Machine Translation

Post-Editing Neural Machine Translation vs. Translation Memory Segments

--Manuscript Draft--

Manuscript Number:	COAT-D-18-00047R4
Full Title:	Post-Editing Neural Machine Translation vs. Translation Memory Segments
Article Type:	S.I. : Human Factors in Neural Machine Translation
Keywords:	Neural machine translation; Translation Memory; quality perception; MT acceptance; translation productivity
Corresponding Author:	Pilar Sánchez-Gijón, Ph.D. Universitat Autònoma de Barcelona Bellaterra, SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universitat Autònoma de Barcelona
Corresponding Author's Secondary Institution:	
First Author:	Pilar Sánchez-Gijón, Ph.D.
First Author Secondary Information:	
Order of Authors:	Pilar Sánchez-Gijón, Ph.D. Joss Moorkens Andy Way
Order of Authors Secondary Information:	
Funding Information:	Ministerio de Economía, Industria y Competitividad, Gobierno de España (PRoJECTA-U, FFI2016-78612-R) — Dr. Pilar Sánchez-Gijón Science Foundation Ireland (IE) (Grant 13/RC/2106) — Dr. Joss Moorkens
Abstract:	The use of neural machine translation (NMT) in a professional scenario implies a number of challenges despite growing evidence that, in language combinations such as English to Spanish, NMT output quality has already outperformed statistical machine translation in terms of automatic metrics scores. This article presents the result of an empirical test that aims to shed light on the differences between NMT post-editing and translation with the aid of a translation memory (TM). The results show that NMT post-editing involves less editing than TM segments, but this editing appears to take more time, with the consequence that NMT post-editing does not seem to improve productivity as may have been expected. This might be due to the fact that NMT segments show a higher variability in terms of quality and time invested in post-editing than TM segments that are 'more similar' on average. Finally, results show that translators who perceive that NMT boosts their productivity actually performed faster than those who perceive that NMT slows them down.
Response to Reviewers:	All changes included.

Authors:

Pilar Sánchez-Gijón
Grup Tradumàtica - Universitat Autònoma de Barcelona, Barcelona, Spain
pilar.sanchez.gijon@uab.cat (corresponding author), https://orcid.org/0000-0001-5919-4629

Joss Moorkens
ADAPT Centre, School of Applied Language and Intercultural Studies, Dublin City University, Dublin, Ireland
joss.moorkens@dcu.ie, http://orcid.org/0000-0003-0766-0071

Andy Way
ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
andy.way@adaptcentre.ie, https://orcid.org/0000-0001-5736-5930

Title:

Post-Editing Neural Machine Translation vs. Translation Memory Segments

Abstract

The use of neural machine translation (NMT) in a professional scenario implies a number of challenges despite growing evidence that, in language combinations such as English to Spanish, NMT output quality has already outperformed statistical machine translation in terms of automatic metric scores. This article presents the result of an empirical test that aims to shed light on the differences between NMT post-editing and translation with the aid of a translation memory (TM). The results show that NMT post-editing involves less editing than TM segments, but this editing appears to take more time, with the consequence that NMT post-editing does not seem to improve productivity as may have been expected. This might be due to the fact that NMT segments show a higher variability in terms of quality and time invested in post-editing than TM segments that are 'more similar' on average. Finally, results show that translators who perceive that NMT boosts their productivity actually performed faster than those who perceive that NMT slows them down.

Keywords

Neural machine translation; translation memory; translation quality perception; MT acceptance; translation productivity.

1. Introduction

Neural machine translation (NMT) seems to be breaking the quality ceiling that previous machine translation (MT) systems were not able to overcome (Castilho et al. 2017; Way 2018). The quality scores obtained by state-of-the-art NMT systems have been improving over time compared with statistical machine translation (SMT) systems. Raw NMT output is perceived as being more fluent and natural than

SMT-proposed translations (Klubička et al. 2017), so it may be inferred that NMT ought to boost post-editing productivity more than when SMT is used to generate the draft output.

This article gathers together data about post-editing productivity using NMT. It tries to shed some light on whether professional translators perceive NMT as a technology that will boost their productivity. Following Moorkens and Way (2016), NMT post-editing is compared with translation assisted by translation memory (TM) with the aim of investigating whether improved MT quality will really result in increased productivity in a professional workflow based on MT post-editing. It presents a study that gathers information from EN—ES professional translators regarding their level of acceptance of NMT contrasted with TM, and any tangible improvement in performance when editing NMT segments rather than TM segments. The study aims to delve more deeply into the differences between editing MT and TM segments, and to examine the extent to which translators' perception of MT correlates with their performance indicators.

2. Motivation and related work

Although the use of MT is increasing within the translation industry, and particularly within the English to Spanish (EN—ES) market (Torres Hostench et al. 2015), many translators still mistrust the idea that the quality provided by MT systems can contribute to an increase in their productivity. Many studies show that, in general, NMT output can achieve higher scores according to a range of automatic metrics when compared to those obtained by other kinds of MT systems. However, only a few translators see a direct correlation between these scores and a boost in their productivity (Shterionov et al. 2018).

There are reports that allow the establishment of a direct correlation between domain-trained SMT systems and a boost to productivity in a professional workflow (Flournoy and Duran 2009; Pinnis et al. 2016). In particular, Pinnis et al. (2016) report that, following the introduction of an MT post-editing workflow with a domain-trained SMT system, there was a marked increase in translation productivity of 200%, although it was not specified whether TM was part of this scenario. In terms of productivity, Sánchez-Torrón and Koehn (2016) also report that for each 1-point increase in BLEU (Papineni et al. 2002) in raw EN—ES MT output, there is a decrease in editing time of 0.16 seconds per word. However, what remains unanswered is whether there is a BLEU score above which MT post-editing may be expected to outperform TM editing.

With regard to NMT use in professional environments, Castilho et al. (2017) give an overview of three studies carried out comparing NMT and SMT performance in terms of quality. One of the conclusions reached by these authors is the fact that, although NMT models can outperform phrase-based SMT with regard to automatic MT quality evaluation scores, they may not be preferred by human evaluators. In another study, Koehn and Knowles (2017) found that an EN—ES NMT system outperformed an SMT system in a study involving training data of 15 million words. The challenges of NMT pointed out by these authors include domain mismatch and rare words. In general, when NMT has access to sufficiently large amounts of training data in accordance with the domain and the vocabulary of the texts that have to be translated, it should outperform SMT. Segment length is also identified by Koehn and Knowles (2017) as an obstacle to obtaining high-quality NMT output; SMT outperformed NMT for sentences longer than 60 sub-word tokens in their EN—ES results.

Nonetheless, if MT (and NMT in particular) achieves – or will achieve in the near future – levels of quality that may substantially increase productivity, why are translators still reluctant to use it? We should not forget that translators were also reluctant to use CAT tools when they first appeared, but most now feel comfortable with them (Moorkens 2017). Other different reasons may come into play, such as the possibility of editing different types of text files; the re-use of human-edited translations; integrated solutions that simplify translation project management and teamwork; as well as data mining and quality assurance (Martín-Mor et al. 2016). Most of these reasons are connected with the fact that (N)MT output editing implies the use of simple editors that do not include many of the features of current CAT tools. Such tools empower translators with features that assist them in their everyday tasks (O'Brien 2012), whereas MT is sometimes used by the industry to cut translation costs by reducing translators' interventions or even producing unedited and unreviewed MT output (Sánchez-Gijón 2016). Other studies

point out that one of the reasons why translators are still reluctant to use MT is their perception of its quality. Aside from its perceived lower quality and the greater effort needed when post-editing output, translators are reluctant to use MT because of falling pay rates, copyright ownership for the resources created during a translation project (Moorkens et al. 2016), as well as technological integration issues with the CAT tools they are already using (Martin-Mor et al. 2016). They feel that MT negatively affects their productivity (Cadwell et al. 2016), or they may just not trust it (Rossetti and Gaspari 2017).

Moorkens and Way (2016) conducted a previous study from English to German which established that high-quality MT (SMT in their case) was preferred by translators instead of a low-quality TM, i.e. TM with a low rate of fuzzy matches. Following their approach, this paper seeks to dig deeper in tackling the same questions, using high-quality TM matches and recording editing effort (technical and temporal effort, Krings 2001). Particular attention will be paid to some issues that have already been identified as challenges for NMT, such as sentences of differing lengths (Koehn and Knowles 2017; Castilho et al. 2018) and quality perception (Cadwell et al. 2016; Rossetti and Gaspari 2017).

3. Methodology

In order to collect data about translators' performance and perceptions with regard to NMT, an empirical test was set up consisting of three different parts: answering a general questionnaire, assessing the quality of a set of segments produced using MT and TM, and editing a second set of segments.

Eight professional EN—ES translators took part in this test. They had regular post-editing experience, varying from 2 to 32 years (with an average of 16 years, and a median of 8 years). They were asked to answer a questionnaire about their general perception of NMT quality and about their performance when using it in post-editing projects. They were then asked to assess the quality of a set of 30 translated segments using the following commonly-used scale (see Moorkens 2018):

1. Needs retranslation.
2. Some editing is needed.
3. Light or no editing is needed.

Finally, they post-edited a set of 30 different segments using the Kanjingo Flow tool (Teixeira et al. 2019). Translators were asked to carry out a blind translation task, editing translation suggestions without knowing whether they came from an MT system or from a TM.

For the post-editing task, two different sets of 30 segments from Autodesk EN—ES TM were used. This text type was chosen because product documentation is one of the content types in which post-editing is regularly applied, even outscoring human translation (Lommel and DePalma 2016). Each set included 15 TM segments and 15 source segments from an Autodesk TM translated into ES using Bing, Microsoft's free online NMT system.[1] The segments used for TM matching and those translated using MT are homogeneous, in that they exhibit similar characteristics that are well within standard deviation using common corpora analyses in the WordSmith WordList tool.[2] For example, the type/token ratio of lexical variation is 56.42 for TM data and 53.77 for MT data, and mean word lengths are 4.81 and 5.1 characters respectively. TM segments for each set included fuzzy matches under 80%, between 80% and 90%, and over 90% (from 90% to 95%). This test did not include any exact matches. Unfortunately, one of the TM segments had to be withdrawn due to an error in creating the XLIFF resource file.

Table 1 presents a global description of the entire group of 59 remaining segments that made up test set 1 and test set 2.

[1] https://www.bing.com/translator
[2] https://lexically.net/wordsmith/

	Less than 10 words	From 10 to 19 words	More than 20 words	Total
MT	5	17	8	30
	17%	57%	27%	100%
TM	5	18	6	29
	17%	62%	21%	100%

Table 1. Segments related to the number of source words

With regard to TM segments, Table 2 describes them by considering both their length and their fuzzy match percentage.

	Less than 10 words	From 10 to 19 words	Over 20 words	Total
TM fuzzy matches under 80%	3	4	0	7
	10%	14%	0%	24%
TM fuzzy matches from 80% to 89%	2	12	2	16
	7%	41%	7%	55%
TM fuzzy matches over 90%	0	2	4	6
	0%	7%	14%	21%

Table 2. Description of TM source segments

As can be seen in Table 2, there were no segments with a fuzzy match proposal of over 90% with less than 10 words. There were no segments with a fuzzy match of under 80% and more than 20 words either. The latter point was discovered after the evaluation had been conducted. It was also found that the MT output included the longest segments used during this test (54 and 49 words each).

These 59 segments (30 from MT and 29 from TM) were proportionally divided into two sets. One set contained 30 segments (15 from MT and 15 from TM) while the other set contained one segment less (15 from MT and 14 from TM). Segments were distributed as shown in the Appendix.

Participants were asked to evaluate one of the segment sets, assigned randomly, and to post-edit the other one to avoid any unwanted learning effect. The data collected for each of the segments consisted of:

- Post-editing time: time invested by participants in post-editing each segment through Kanjingo Flow.
- Post-editing distance.[3]
- Post-edited characters: the number of insertions, deletions or substitutions introduced in the post-edited segments.
- Quality assessment of each of the proposed translations.

The data obtained allowed a comparison between MT and TM segments globally, with regard to all these indicators, as well as a more detailed analysis taking into account the quality level of the proposed translation as perceived by translators, the fuzzy match percentage of the TM segments, and the number of words in the source segments, as well as the translators' perception of the effect of MT on their productivity. The comparison among MT segments and each of the fuzzy matched sets (under 80%, from 80% to 90%, and over 90% as presented in Figure 5) took into account only MT segments within the same range (similar number of words) as the TM segments with which they were being compared. TM segments with a fuzzy match score of under 80% contained segments of up to 19 words; TM segments with a fuzzy match from 80% to 90% contained segments of up to 40 words; and TM segments with a fuzzy match over 90% contained segments with 14 to 31 words.

[3] This is approximated using the difflib library (see https://docs.python.org/3/library/difflib.html), which measures the difference between two strings, rather than the minimum number of edits required to transform one string to another.

Analyses were carried out comparing only two dependent samples each time. For that purpose, a Wilcoxon rank sum test with continuity correction was run in order to assess whether the data from each of the two compared groups were statistically different ($p < 0.05$).

4. Results

In this section, we analyse the data obtained in three different ways. Firstly, attention is paid to the quality of the proposed translations as perceived by translators. Secondly, the analysis focuses on performance bearing in mind the origin of the edited proposed translations. Finally, data regarding performance is approached from the point of view of the translators' perception of any change in their productivity using MT.

4.1 Quality perception

The quality of each of the segments was assessed by more than one participant. For that purpose, a scale of three categories was used. Participants had to choose from the three categories as described in Section 3 (1st category: "Needs retranslation"; 2nd category: "Some editing is needed"; 3rd category: "Light or no editing is needed"). Sometimes participants did not agree, but still they chose categories that were sequential and close to each other (1st and 2nd, or 2nd and 3rd). Some other segments were assigned to categories that were not sequential (1st and 3rd). Finally, some segments were assigned to all three categories. Table 3 summarizes the percentage agreement among participants assessing segment quality.

	MT	TM	Percentage
Segments assigned to all three assessment categories	1	2	5.08%
Segments assigned to two non-sequential categories	3	0	5.08%
Segments assigned to two sequential categories	17	21	64.41%
Full **agreement**	9	6	25.42%

Table 3. Agreement when assessing segments

As can be seen, over 89% of the segments were assessed as belonging to the same category or into two sequential categories (1st and 2nd categories, or 2nd and 3rd categories), while the remaining 10% or so did not achieve this level of agreement (1st and 3rd, or 1st, 2nd and 3rd).

Figure 1 shows the assessment average for each segment bearing in mind the origin of the proposed translation. All segments are perceived as middle- or high-quality proposed translations, except for a small percentage of TM segments with a low fuzzy match (fuzzy matches under 80%) and of MT segments.

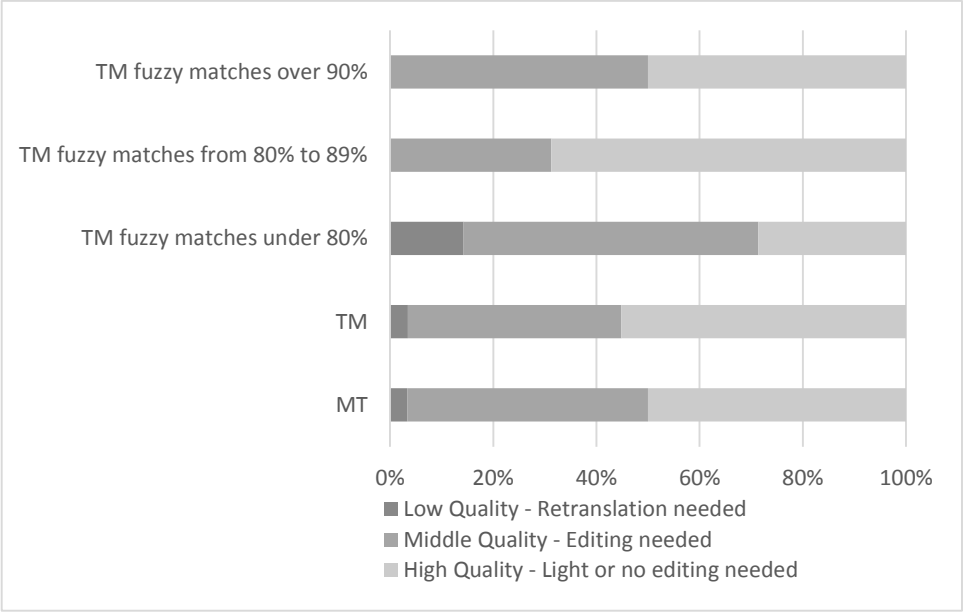


Figure 1. Segment quality averages with regard to the origin of the proposed translation

TM fuzzy matches over 90% seem to be perceived as segments that need greater editing than TM fuzzy matches from 80% to 89%. This might be due to the fact that all fuzzy segments over 90% are over 10 words, while shorter segments may need very little editing even though their fuzzy match was in the region of under 90% similarity. Bearing in mind the number of words in the source segments, the results show some different nuances that may be worth describing. In this case, our analysis will take into account not the assessment average for each segment, but rather all the individual assessments for each segment.

