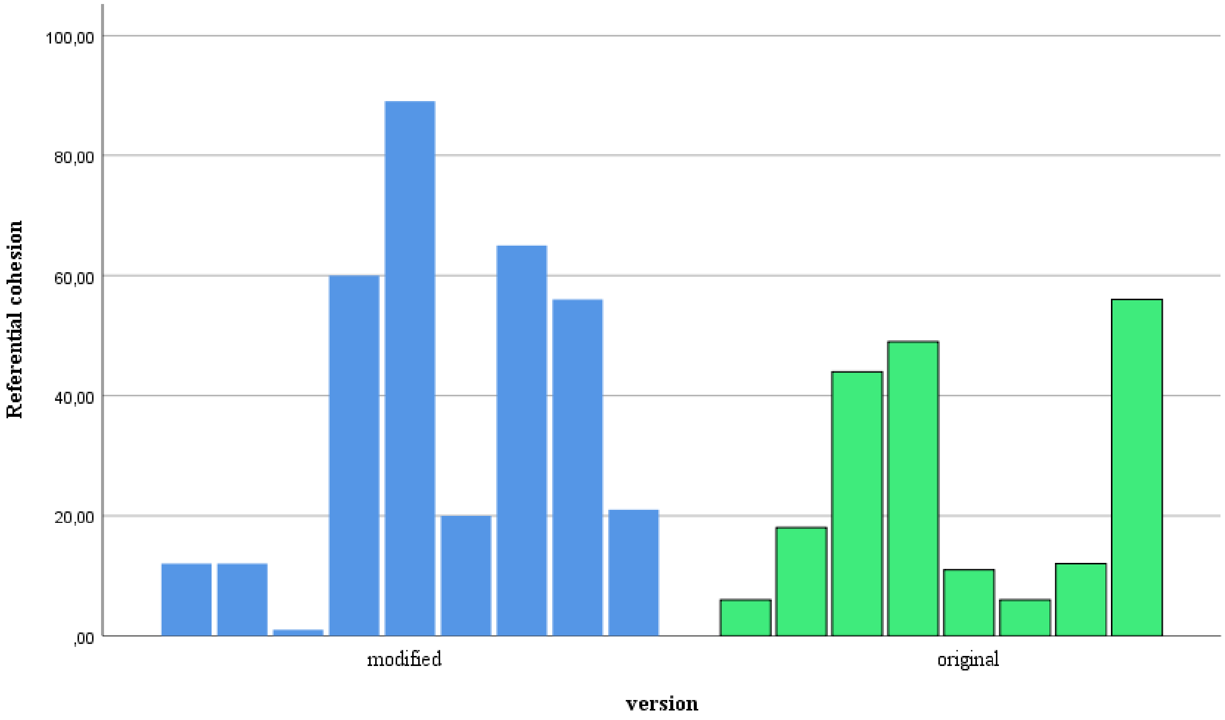

Figure 29. C1 task 2 - TERA results for *Referential cohesion*

Figure 29. presents the results for *Referential cohesion*. In this case, again, there seems to be a difference between the two groups in that there appears to be a tendency in the performances on the modified task to receive higher scores for this text property.

Results for *Deep cohesion* are provided in Figure 30. As is apparent, the two groups seem not to differ very much on this text characteristic. Scores appear to be predominantly high in both sets of performances.