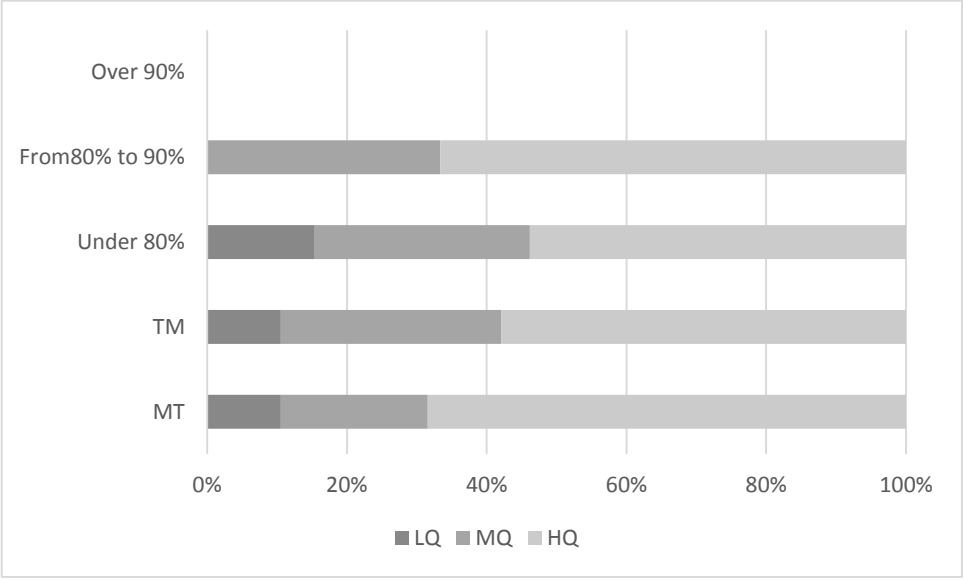


Figure 2. Quality assessment of segments of under 10 words with regard to the origin of the proposed translation

For segments under 10 words, Figure 2 shows that only a low percentage of MT segments (11%) and of TM segments with a low fuzzy match (fuzzy matches under 80%) were assessed as low-quality proposed translations. In contrast, among the high-quality segments (MT segments and TM segments with a fuzzy match percentage between 80% and 90%), the best score achieved was 68% and 67%, respectively. There were no fuzzy matches over 90% among TM segments under 10 words. Accordingly, we can see that MT and TM in the 80% to 90% range present the same amount of high-quality segments, according to our raters. However, a small amount of MT segments may be perceived as low-quality suggestions, while this does not happen with TM fuzzy matches from 80% to 90%.

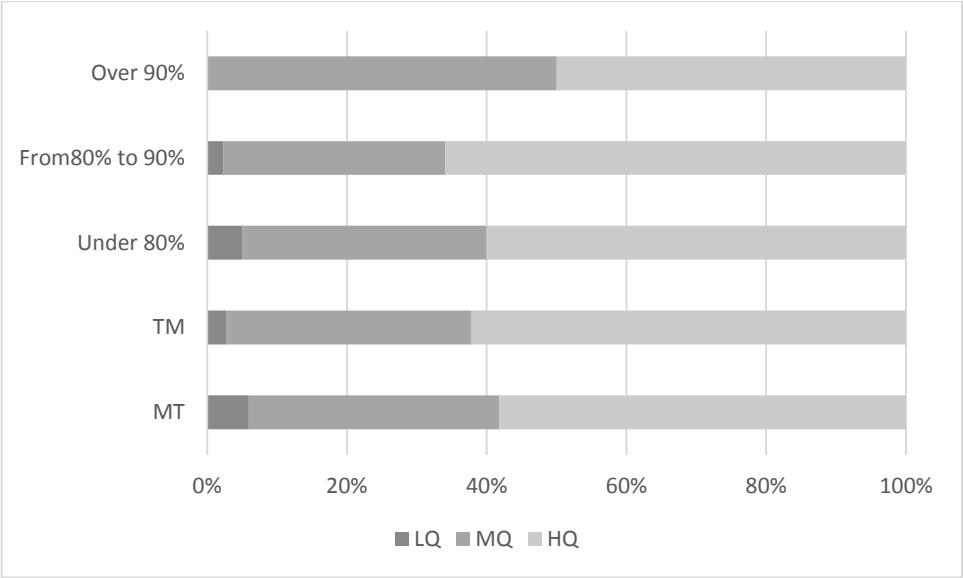


Figure 3. Quality assessment of segments of between 10 and 19 words with regard to the origin of the proposed translation

Regarding longer segments (from 10 to 19 words), only segments over 90% are perceived as medium- or high-quality segments. In all other cases, they might also be perceived as low-quality segments, particularly in the case of MT segments. Comparably, MT segments and TM fuzzy matches under 80% are rated very similarly.

In contrast, as can be inferred from Figure 4, for segments of over 20 words TM scores better, and particularly on segments with a fuzzy match over 90%. Unfortunately, the set of TM segments used in this study did not include any fuzzy matches under 80% from segments with more than 20 words.

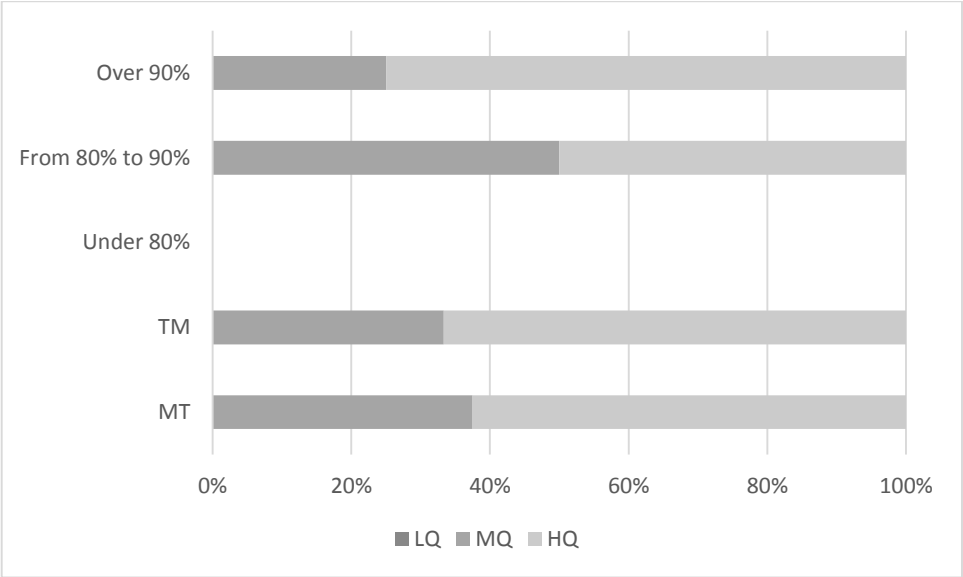


Figure 4. Quality assessment of segments of over 20 words regarding the proposed translation origin

It is of particular note that there are no segments perceived as being low-quality suggestions. Focusing on MT segments, this would mean that these MT suggestions are valid for assimilation purposes, and that they are edited merely to achieve a 'more human' quality.

In sum, quality perception averages show that only TM segments with a fuzzy match of under 80% obtained worse results than MT segments, while MT segments scored better for short segments (under 10 words) than for longer ones, even though they are rarely perceived as low-quality translations *per se*. From this point of view, the perceived quality of MT and TM segments is comparable and ought not to justify any huge differences in terms of productivity or acceptability.

4.2 Performance

Performance is analysed in this section with regard to post-editing time, edit distance, and characters edited. The results presented in this section indicate trends that on some occasions demonstrate statistical significance, even though the amount of data analysed may not allow us to reach any definitive conclusions.

4.2.1. MT vs. TM results for post-editing time

When it comes to post-editing time (or temporal effort), no significant difference can be seen when comparing editing times for MT and TM segments. This is true of all segments, whether short, medium, or long, as shown in Table 4.

	MT	**TM**
All segments	22.53	20.47
Segments of under 10 words	13.47	15.08
Segments from 10 to 19 words	20.63	16.54
Segments of over 20 words	33.24	35.40

Table 4. PE average editing time (in seconds) of MT vs. TM segments

Nevertheless, if close attention is paid to the different TM segments with regard to their fuzzy match percentage vs. MT segments, the results show that, although differences are not significant for either, values expressing time invested in MT post-editing are more dispersed than in the higher quartiles[4] (Q3 and Q4), while they show less dispersion than in Q1 and Q2 values, i.e. the values between the lowest and the median register are very close, meaning that segment editing was homogeneously fast among the different post-editors. None of these differences are statistically significant.

[4] A quartile is a kind of quantile. It's a descriptive measure. The first quartile (Q1) is the middle value of a range between the smallest value and the median. The median corresponds to the second quartile (Q2). The third quartile (Q3) corresponds to the 75% of the range, and the fourth quartile (Q4) to the 100%. Quartiles help describing the distribution of the population of the sample.

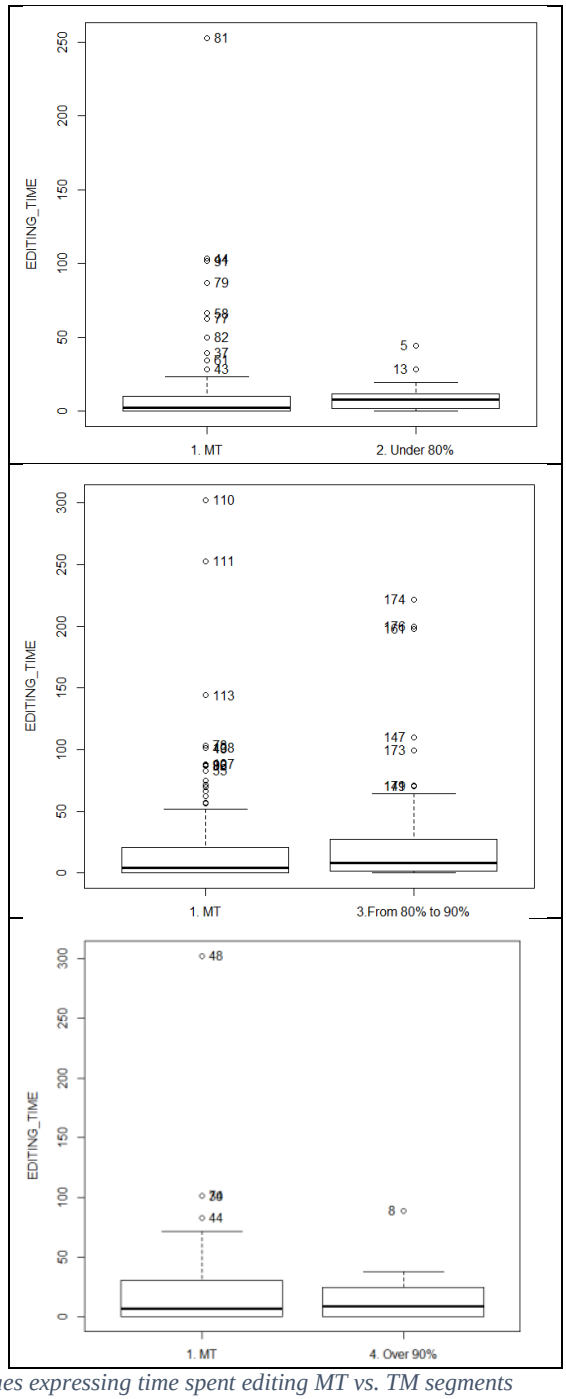


Figure 5. Distribution of values expressing time spent editing MT vs. TM segments

Figure 5 features a comparison among MT segments and each of the fuzzy matched sets as described in Section 3. The first chart in Figure 5 (MT vs. TM under 80%) shows that the first 50% of MT segments were completed very quickly as well as being homogeneously edited.[5] At the same time, the first 50% of the MT segments were edited more rapidly than the first 50% of the TM segments. The opposite occurs with the second 50% of the segments; TM Q3 and Q4 values are very close, and there are just two values over Q4, whereas MT Q3 and Q4 are farther from the median value of the sample and there are many outlier values. That means that editing TM fuzzy matches under 80% seems to imply a constant time investment, while editing MT segments is not as consistent, so in some cases may need a shorter time investment, while in others they may need a very much longer one. The second chart compares the editing

[5] The Q2 value and the average of the MT set are very close to each other and are lower than those of TM under 80% set.

times for all MT segments against those of TM segments with fuzzy matches from 80% to 90%. In this case, all values indicate that editing MT segments is slightly faster than editing TM segments. Dispersion is similar in both cases, even though MT segment editing presents a greater number of higher values (meaning that there are some MT segments that took much longer to edit than TM segments). The third chart (MT vs. TM over 90%) also shows that MT editing is faster on average, even though it also includes those segments that took longer to be edited.

We infer that this lack of homogeneity in editing MT segments may contribute to the reluctance shown by translators. The lack of predictability of MT output quality has been highlighted previously (e.g. by Moorkens and Way 2016), and is the basis for work on confidence or quality estimation and research on visual representation of confidence within the interface (Alabau et al. 2013; Moorkens et al. 2015) in order to make MT post-editing more acceptable to translators.

4.2.2. MT vs. TM results for edit distance

Unlike temporal post-editing effort results, MT data concerning edit distance (or technical effort) would suggest that MT boosts productivity. The results are more homogeneous and show lower average values than the TM data. This means that the global editing distance of MT segments is smaller than that of TM segments.

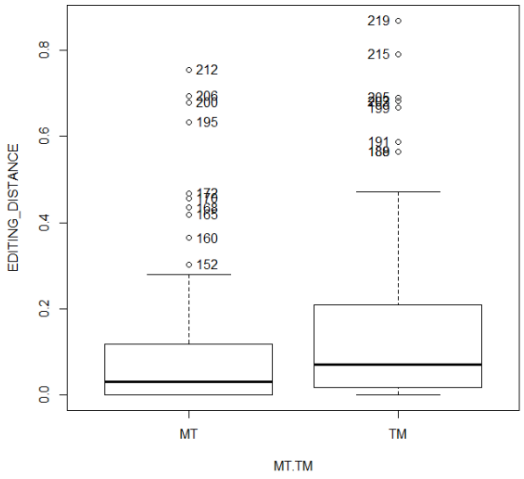


Figure 6. Distribution of edit distance values for MT vs. TM segments

The differences between MT segments and TM segments in terms of edit distance (Levenshtein 1966) are statistically significant. The edit distance of at least 25% of the MT segments was 0, meaning that no editing was needed. Following the progression, MT edit distances show lower values in all quartiles. Therefore, edit distance is significantly lower for MT when compared with TM segments, where Z = 5250.5, p = 0.001449.

Figure 7 shows MT and TM edit distance values categorised by source segment length. The above findings are confirmed; in all cases, Q1, Q2 and Q3 values for MT segments are lower than those for TM segments. Only Q4 values in segments of over 20 words differ, probably because the two longest source segments belonged to the MT sample.

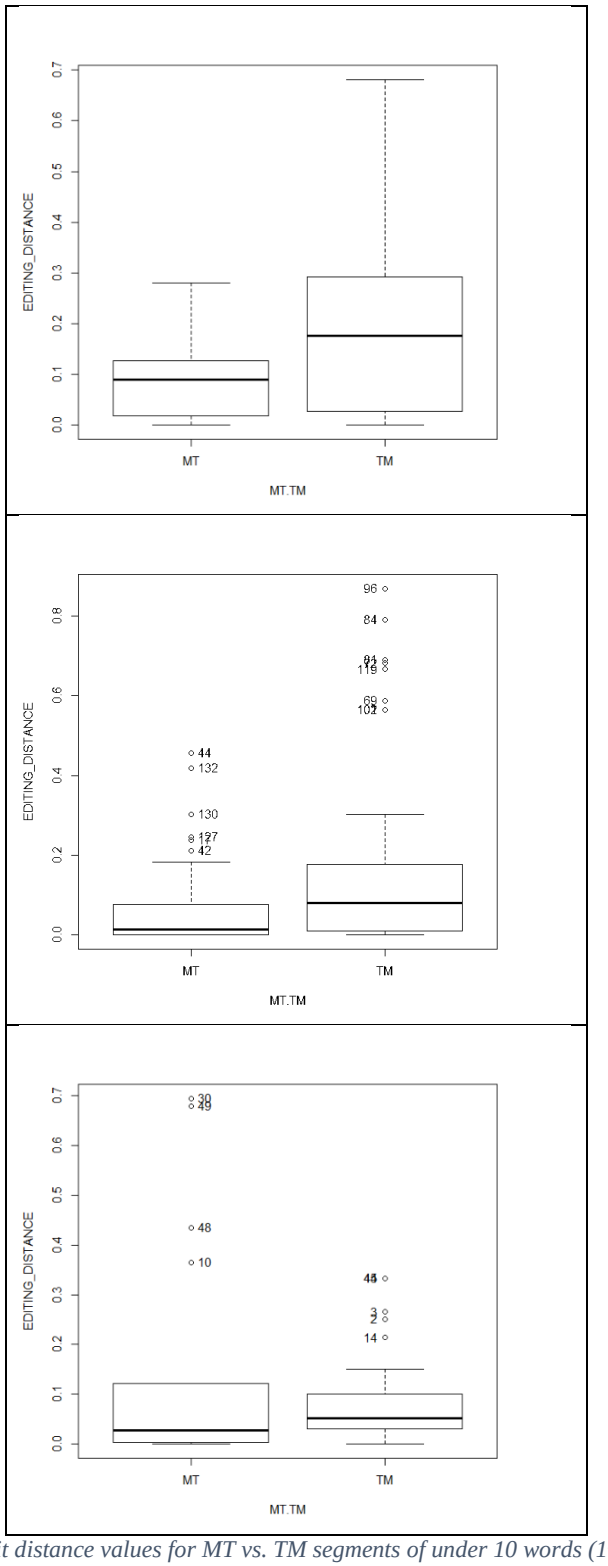


Figure 7. Distribution of edit distance values for MT vs. TM segments of under 10 words (1st chart), from 10 to 19 words (2nd chart), and of over 20 words (3rd chart)

Figure 7 shows that in all cases MT edit distance is lower than for TM segments. None of these differences, however, are statistically significant, except for segments of length of from 10 to 19 words ($Z = 1551.5$, $p = 0.0006214$). It may be worth pointing out that segments from 10 to 19 words represent 60% of the whole sample, so the amount of data related to this kind of segment is higher than those from longer and shorter segments.

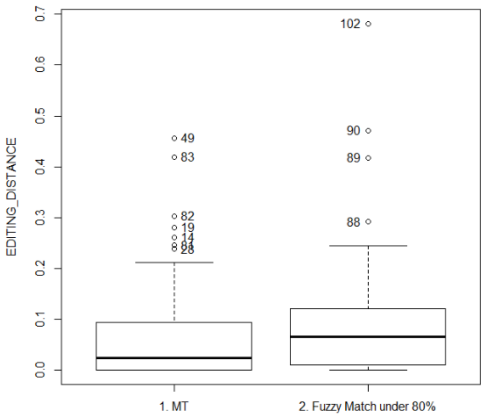


Figure 8. Distribution of edit distance values for MT segments of under 19 words vs. Segments with TM fuzzy matches under 80%

MT edit distance is lower than that of TM fuzzy matches. Comparing MT segments against segments with TM fuzzy matches of under 80% (Figure 8), these differences are not significant. In contrast, when comparing MT segments with TM fuzzy matches from 80% to 90% (see Figure 9), the edit distance for MT segments is significantly lower than that of the TM set ($Z = 2724.5$, $p = 0.0004579$). This means that the amount of valid text from the proposed translation is significantly higher in the case of MT segments than in the case of TM fuzzy matches from 80% to 90%. Post-editors needed to perform fewer edits on MT segments than on TM fuzzy matches from 80% to 90%.

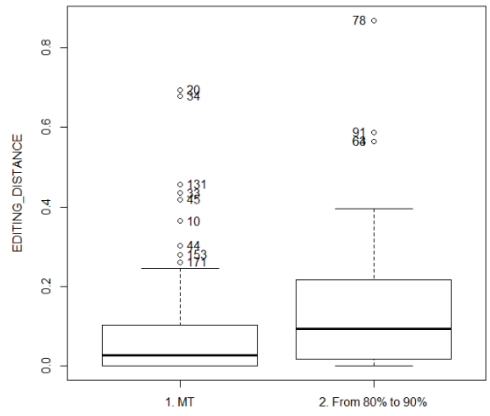


Figure 9. Distribution of edit distance values for MT segments under 40 words vs. segments with TM fuzzy matches from 80% to 90%

Comparing the edit distance for MT segments with TM fuzzy matches of over 90%, the results are not so clear. As Figure 10 shows, even though the MT segments have lower average values. These results are not statistically significant, but this irregular distribution of MT results coincides with the small percentage of MT segments perceived as being of low quality (see Figure 1 in Section 4.1).

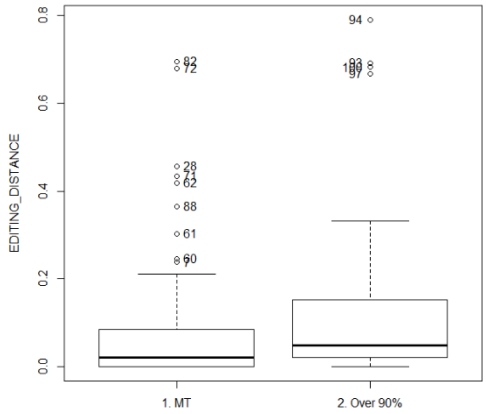


Figure 1. Distribution of edit distance values for MT segments of over 10 words long vs. segments with TM fuzzy matches of over 90%

In summary, MT edit distance is lower than for TM fuzzy matches in all categories. Furthermore, the distribution of the results of MT edit distance in all scenarios shows that, in most cases (50% or even 75%), edit distances for MT are far more consistent than the corresponding TM editing distances.

4.2.3. MT vs. TM results for characters edited

In terms of deletions, insertions and substitutions, there are no significant differences between edits of MT and TM segments globally, although MT editing seems to be more productive. However, when only those MT segments under 40 words are considered (and both TM and MT segments sets are comparable in terms of length), differences arise that are statistically significant ($Z = 4939.5$, $p = 0.001343$).

Globally, editing MT segments required the introduction of fewer changes than for TM segments. If we focus on segment lengths, differences are not significant, although MT results show the same pattern (see Figure 11).

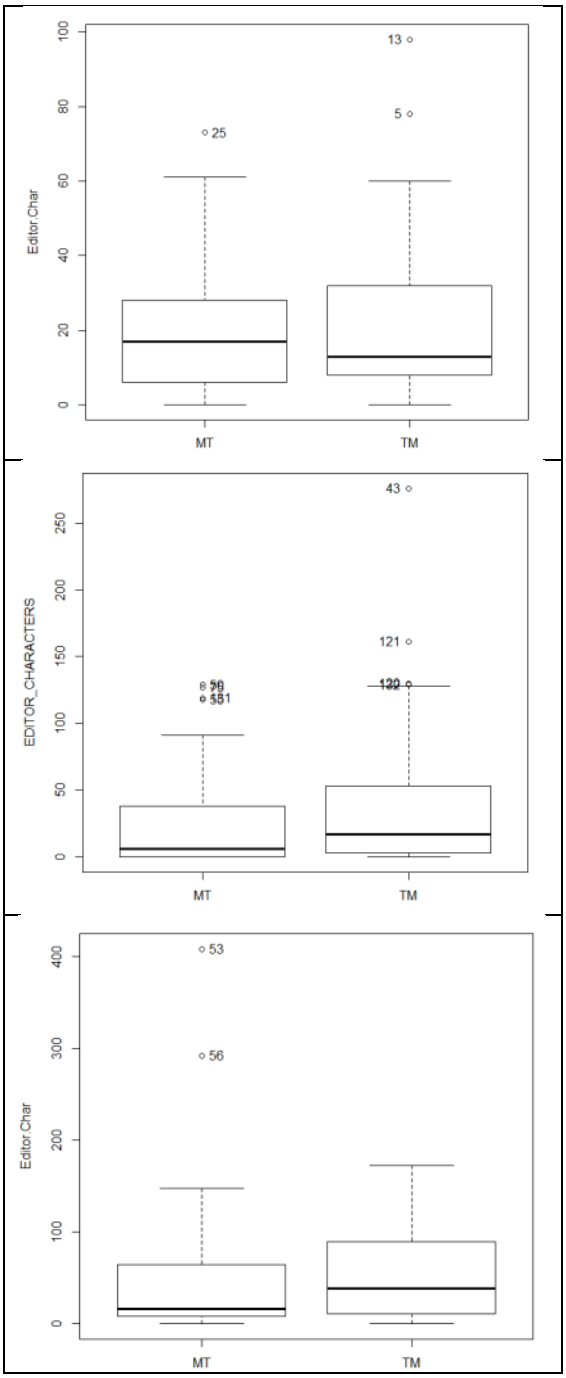


Figure 11. Distribution of edited character values for MT vs. TM segments of under 10 words (1st chart), from 10 to 19 words (2nd chart), and of over 20 words (3rd chart)

The first chart in Figure 11 (concerning segments of under 10 words) shows that MT and TM data are very similar, even though the average for MT segments is slightly lower. As for the second chart dealing with segments from 10 to 19 words, and the third chart showing segments of over 20 words, the results are not significant, but are very similar. In this case, all average MT values are lower, but there are some values above Q4, meaning that some segments took much longer to edit, representing a statistically abnormal distance compared with most of the sample.

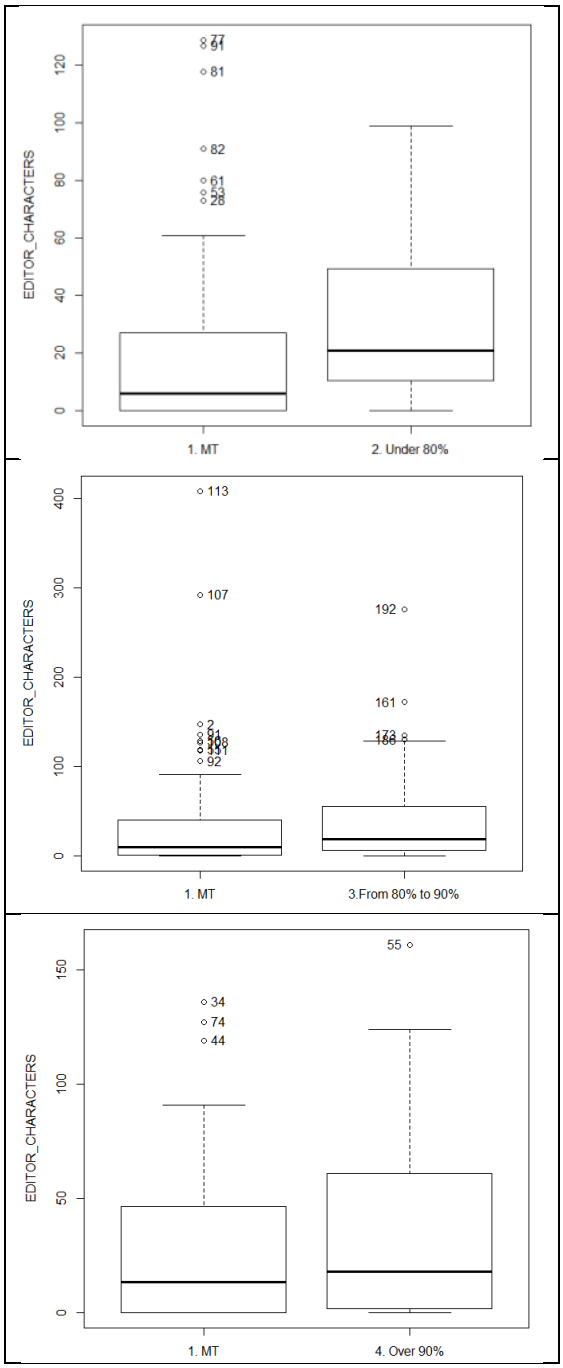


Figure 12. Distribution of edited character values for MT segments vs. segments with TM fuzzy matches of under 80% (1st chart), from 80% to 90% (2nd chart) and of over 90% (3rd chart)

Finally, Figure 12 shows that significant differences were not observed between MT data and those from TM with different levels of fuzzy matches. Again, the MT figures are lower and show less dispersion (average values are closer) than any TM values, although the MT figures show a higher level of values above that of Q4. This would suggest that editing 50% or even 75% of the MT sample needs a progressively increasing amount of editing. However, the final 25% of the MT sample requires a proportionally higher amount of editing than the rest of the sample. These results coincide with the percentage of MT segments perceived as being of low quality.

Nevertheless, as Table 5 shows, all MT average values and IQR values[6] are lower than the TM fuzzy match values with which they are being compared. That means that editing MT segments is more homogeneous in terms of keystrokes (IQR values are smaller) than editing TM segments. These results show no statistical significance.

	Mean	Sd	IQR
MT (segments under 19 words)	20.11	30.54	26.75
TM fuzzy matches under 80%	**31.17**	**28.18**	**39.00**
MT (all segments)	31.62	54.51	39.00
TM fuzzy matches from 80% to 90%	39.79	49.70	48.00
MT (segments from 14 to 40 words)	29.55	35.56	45.75
TM fuzzy matches over 90%	**36.95**	**46.46**	**55.75**

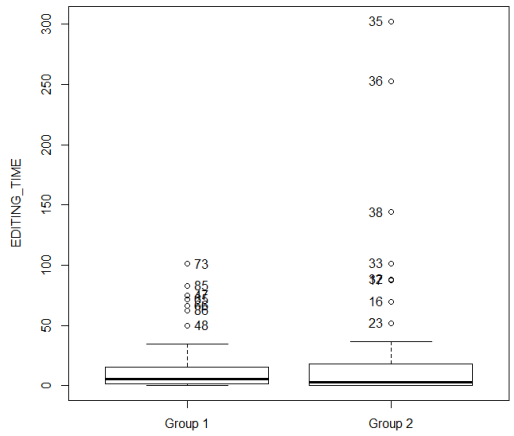
Table 5. Mean Standard deviation and IQR values in regard to edited characters of MT segments vs.TM fuzzy matches

In sum, performance in terms of the amount of edited characters shows that MT editing involves less effort than editing any kind of TM fuzzy matches (under 80%, from 80% to 90% and over 90%). Most of the MT sample needs consistently less editing except for a small part of the sample (mostly outlier values) which required as much editing or even more editing than TM segments. This lack of consistent amounts of editing needed in this small part of the MT sample could lead translators to perceive that these MT segments were of low quality.

4.3 MT perception and performance

Sections 4.1 and 4.2 established that MT segments take on average less time and a lower number of edits that TM segments, even though a small number of MT segments need longer to be edited than TM segments. In order to check whether translators' perception of MT correlates with their own experience, participants were asked to describe their agreement with the sentence "In translation projects, MT slows me down and I prefer not to use it" through a Likert-type 5-point scale, in which 1 stood for "I fully disagree" and 5 for "I fully agree". Three of the participants were in partial or total disagreement with this statement (answers 1 and 2), while three of them partially or totally agreed (answers 4 and 5). Two participants did not state whether they agreed or disagreed (answer 3).

MT performance results were divided into two different sets in relation to the participants' perception. Group 1 was defined as participants who perceived that MT does not tend to decrease their productivity, while Group 2 includes results from participants who stated that MT does tend to decrease their productivity. Results from the participants who did not state a preference are not considered for the remainder of this section.



[6] The interquartile range (IQR) measures statistical dispersion. A low IQR stands for a low level of variability.

Figure 13. Distribution of MT post-editing time values with regard to participants' perception

Each participant post-edited 15 MT segments, so Group 1 and 2 included 45 segments each. Figure 13 shows the differences between the post-editing time of both participant groups with regard to all MT segments. Even though differences are not very evident, Group 2 presents a higher number of values clearly above Q4. When observing this data more carefully (see Table 6), the mean editing time for MT segments by participants who do not perceive that MT lowers their productivity is far below that of the participants who did perceive MT as a handicap in terms of productivity. Group 1 results for MT post-editing time show a progressive increase from the lowest value (Q1, 25%) to the highest (Q4 and outliers, 100%) meaning that the progression is fairly homogeneous. In contrast, Group 2 results show a homogeneous progression only until Q2 (the first 50% of the sample), but not in the second part of the sample (values regarding 75% and 100%).

	Mean editing time	Sd	IQR	0%	25%	50%	75%	100%
Group 1	17.83	26.00	14.31	0	1.34	5.41	15.65	101.11
Group 2	29.32	63.04	18.20	0	0.00	2.70	18.20	302.59

Table 6. Summary of statistics comparing MT segment editing time in relation to both groups of participants

Even though these results are not statistically significant, it struck us that the participants' perception of NMT post-editing effort matched their performance in general. Studies of SMT post-editing mostly found users' perceptions to be a poor predictor of PE effort (Plitt and Masselot 2010; Läubli et al. 2013).

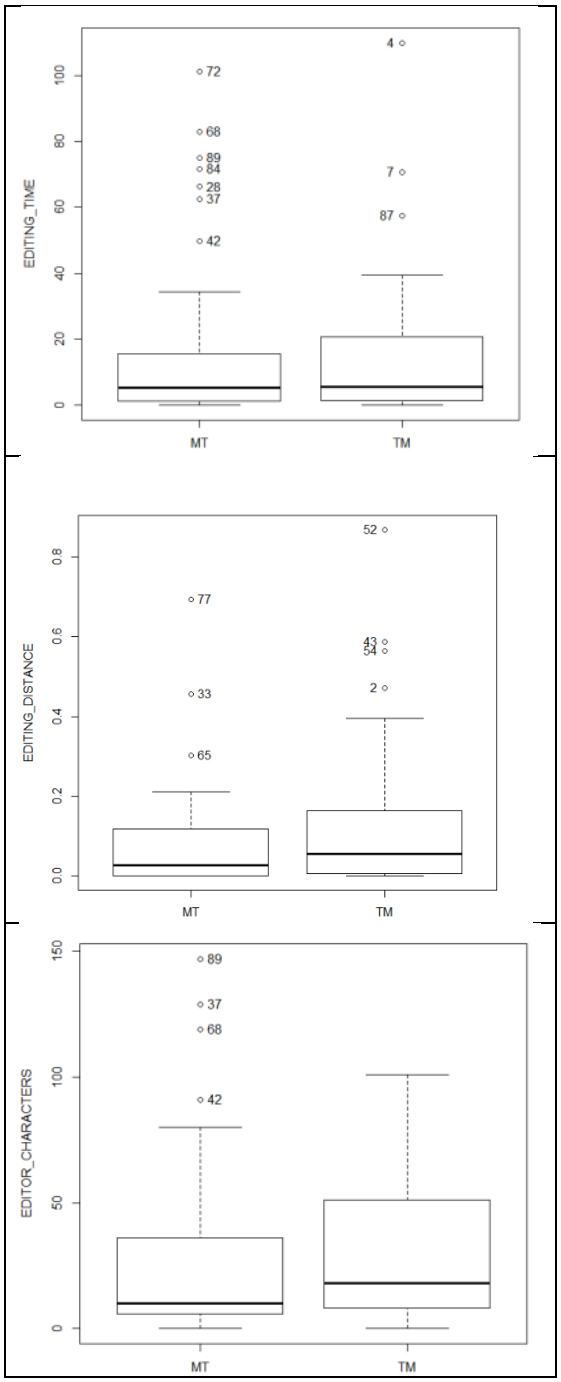


Figure 14. Distribution of Group 1 results regarding editing time (first chart), editing distance (second chart) and edited characters (third chart)

As can be inferred from Figure 14, Group 1 showed more productive results when editing MT-assisted translations than TM-assisted translations, both in terms of editing time, edit distance, and number of edited characters. The first chart (editing time) shows only slight differences between MT and TM results. In terms of edit distance (second chart), differences are not statistically significant ($Z = 817.5$, $p = 0.2085$). In conclusion, as they have already perceived, these translators seem to be more productive using MT than TM in their translations, although the differences between both groups are not significant.