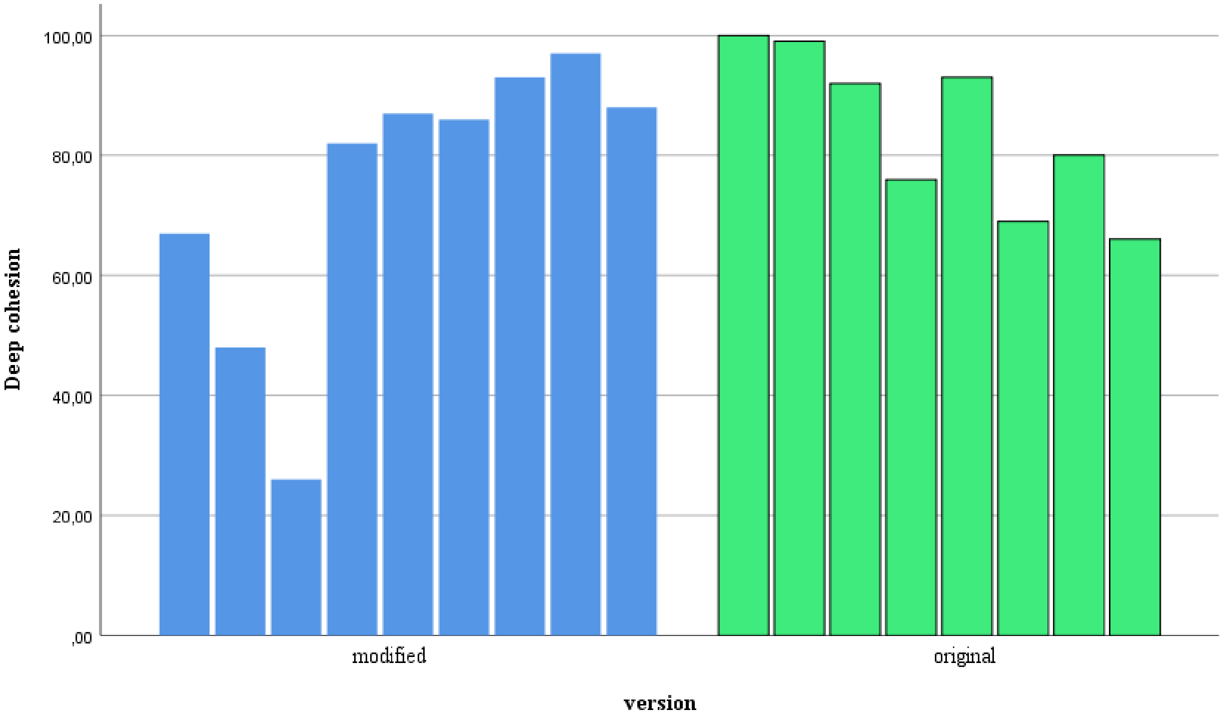

Figure 30. C1 task 2 - TERA results for *Deep cohesion*

The final text property to investigate is *L2 readability*, with results presented in Figure 31. The pattern for this text property is that the two groups display similarities. Scores, though rather low, tend to be relatively homogeneous in both sets of performances.

Having reviewed all text characteristics related to the second C1 task, the significance of the differences identified needs to be examined. The results of Mann-Whitney's U-test are found in Table 5.

As was the case for all other tasks at the other levels, no significant differences were detected for any of the text properties examined.

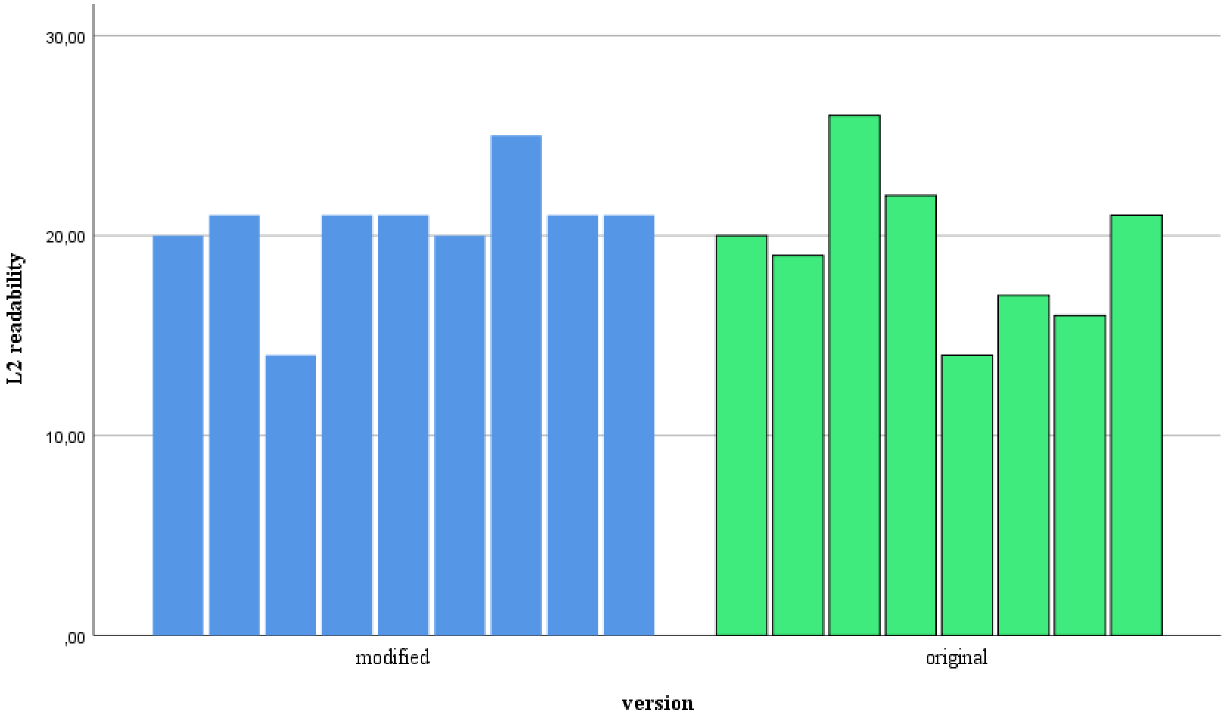

Figure 31. C1 task 2 - Coh-Metrix *L2 Readability* results

	Narrativity	Syntactic simplicity	Word concreteness	Referential cohesion	Deep cohesion	L2 readability
Mann-Whitney U	16.000	18.500	27.500	23.500	28.500	28.000
Wilcoxon W	52.000	54.500	72.500	59.500	73.500	64.000
Z	-1.928	-1.686	-.820	-1.207	-.722	-.789
Asymp. Sig. (2-tailed)	.054	.092	.412	.227	.470	.430
Exact Sig. [2*(1-tailed Sig.)]	.059^b	.093^b	.423^b	.236^b	.481^b	.481^b

b. Not corrected for ties.

Table 5. C1 task 2 – Mann-Whitney's U-test results

At this point it is worth clarifying a seemingly controversial issue. On might wonder why the raw data at all three levels seemed to suggest that there are differences between the two groups on a number of occasions, while the statistical analyses detected no significant differences. The answer most probably lies in the fact that, although occasional differences involving differing ranges and extreme values did occur, there was no tendency in these differences, and they can most likely be attributed to the unlikely performance of individual test takers. This is all the more likely, as the lack of significant differences actually means that any difference detected can most probably be attributed to chance.

5 Conclusions

Based on the above analysis, it is safe to claim that the performances on the original and the modified tasks, regardless of the required number of words, showed no significant differences for the text characteristics examined at any of the test levels. In other words, the different task versions generated responses that are no different in any respect except the required number of words. What follows from this observation is that out of the three potential problems discussed in Section 2, none actually materialized. While performances were shorter in the modified tasks, this, apparently, did not make the tasks easier; thus, the level of the exam did not change. It can also be asserted that the structure of the performances showed no discrepancies along the different versions of the tasks. Therefore, it can be demonstrated that the construct of the exam was not affected by the modifications in the tasks. Finally, the lack of significant differences in text characteristics also proves that the sample taken from candidates' proficiency was not significantly different in the modified tasks, so the content validity of measurement was retained. Based on the results of the study and the conclusions drawn, it can thus be stated that changing the required number of words did not result in any discernible difference in candidate performances, and, accordingly, it had no significant impact on the content and result of measurement.

As the study has demonstrated, automated text analysis in the framework of the Coh-Metrix platform can be utilized as a powerful tool for examining text characteristics for a variety of purposes. Whether in teaching or testing, it can aid text selection as well as a wide range of other assessment objectives by providing a highly detailed and fully objective picture of text properties. Using Coh-Metrix provides an opportunity for professionals to have additional support both in research and in teaching and testing activities. Hopefully, this resource will be utilized more frequently by practitioners in education, adding a further dimension to computer-based forms of teaching and assessment.

Proofread for the use of English by: Peter A. Sabath, Foreign Language Centre, University of Pécs

References

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

Aryadoust, V., & Liu, S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study. *Assessing Writing*, 24, 35–58.

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

Chun, D., Kern, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. *Modern Language Journal*, 100, 64–80.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Strasbourg: Council of Europe. URL: https://rm.coe.int/cefr-companion-volumewith-new-descriptors-2018/1680787989 (accessed March 27, 2020).

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

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., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35, 115–135.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

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.

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

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

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

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). Routledge.

Jarvis, S., Bestgen, Y., Crossley, S. A., Granger, S., Paquot, M. Thewissen, J. & McNamara, D. S. (2012). The comparative and combined contributions of N-grams, Coh-Metrix indices, and error types in the L1 classification of learner texts. In S. Jarvis and S. A. Crossley (Eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-Based Approach*. (pp. 154-177). Multilingual Matters.

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

McNamara, D.S., Graesser, A.C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2013). *Coh-Metrix version 3.0*. URL: http://cohmetrix.com (accessed: April 3, 2019)

Smith, B. E., Pacheco, M. B., & de Almeida, C. R. (2017). Multimodal codemeshing: Bilingual adolescents' processes composing across modes and languages. *Journal of Second Language Writing*, 36, 6–22.