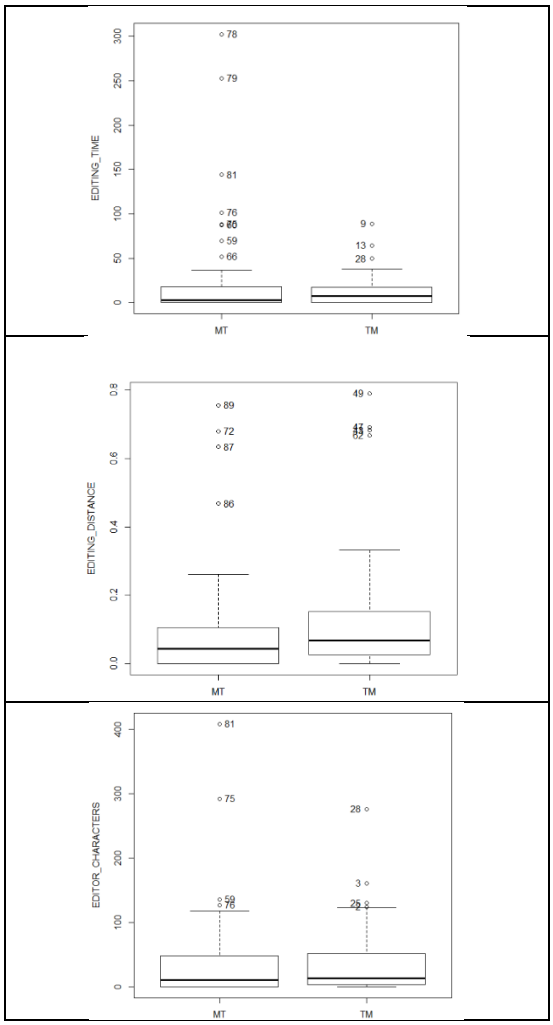


Figure 15. Distribution of Group 2 results in regard to editing time (first chart), editing distance (second chart) and edited characters (third chart)

Unlike Group 1, Group 2 results do not demonstrate the same degree of MT productivity. In terms of editing time (Figure 15, first chart), the average MT value is substantially higher than the TM mean value (29.32s vs. 12.78s). MT also presents many values over Q4. As regard to edit distance (Figure 15, second chart), MT shows some better results on average without outperforming TM results. Finally, as for edited characters (Figure 15, third chart), both MT and TM results are very similar, but MT results include higher values than TM results.

Comparing Figure 14 and Figure 15 allows us to conclude that, with regard to this data, translators' perception of the effect of MT on their productivity is actually also reflected in their performance, even though they were asked to carry out a blind task.

In order to shed some light on where these differences are obvious, the performance results of both groups were examined with regard to segment length. None of these results show statistical significance, probably because sample sizes are quite small.

Group 1 Group 2

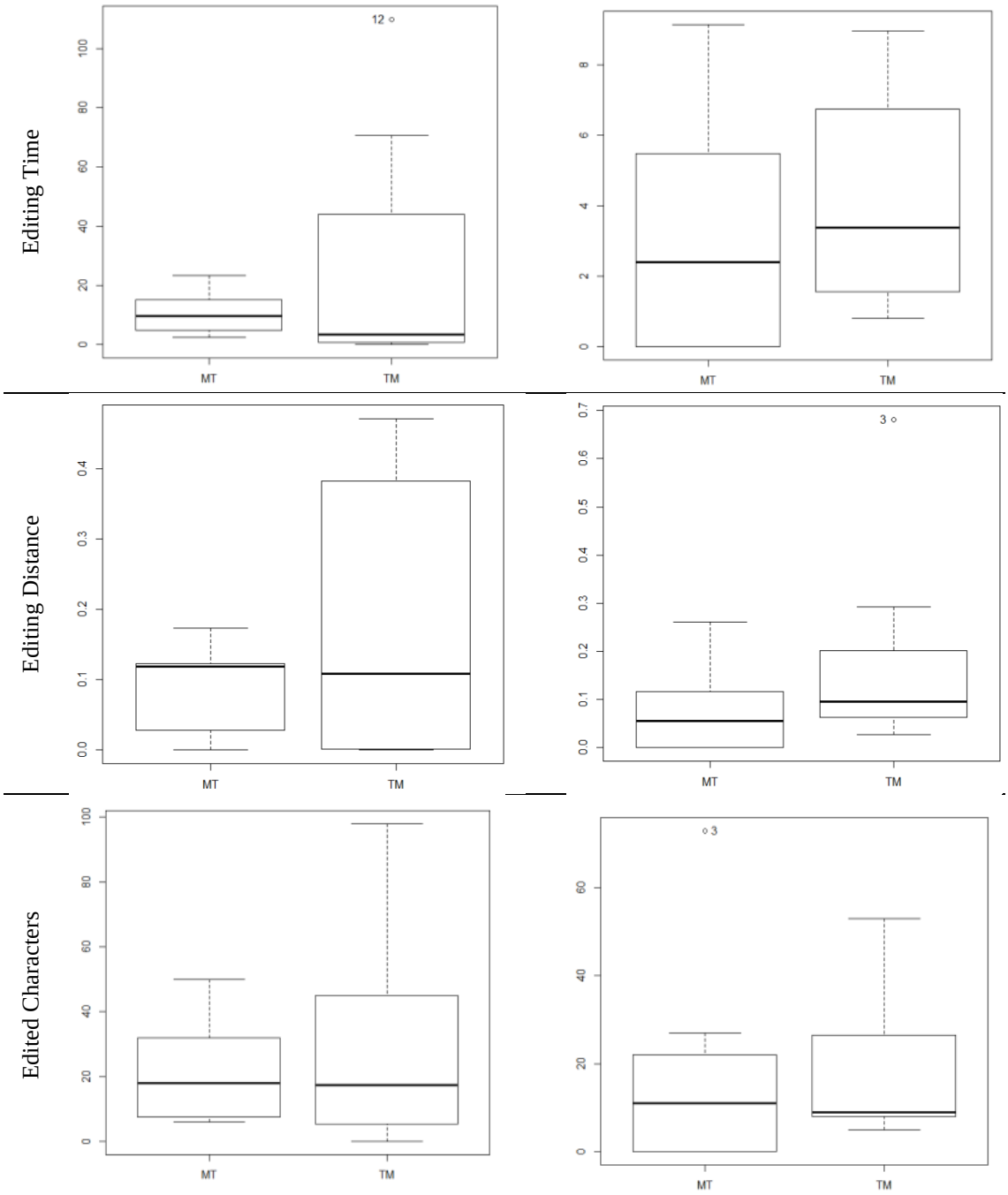


Figure 16. Performance on segments under 10 words

Figure 16 shows the results of both groups with regard to segments under 10 words. The results for Group 1 (who perceive that MT helps their productivity) clearly favour MT in terms of editing time, which globally has been an indicator independent of results for edit distance and characters edited. However, in Group 2 (who perceive that MT hinders their productivity), even though it can be inferred that MT requires less editing (in terms of edit distance and number of edited characters) the differences in terms of editing time are not so clear: on average, MT seems to be slightly faster, but results are not very consistent. The MT results for Group 2 also show a higher level of dispersion and are closer to the TM results in Group 1. Accordingly, Group 1 does obtain a perceptible enhancement using MT, while Group 2's performance does not really show much difference between MT and TM, with any difference being mostly in terms of editing time.

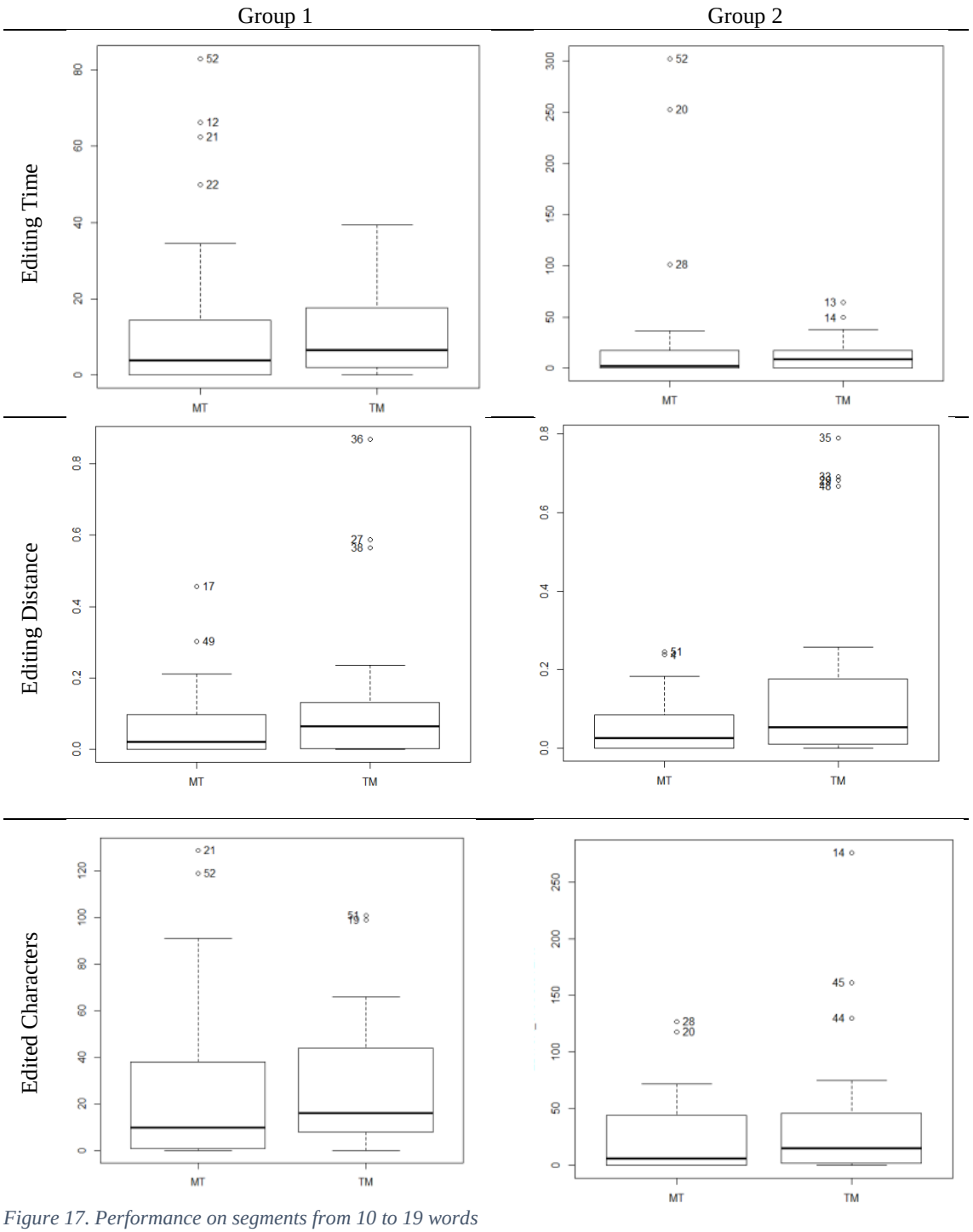


Figure 17. Performance on segments from 10 to 19 words

Figure 17 shows the results of both groups considering segments from 10 to 19 words. In this group of segments, MT editing time shows high values both for Group 1 and for Group 2. Group 1 shows lower MT values on average than TM values. The same thing occurs with Group 2 results, even though in this case there are some high values indicating that particular TM segments took very much longer to edit. When we turn to edit distance, both groups show substantially lower values in MT Q1 and Q2.

In sum, results for these test sets of segments show similar ranges for both MT and TM results in both groups, and MT dispersion is higher in the second half of each range where there are outlier values, while results for the first part of the range are better. This allows us to infer that a few MT segments show

worse results than TM segments, but on average many MT segments perform as well as – or even better than – TM-proposed translations in both groups.

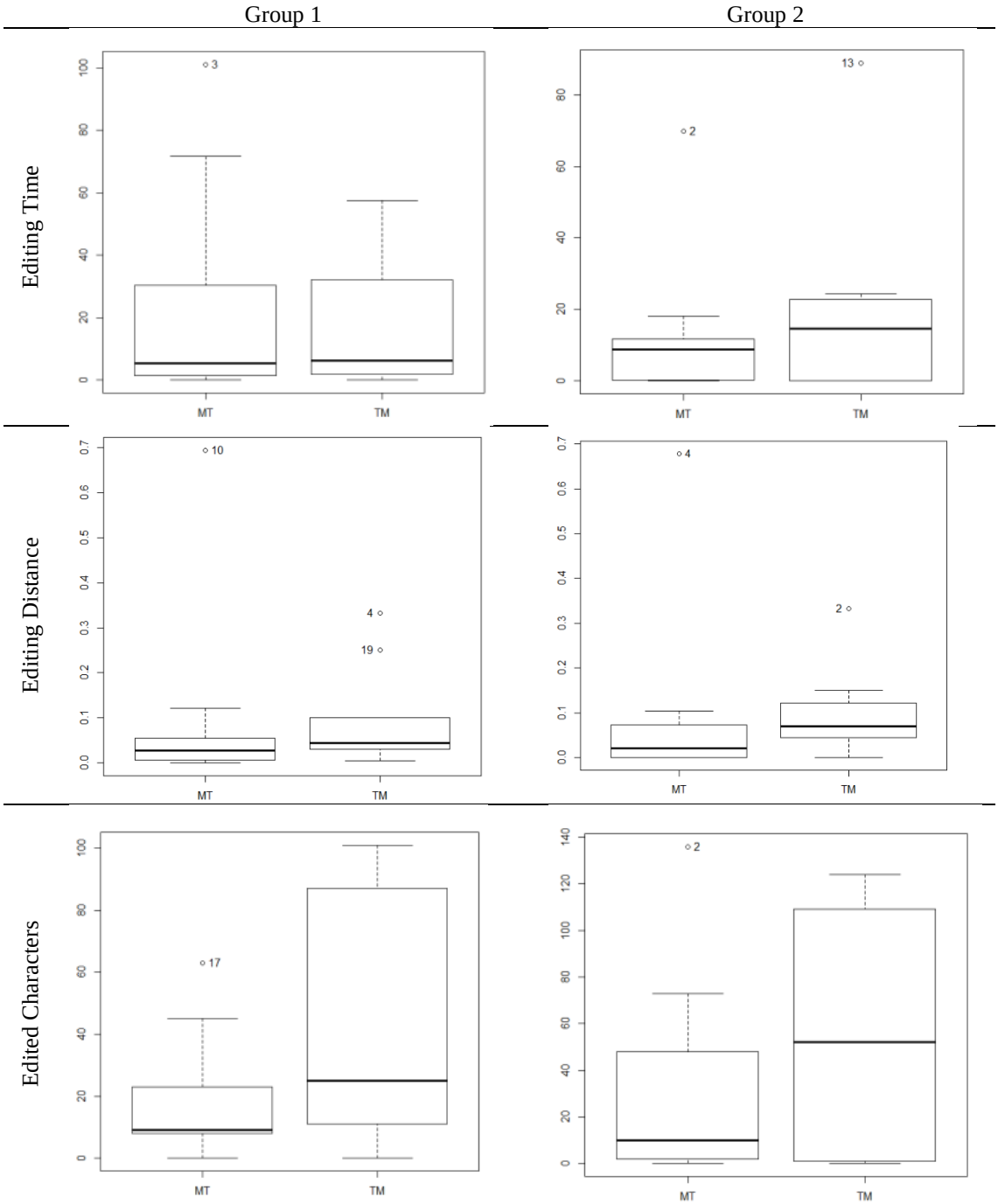


Figure 18. Performance on segments from 20 to 40 words

Figure 18 shows results for segments of over 20 words. In this case, those MT segments having a substantially higher number of words (49 and 54 each) were not considered. In this case, the results from edit distance and edited characters for Group 1 are clearly lower in the MT set than the results for Group 2. Group 2 MT results show a higher range of values, even though the results are always less dispersed than the TM results. Unlike edit distance and edited characters, editing time does not present such a clear disparity. In fact, Group 1 shows very similar results for both MT and TM in terms of time, while for Group 2, MT PE is somewhat faster and with less dispersion, mainly in the second half of the range.

5. Concluding remarks

In this paper, we have investigated professional translators' perceptions of NMT as a technology capable of boosting their productivity. We also sought to establish whether this perception can be confirmed by the data relating to their performance. The results have given us some insight, but not in an entirely conclusive manner.

Quality perception of proposed segments from MT and TM is quite similar. In general, both TM and MT show a similar amount of low-, medium-, and high-quality segments as assessed by translators. All TM proposals perceived as being of low quality by translators are fuzzy matches under 80%. Consequently, it could be assumed that post-editing MT and TM segments that are perceived as comparable in terms of quality should also show the same results in terms of performance. However, our findings do not allow us to draw that conclusion.

As regards performance, MT segments take slightly longer to edit on average than TM segments. This is particularly noticeable for segments from 10 to 19 words long, while the opposite happens for shorter (under 10 words) and longer (over 20 words) segments. MT segment length has previously been shown to be a surprisingly good parameter for PE effort estimation (Forcada et al. 2017). However, when attention is paid to the distribution, MT results are better in the shortest 50% (or even 75%) of segments, but for the remaining 25% the results include a substantially high number of outliers above Q4. In other words, even though 50% (or even 75%) of the MT segments are edited as fast as or even faster than TM segments, there is a 25% portion of the segments that take a great deal longer to be edited than TM segments.

While results for editing time do not really confirm nor refute whether MT boosts productivity, performance indicators related to the amount of edits incorporated into the proposed translations do. These results seem to be clearer when taking into account the level of TM similarity (in terms of fuzzy matches) as well as the number of words in the original segment. From this perspective, fuzzy matches may only be a relevant criterion with regard to productivity when the number of words in the segments compared is normalized (see Section 3. Methodology). Otherwise, a potential bias in favour of shorter segments should be taken into consideration.

Regarding segment length, MT performance results were compared with TM proposals with a variety of fuzzy match values. Here, MT performs better than TM at all levels of similarity (under 80%, from 80% to 90% and over 90%) in almost all quartiles of segment length for editing time, edit distance, and number of edited characters. Only in edit time does MT show higher and more dispersed results for the longest 50% of segments, which means that just a few MT segments perform much worse than TM segments, causing a reduction in the overall effectiveness of MT as shown in the results.

It is also worth mentioning that the results for edit distance and the number of edited characters clearly favour MT, but these results are not consistent with edit time values. Translators were not provided with information about the provenance of the proposed translations (whether they originated from MT or TM). This means that, when provided with information about the points where edits should be considered in TM-proposed translations (as CAT tools usually do), TM edit time results may be lower than those observed in this study. Research on word-level quality estimation for MT that may produce a similar visual aid for translators is ongoing but has not yet been widely deployed (Specia et al. 2018).

When translators' perceptions of the effect of on their productivity are taken into account, the study reveals that those translators who think that MT boosts their productivity (Group 1) actually perform a little faster when editing MT-proposed translations than those who think that MT slows them down (Group 2). This was also observed for edit distance and number of edited characters: in general, MT post-editing effort is lower than TM editing effort in Group 1, whereas this was not the case with Group 2. Bearing in mind the lack of translation proposal provenance information, this is an unexpected conclusion from the study that merits further investigation. In particular, it would be worth finding out whether these perceived and actual performance results are the same when translators are faced both with a blind and a non-blind task. Being aware of the origin of a translation proposal might have a direct impact on the translators' perception of their actual performance. In principle, the setting of our test allows us to consider whether these results are generalisable to other language combinations. However, further research should be carried out to dig deeper into different scenarios and settings, particularly using different text genres or focusing on different translator profiles.

Even though NMT allows a translator to achieve better quality output than other MT systems, the results obtained suggest that NMT output does not seem to boost productivity as much as might be expected. NMT output achieves good performance results in terms of edit distance and number of edits required, meaning that the output requires less editing effort than TM fuzzy matches in general. However, the time invested in post-editing NMT output is, in general, higher. A caveat here is that throughput may increase as post-editors become more accustomed to the types of errors that are produced by an NMT system. Furthermore, those translators who perceived that MT boosts their productivity actually performed better when post-editing MT segments than those translators who perceived MT as a poor resource. To delve deeper into the reasons why the correlation between an increase in quality and a decrease in editing and editing time does not occur, the results presented suggest that it might not be enough to merely collect data on performance and perception. The results point to the fact that translators' perceptions of MT tends to match their real productivity, even when they carry out a blind task and there are no hints as to the provenance of the segments they are editing. This suggests that this issue should also be addressed by gathering data of a cognitive nature on the process of post-editing MT output.

Acknowledgements:

This work has been supported by the ProjecTA-U project, grant number FFI2016-78612-R (MINECO / FEDER / UE), and by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2016) and is co-funded under the European Regional Development Fund.

References

Alabau V, Bonk R, Buck C, Carl M, Casacuberta F, García-Martínez M, González J, Koehn P, Leiva L, Mesa-Lao B, Ortiz D, Saint-Amand H, Sanchis G, Tsoukala C (2013) CASMACAT: an open source workbench for advanced computer aided translation. Prague Bull Math Linguist 100:101–112

Cadwell P, Castilho S, O'Brien S, Mitchell L (2016) Human factors in machine translation and post-editing among institutional translators. Translation Spaces 5(2):222-243

Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A (2017) Is neural machine translation the new state of the art? Prague Bull Math Linguist 108:109-120

Castilho S, Moorkens J, Gaspari F, Sennrich R, Way A, Georgakopoulou P (2018) Evaluating MT for massive open online courses. Mach Transl 32(3):255–278

Flournoy R, Duran C (2009) Machine translation and document localization at Adobe: From pilot to production. Proceedings of MT Summit XII, pp 425-428

Forcada M, Esplà-Gomis M, Sánchez-Martínez F, Specia L (2017) One-parameter models for sentence-level post-editing effort estimation. Proceedings of MT Summit XVI, vol.1: Research Track, pp 132-143

Klubička F, Toral A, Sánchez-Cartagena VM (2017) Fine-grained human evaluation of neural versus phrase-based machine translation. Prague Bull Math Linguist 108(1):121-132

Koehn P, Knowles R (2017) Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation, Vancouver, Canada, pp 28—39

Krings H P (2001). Repairing texts: Empirical investigations of machine translation post-editing process. The Kent State University Press, Kent

Läubli S, Germann U (2016) Statistical Modelling and Automatic Tagging of Human Translation Processes. In: Carl M, Bangalore S, Schaeffer M (eds) New Directions in Empirical Translation Process Research. Springer, Heidelberg, pp 77–94

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Doklady 10(8): 707–710

Lommel A, DePalma D (2016) Europe's Leading Role in Machine Translation. Common Sense Advisory – Cracker Project.

Martín-Mor A, Piqué Huerta R, Sánchez-Gijón P (2016) Tradumàtica: Tecnologies de la traducció. Eumo Editioral, Vic

Moorkens J, O'Brien S, Silva IAL, Fonseca N, Alves F (2015) Correlations of perceived post-editing effort with measurements of actual effort. Machine Translation 29(3-4):267-284

Moorkens J, Lewis D, Reijers W, Vanmassenhove E, Way A (2016) Translation Resources and Translator Disempowerment. Proceedings of ETHI-CA² 2016: ETHics In Corpus collection, Annotation and Application, pp 49-53.

Moorkens J, Way A (2016) Comparing Translator Acceptability of TM and SMT outputs. Baltic J. Modern Computing, 4(2):141-151

Moorkens J (2017) Under pressure: Translation in times of austerity. Perspectives, 25(3):464-477

Moorkens J (2018) Eye-Tracking as a Measure of Cognitive Effort for Post-Editing of Machine Translation. In: Walker C and Federici F (eds), Eye Tracking and Multidisciplinary Studies on Translation. John Benjamins, Amsterdam, pp 55-69

O'Brien S (2012) Translation as human-computer interaction. Translation Spaces 1:101–122

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, Philadelphia, Pennsylvania, pp 311–318

Pinnis M, Kalnis R, Skandis R, Skadina I (2016) What Can We Really Learn from Post-editing. Proc. of AMTA 2016 vol. 2: MT Users' Track, pp. 86-91.

Plitt M, Masselot F (2010) A productivity test of statistical machine translation post-editing in a typical localization context. Prague Bull Math Linguist 93:7–16

Rossetti, A, Gaspari, F (2017) Modelling the analysis of translation memory use and post-editing of raw machine translation output: A pilot study of trainee translators' perceptions of difficulty and time effectiveness. In: Hansen-Schirra, S, Czulo, O, Hoffmann, S (eds) Empirical modelling of translation and interpreting. Berlin: Language Science Press, pp. 41-67 http://langsci-press.org/catalog/view/132/360/902-1

Sánchez-Gijón P (2016) La posedición: hacia una definición competencial del perfil y una descripción multidimensional del fenómeno. Sendebar, 27:151-162

Sánchez-Torrón, M, Koehn, P (2016) Machine translation quality and post-editor productivity. Proceedings of AMTA 2016, pp.16

Shterionov D, Superbo R, Nagle P, Casanellas L, O'Dowd T, Way A (2018) Human versus automatic quality evaluation on NMT and PBSMT. Mach Transl 32(3): 217-235

Specia L, Blain F, Astudillo RF, Logacheva V, Martins A (2018) Findings of the WMT 2018 Shared Task on Quality Estimation. In: Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, Belgium, Brussels, pp 689–709

Teixeira CSC, Moorkens J, Turner D, Vreeke J, Way A (2019) Creating a multimodal translation tool and testing machine translation integration using touch and voice. In: Macken L, Daems J, Tezcan A (eds) Informatics 6(1) Special Issue on Advances in Computer-Aided Translation Technology.

Torres-Hostench O, Presas M, Cid P et al. (2016). El uso de traducción automática y posedición en las empresas de servicios lingüísticos españolas: informe de investigación ProjecTA 2015. https://ddd.uab.cat/record/148361

Way A (2018) Quality expectations of machine translation. In: Moorkens J, Castilho S, Gaspari F, Doherty S (eds) Translation Quality Assessment, Berlin: Springer, pp 159–178

Appendix

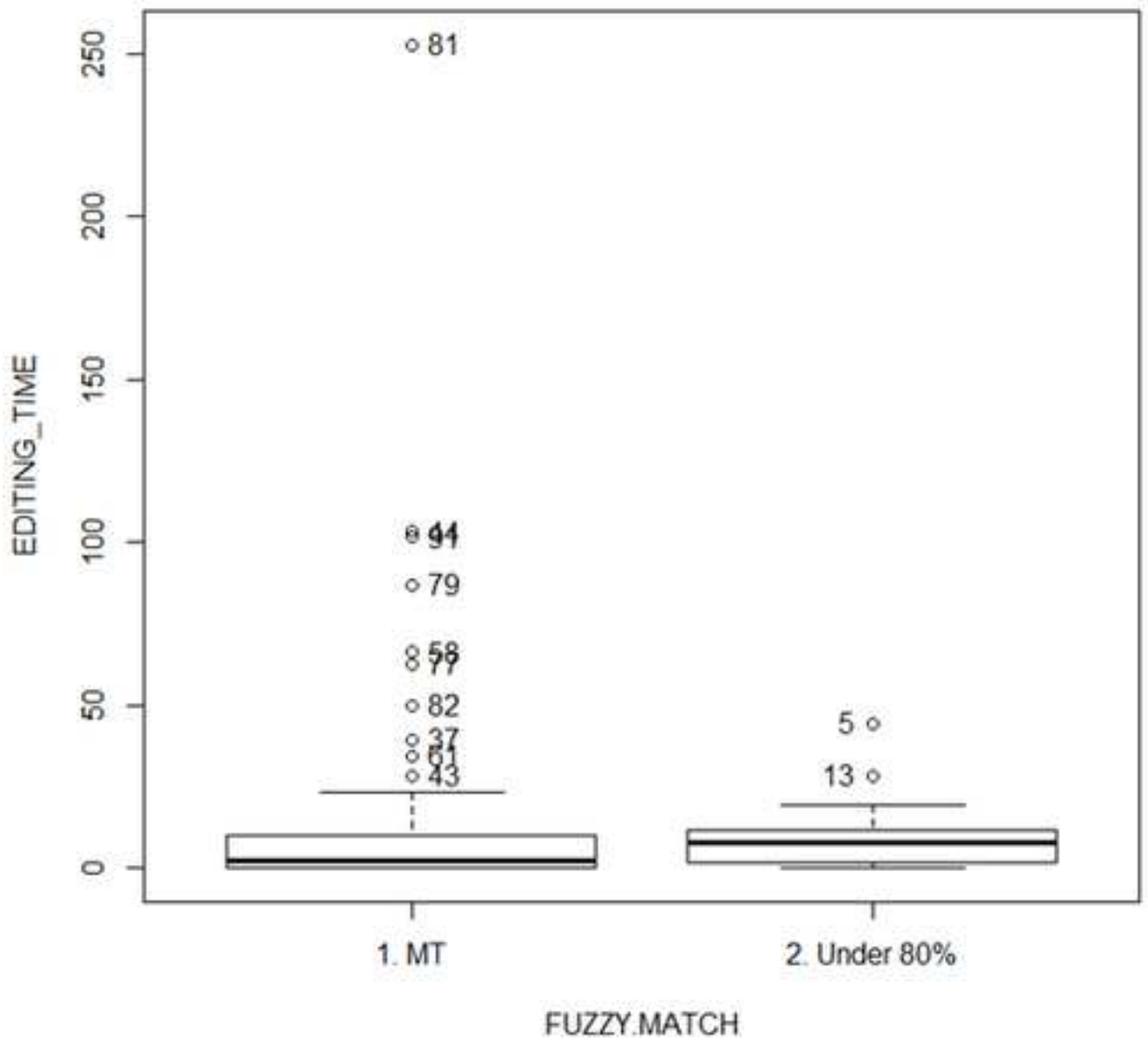
Segment distribution into two sets:

#	Origin	Number of words
1-1	MT	12
1-2	TM (fuzzy match under 80%)	5
1-3	MT	21
1-4	TM (fuzzy match over 90%)	24
1-5	MT	12
1-6	TM (fuzzy match from 80% to 89%)	25
1-7	MT	16
1-8	TM (fuzzy match from 80% to 89%)	23
1-9	MT	33
1-10	TM (fuzzy match from 80% to 89%)	21
1-11	MT	15
1-12	TM (fuzzy match from 80% to 89%)	35
1-13	MT	17
1-14	TM (fuzzy match from 80% to 89%)	10
1-15	MT	11
1-16	TM (fuzzy match from 80% to 89%)	16
1-17	MT	13
1-18	TM (fuzzy match from 80% to 89%)	16
1-19	MT	25
1-20	TM (fuzzy match from 80% to 89%)	17
1-21	MT	17
1-22	TM (fuzzy match over 90%)	24
1-23	MT	26
1-24	TM (fuzzy match from 80% to 89%)	16
1-25	MT	22
1-26	TM (fuzzy match from 80% to 89%)	20
1-27	MT	10
1-28	TM (fuzzy match from 80% to 89%)	7
1-29	MT	16
1-30	TM (fuzzy match from 80% to 89%)	10
2-1	TM (fuzzy match from 80% to 89%)	16
2-2	MT	25
2-3	TM (fuzzy match from 80% to 89%)	22

#	Origin	Number of words
2-4	MT	63
2-5	TM (fuzzy match under 80%)	7
2-6	MT	15
2-7	TM (fuzzy match over 90%)	29
2-8	MT	19
2-9	TM (fuzzy match from 80% to 89%)	14
2-10	MT	17
2-11	TM (fuzzy match under 80%)	21
2-12	MT	17
2-13	TM (fuzzy match under 80%)	22
2-14	MT	16
2-15	TM (fuzzy match under 80%)	17
2-16	MT	68
2-17	TM (fuzzy match from 80% to 89%)	23
2-18	MT	12
2-19	TM (fuzzy match under 80%)	18
2-20	MT	11
2-21	MT	12
2-22	TM (fuzzy match under 80%)	5
2-23	MT	24
2-24	TM (fuzzy match over 90%)	18
2-25	MT	19
2-26	TM (fuzzy match over 90%)	27
2-27	MT	26
2-28	TM (fuzzy match over 90%)	16
2-29	MT	6

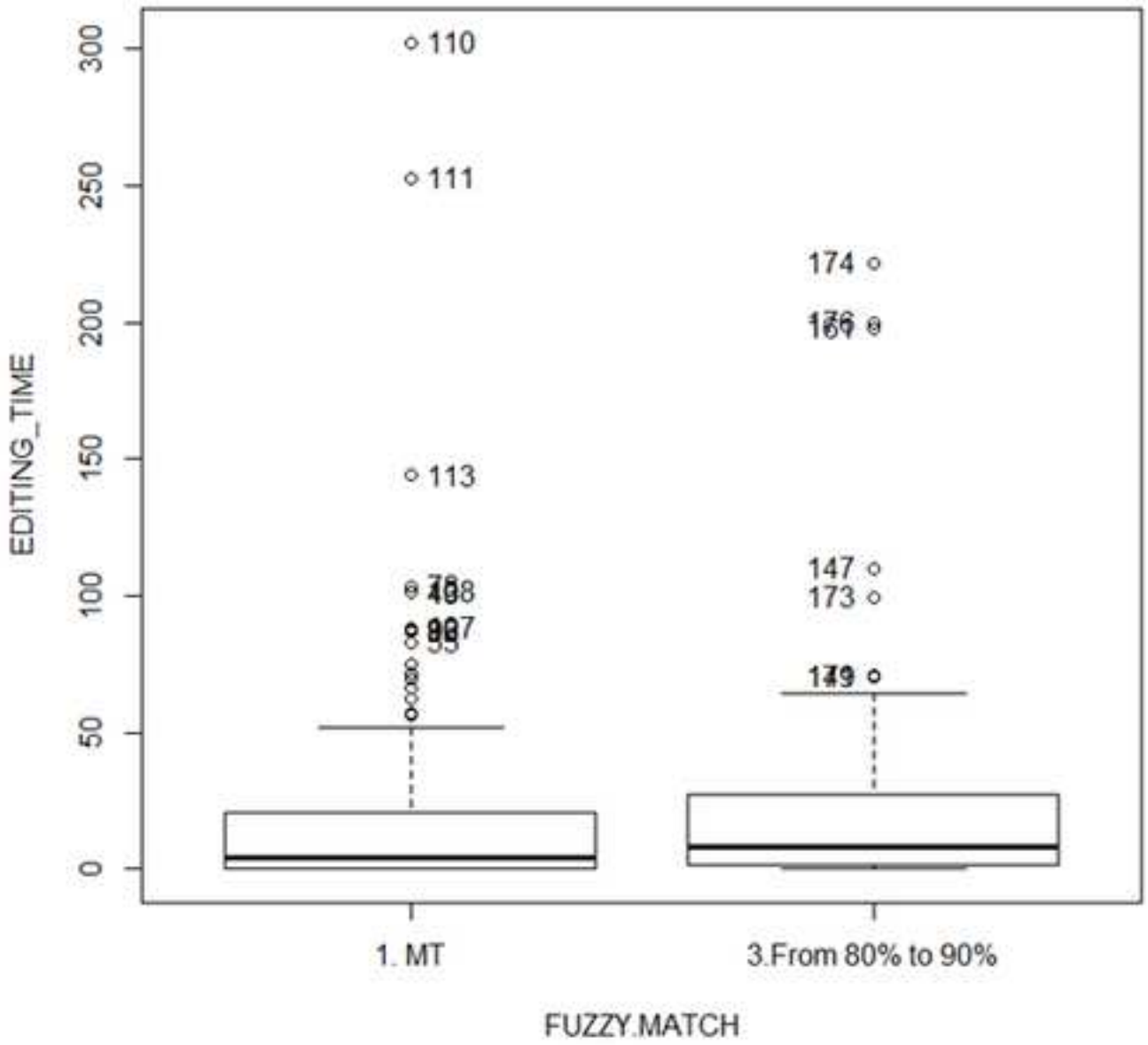
Figure

Click here to access/download;Figure;Figure_05-a.bmp



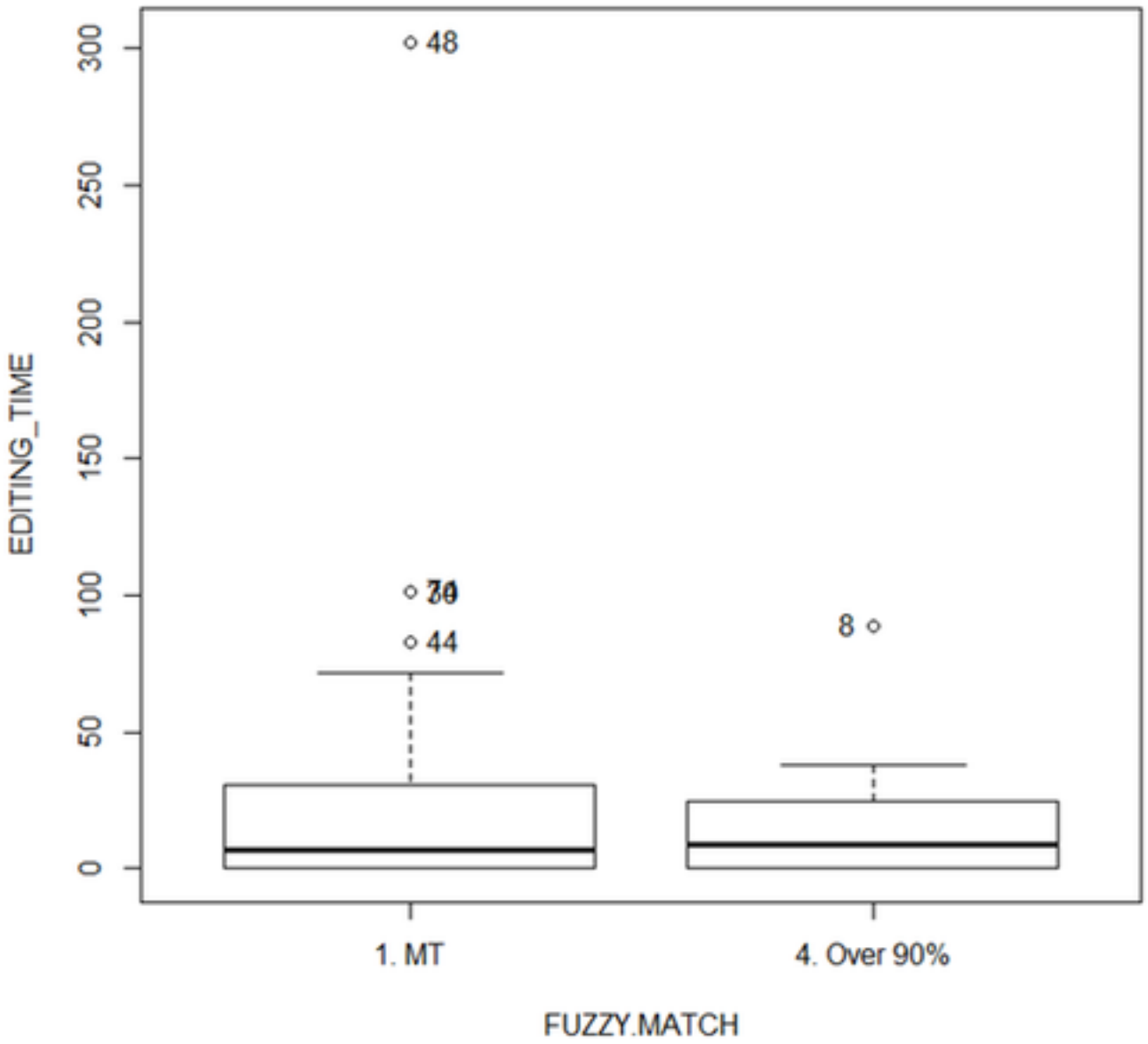
Figure

Click here to access/download;Figure;Figure_05-b.bmp



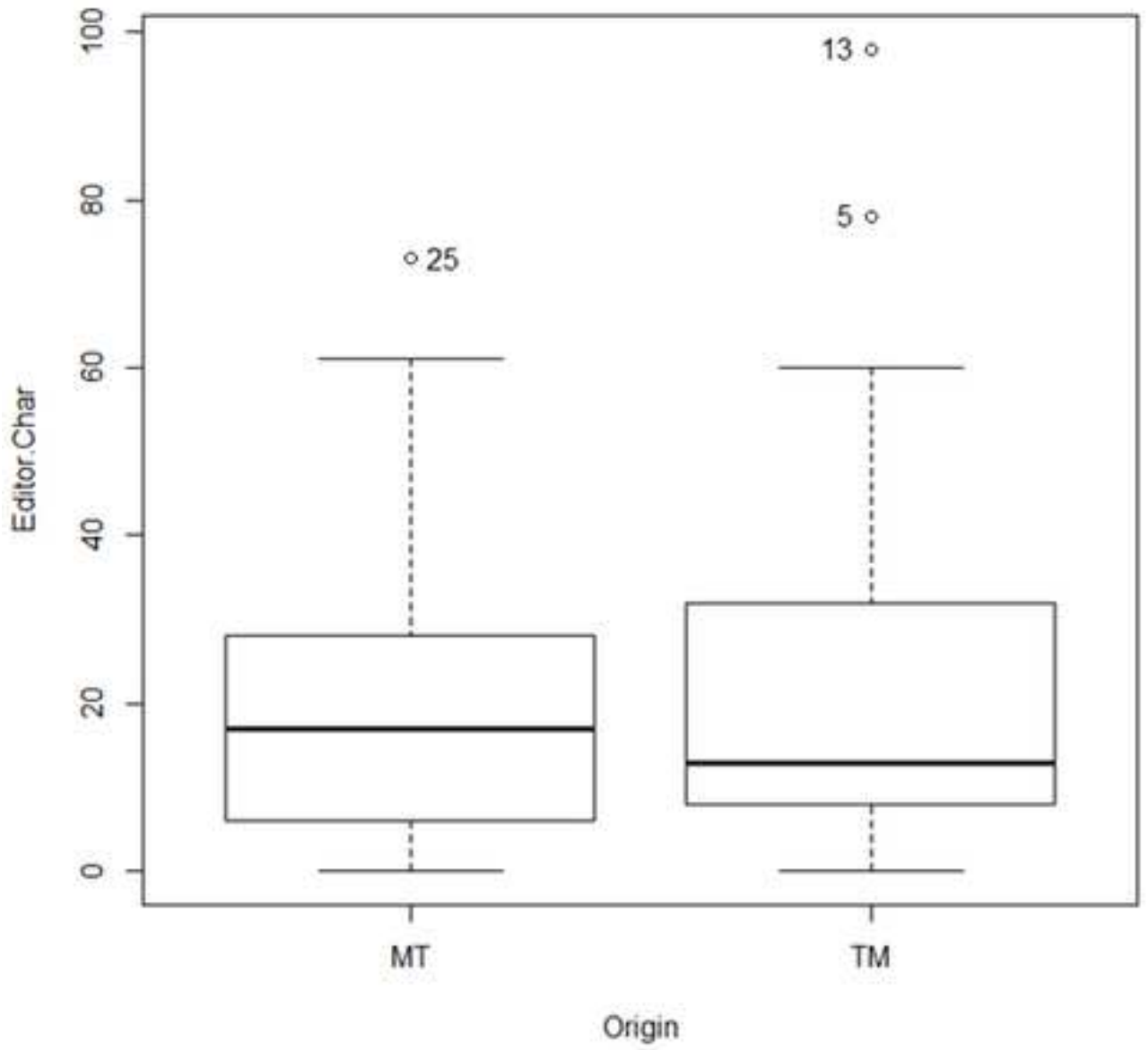
Figure

Click here to access/download;Figure;Figure_05-c.bmp



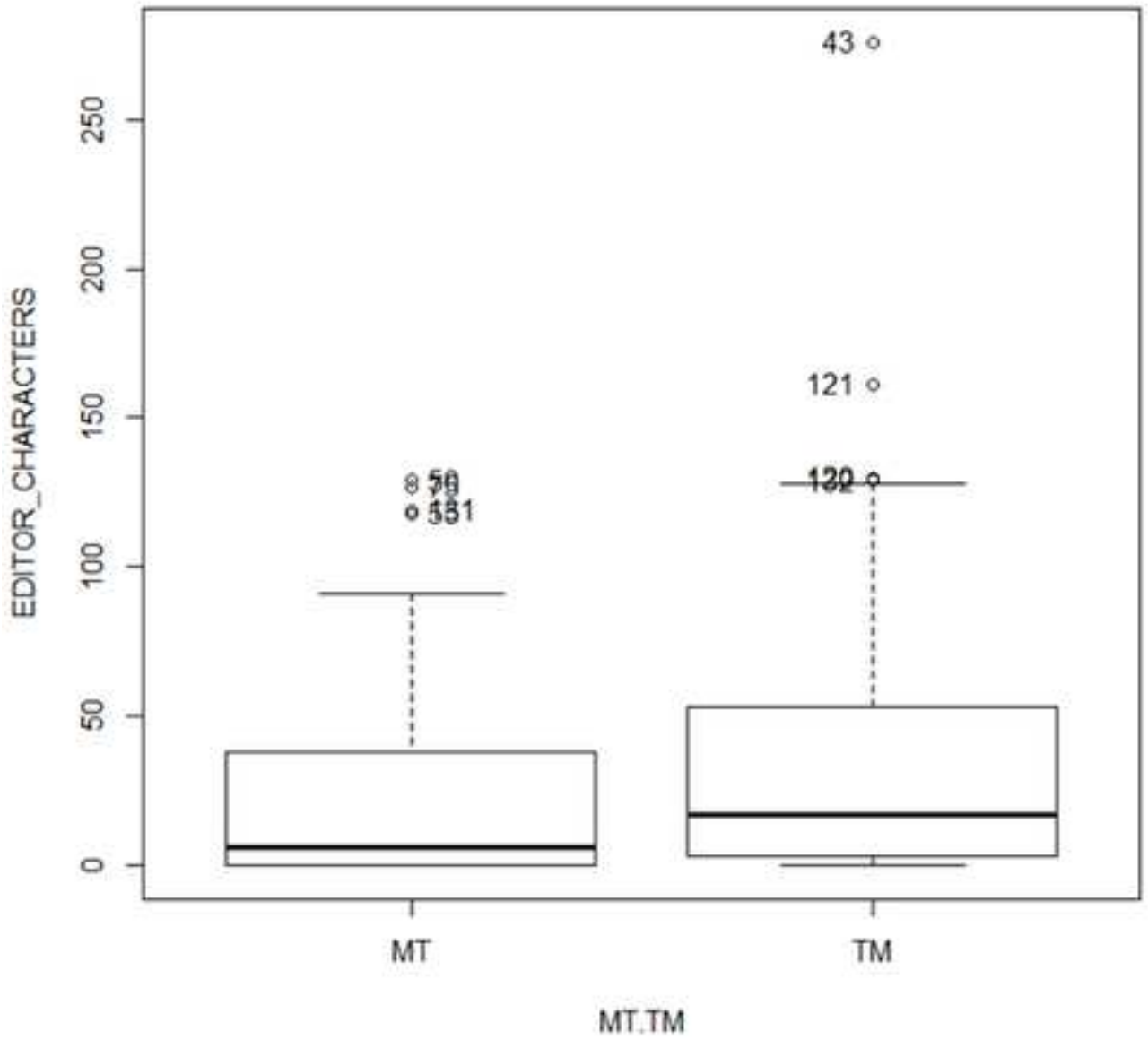
Figure

Click here to access/download;Figure;Figure_11-a.bmp



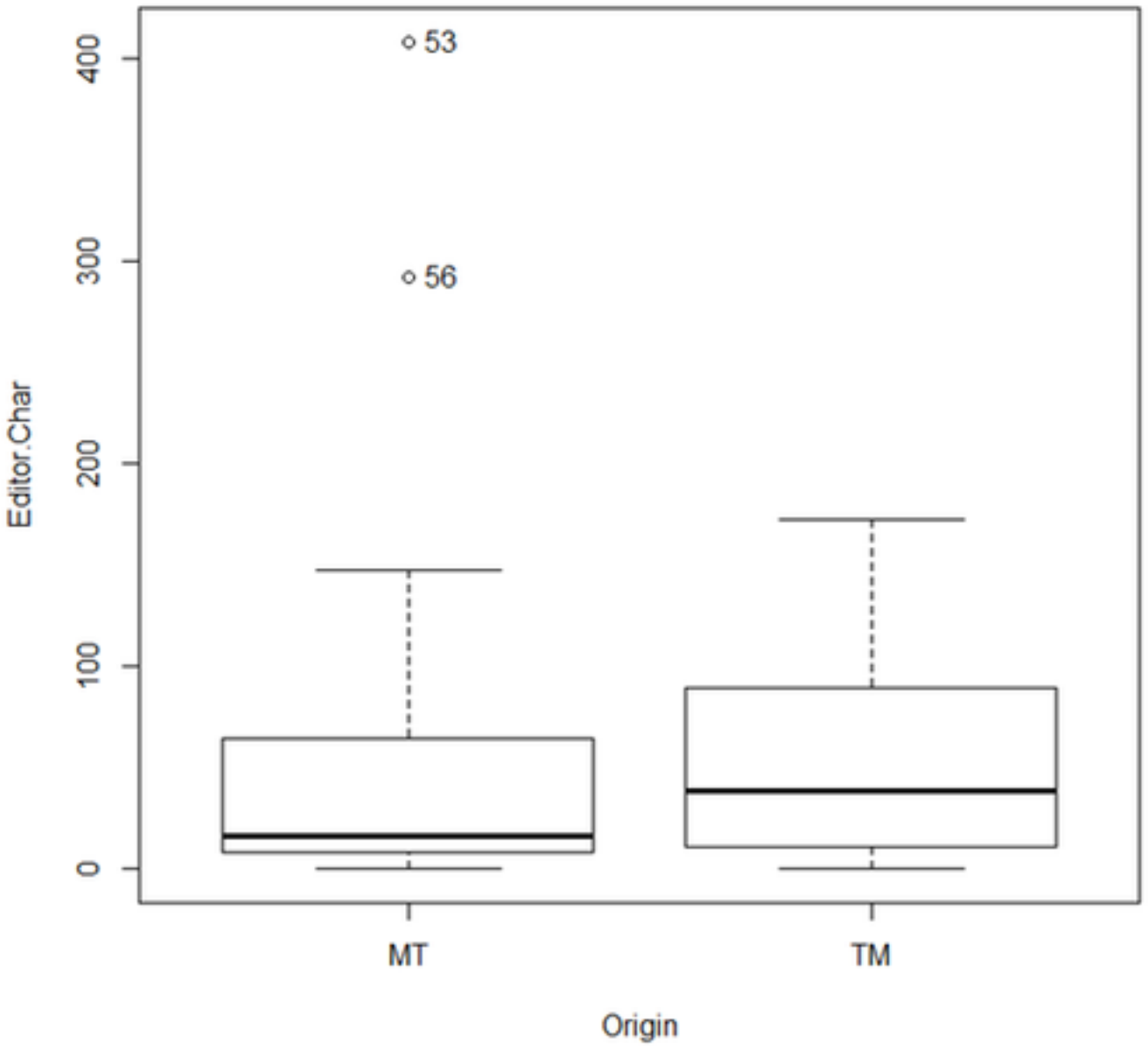
Figure

Click here to access/download;Figure;Figure_11-b.bmp



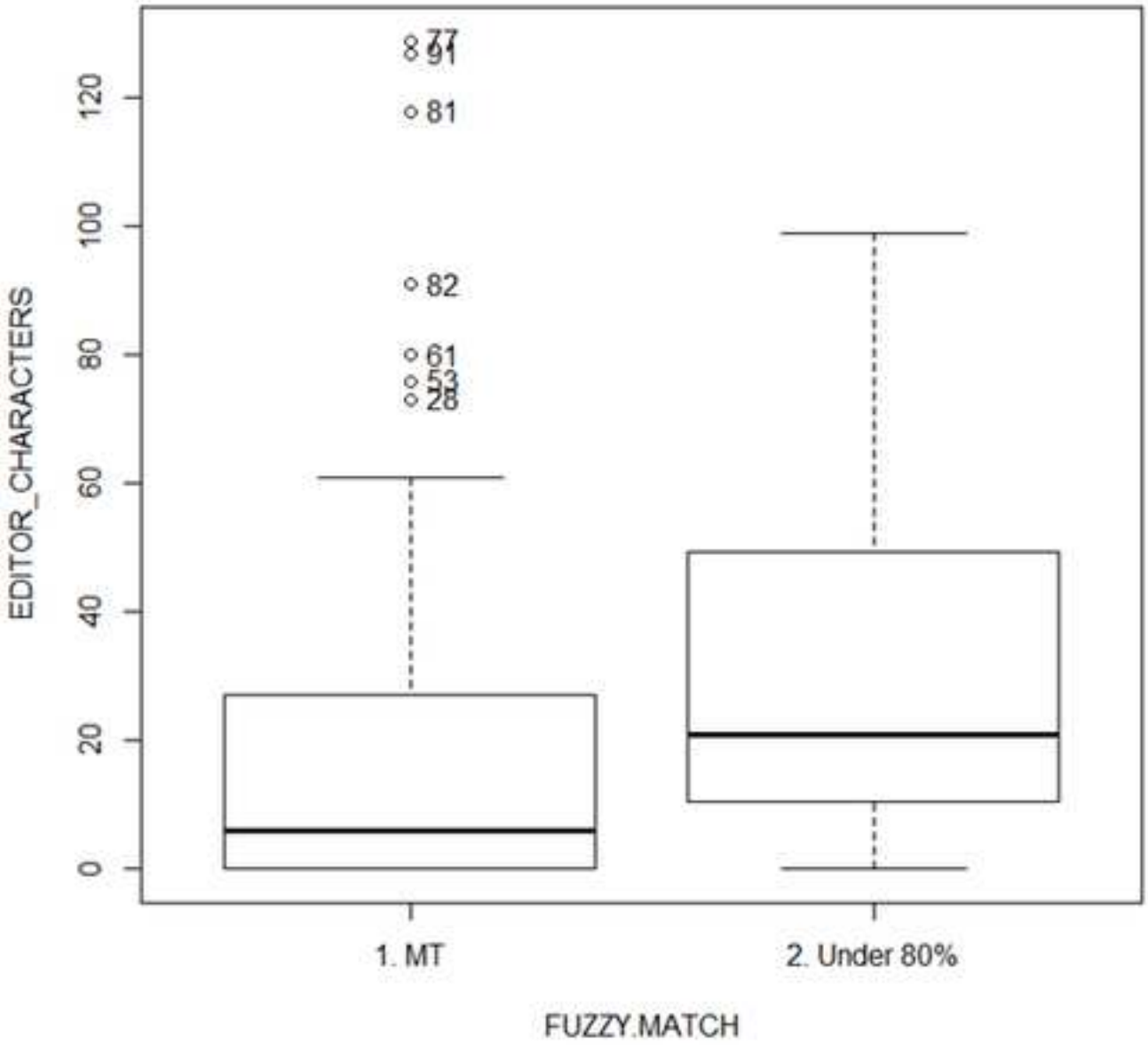
Figure

Click here to access/download;Figure;Figure_11-c.bmp

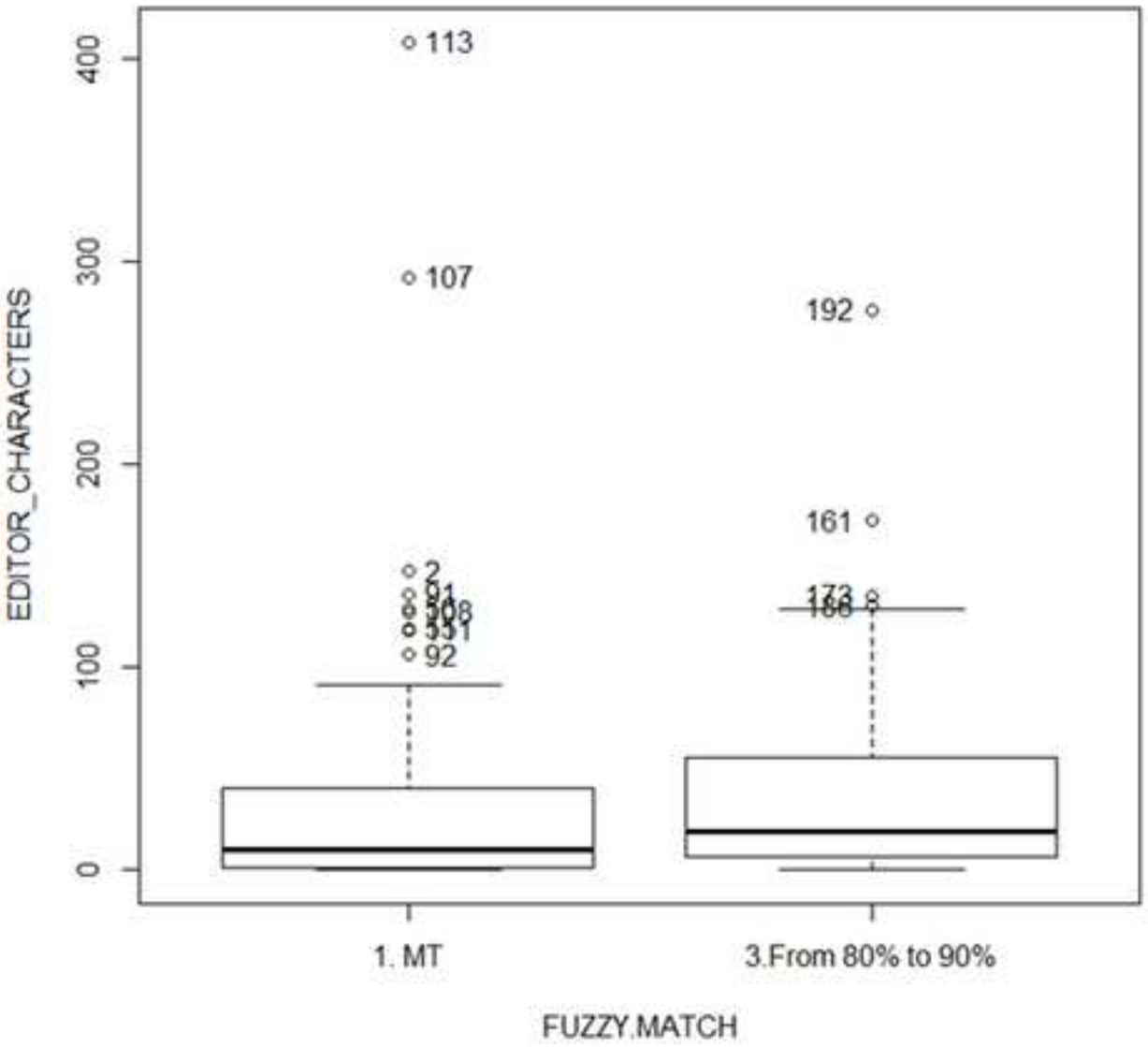


Figure

Click here to access/download;Figure;Figure_12-a.bmp

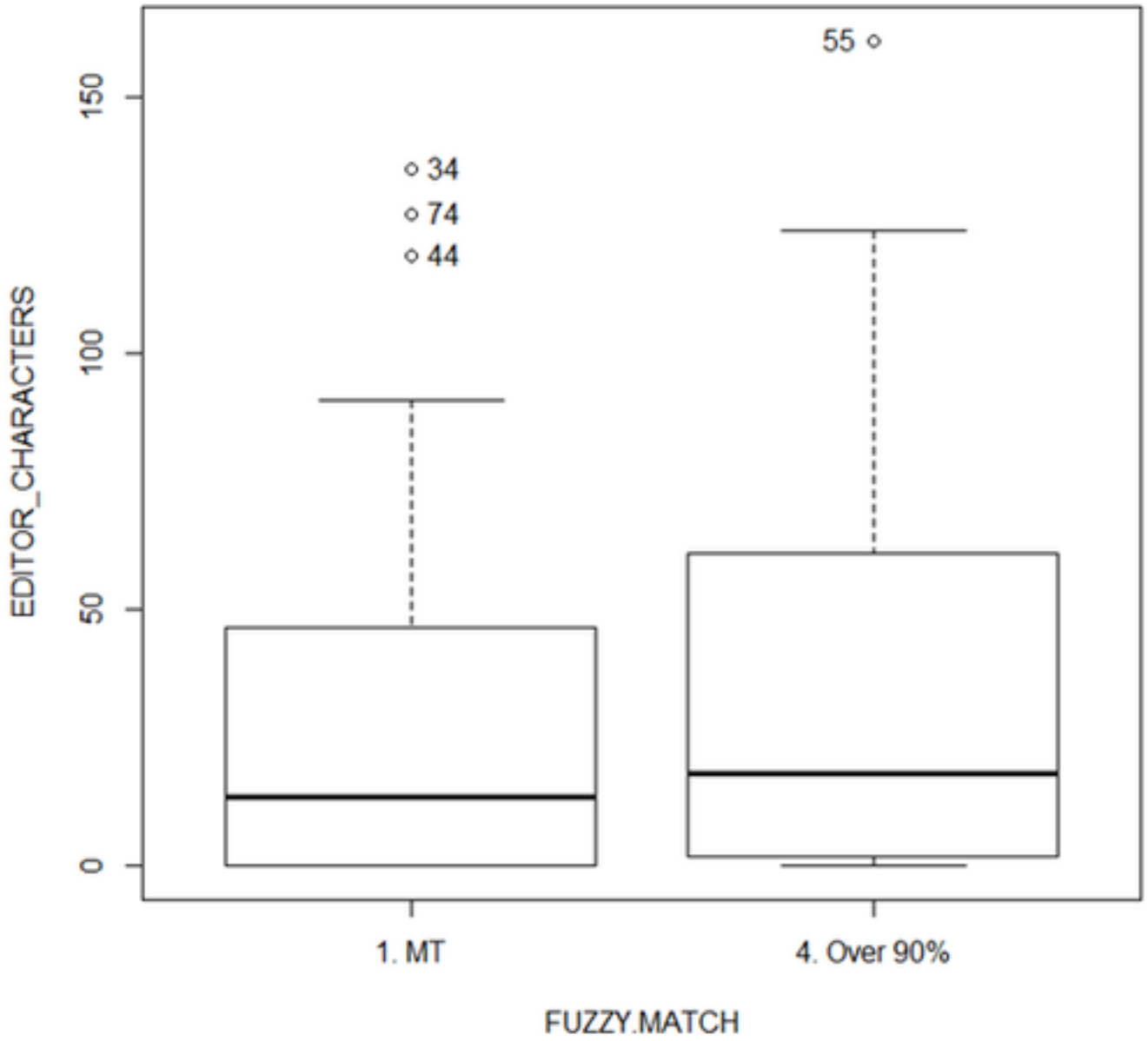


Figure



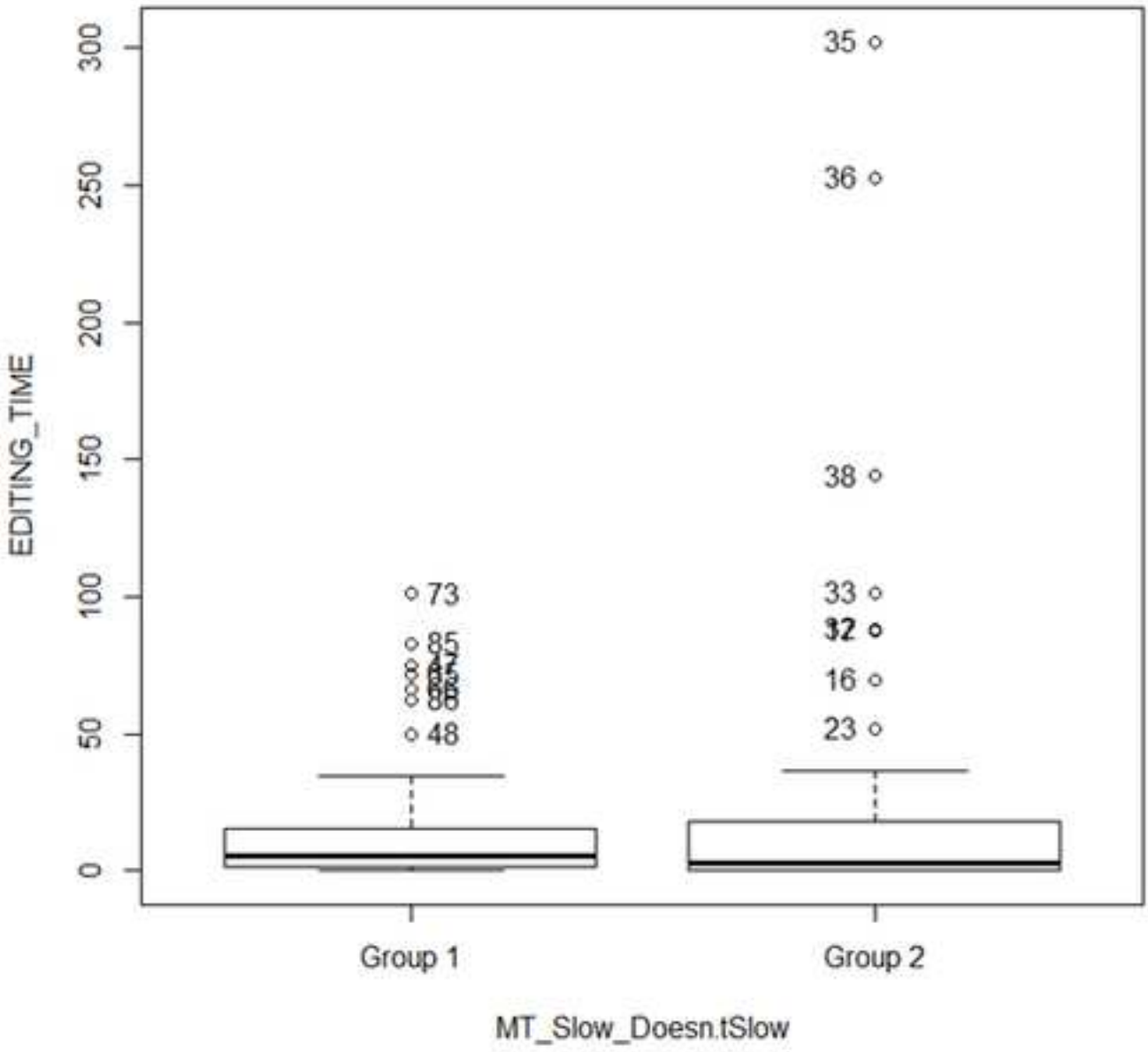
Figure

Click here to access/download;Figure;Figure_12-c.bmp



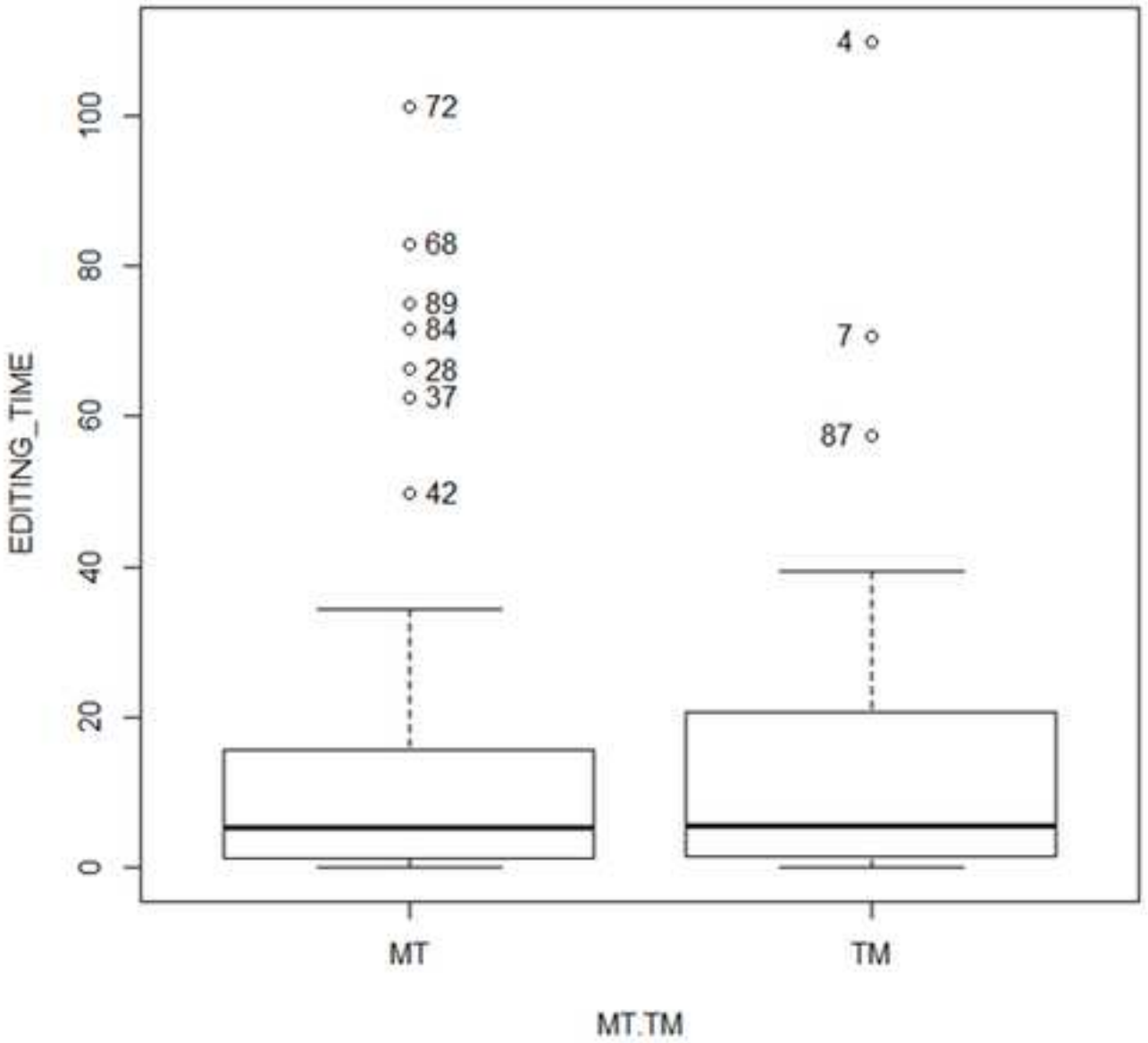
Figure

Click here to access/download;Figure;Figure_13.bmp



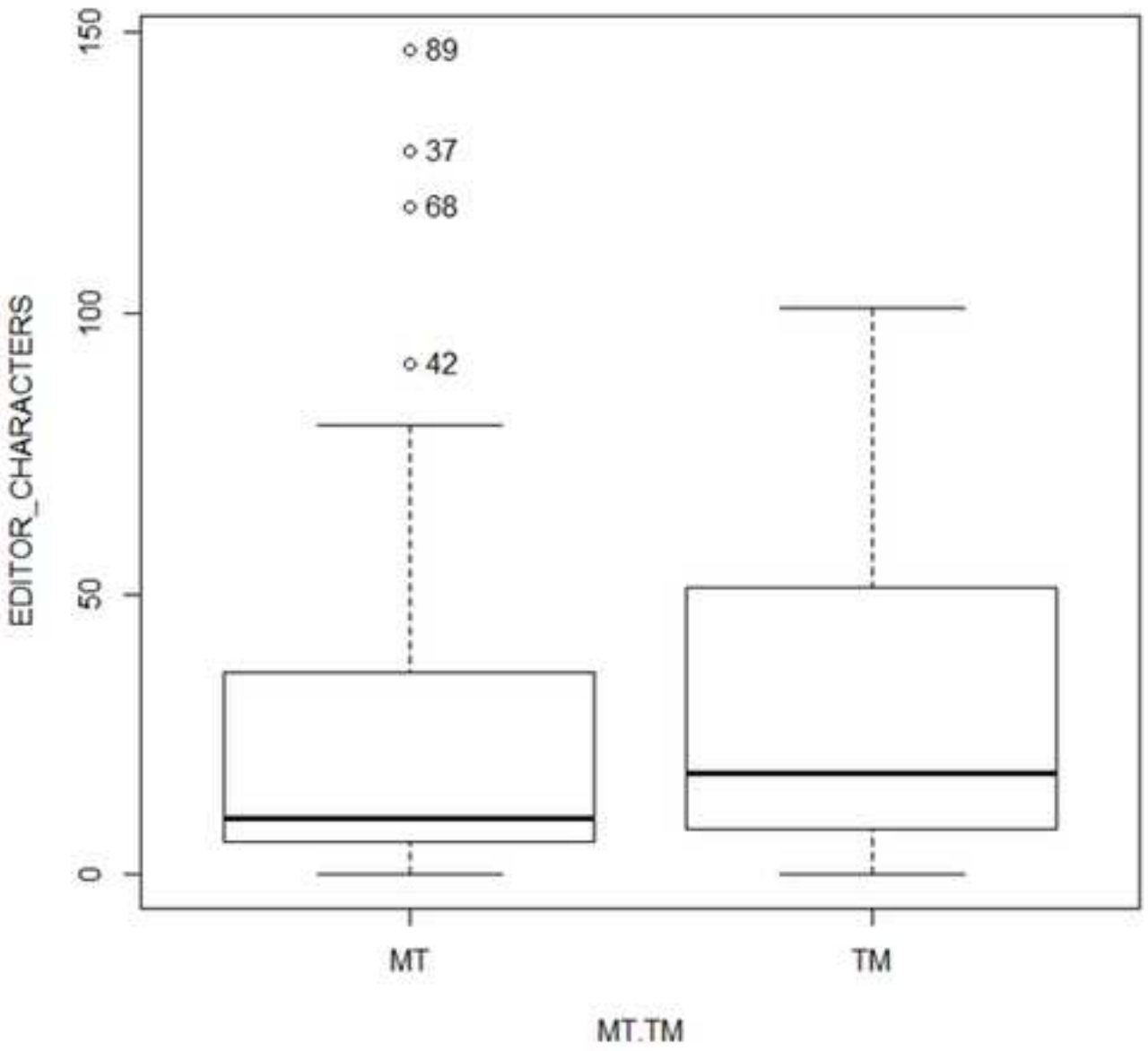
Figure

Click here to access/download;Figure;Figure_14-a.bmp



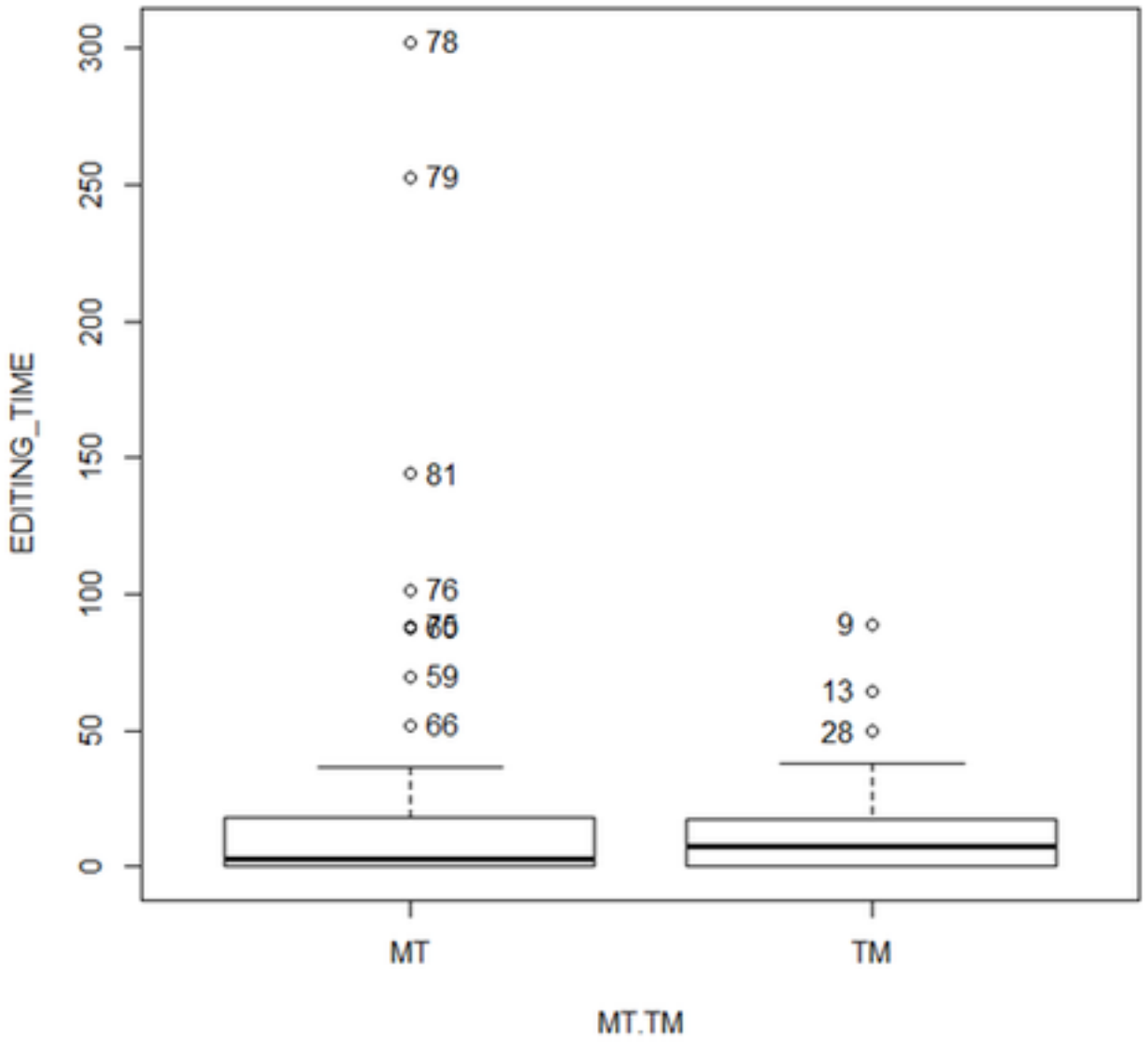
Figure

Click here to access/download;Figure;Figure_14-c.bmp



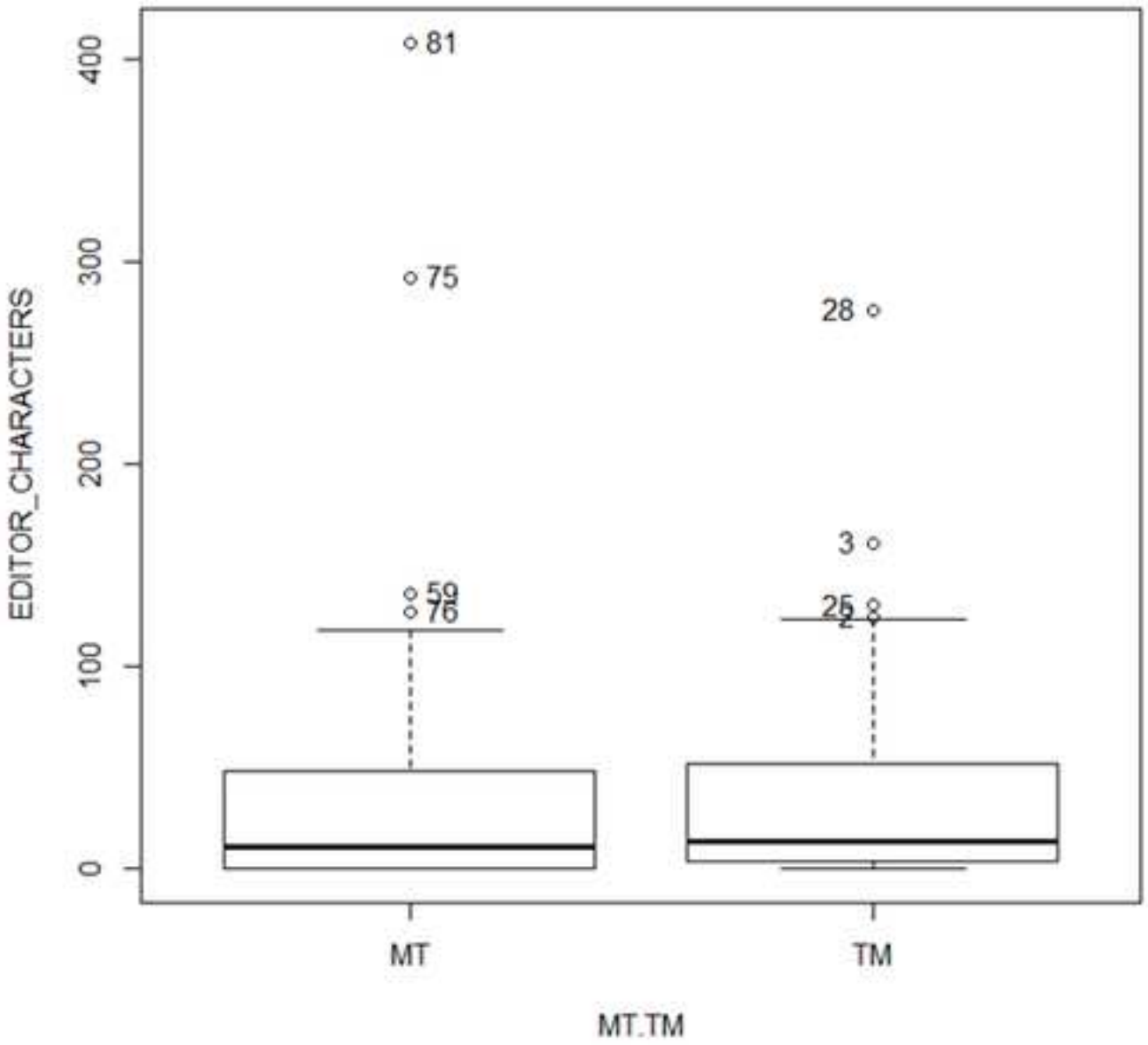
Figure

Click here to access/download;Figure;Figure_15-a.bmp



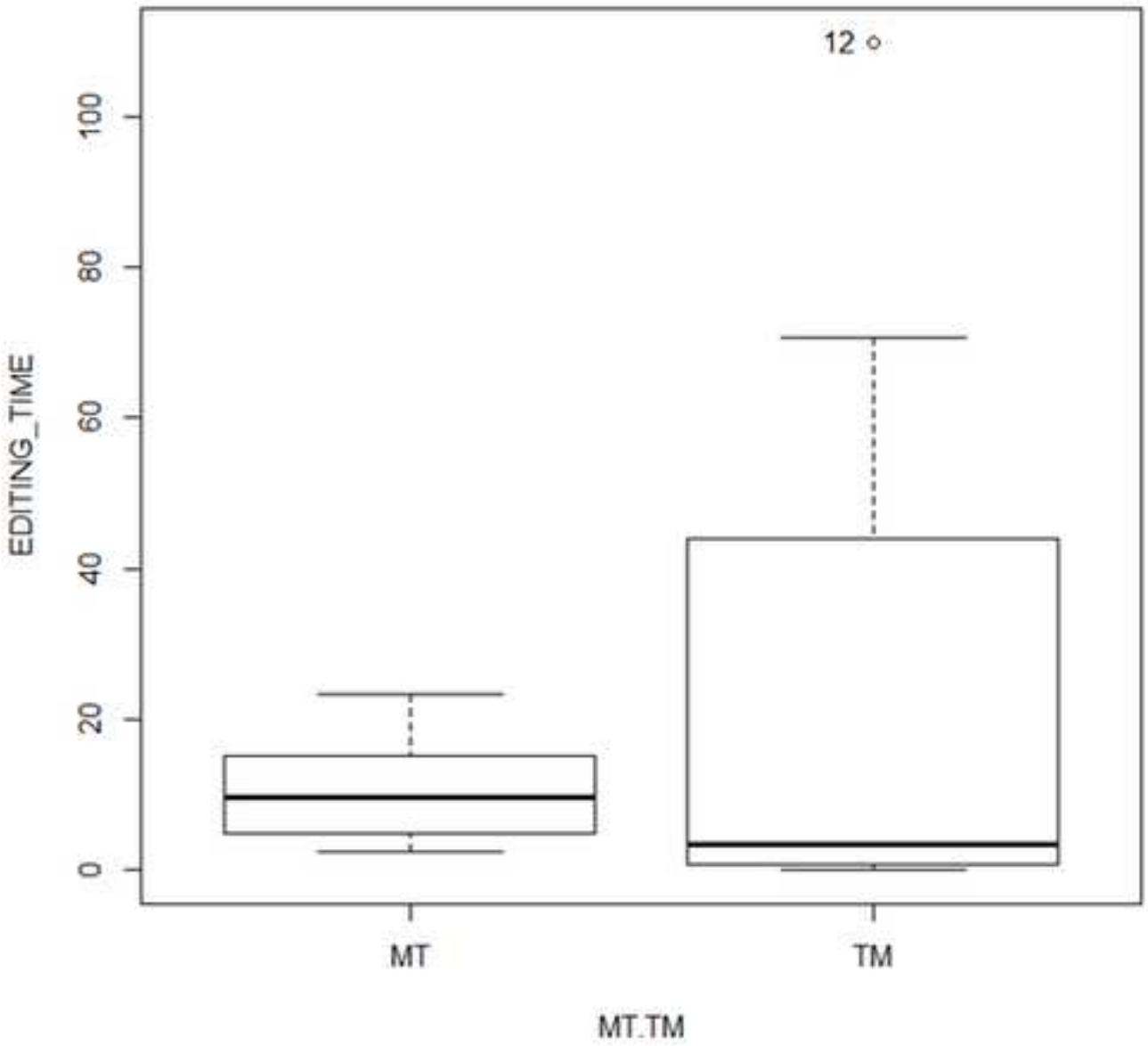
Figure

Click here to access/download;Figure;Figure_15-c.bmp



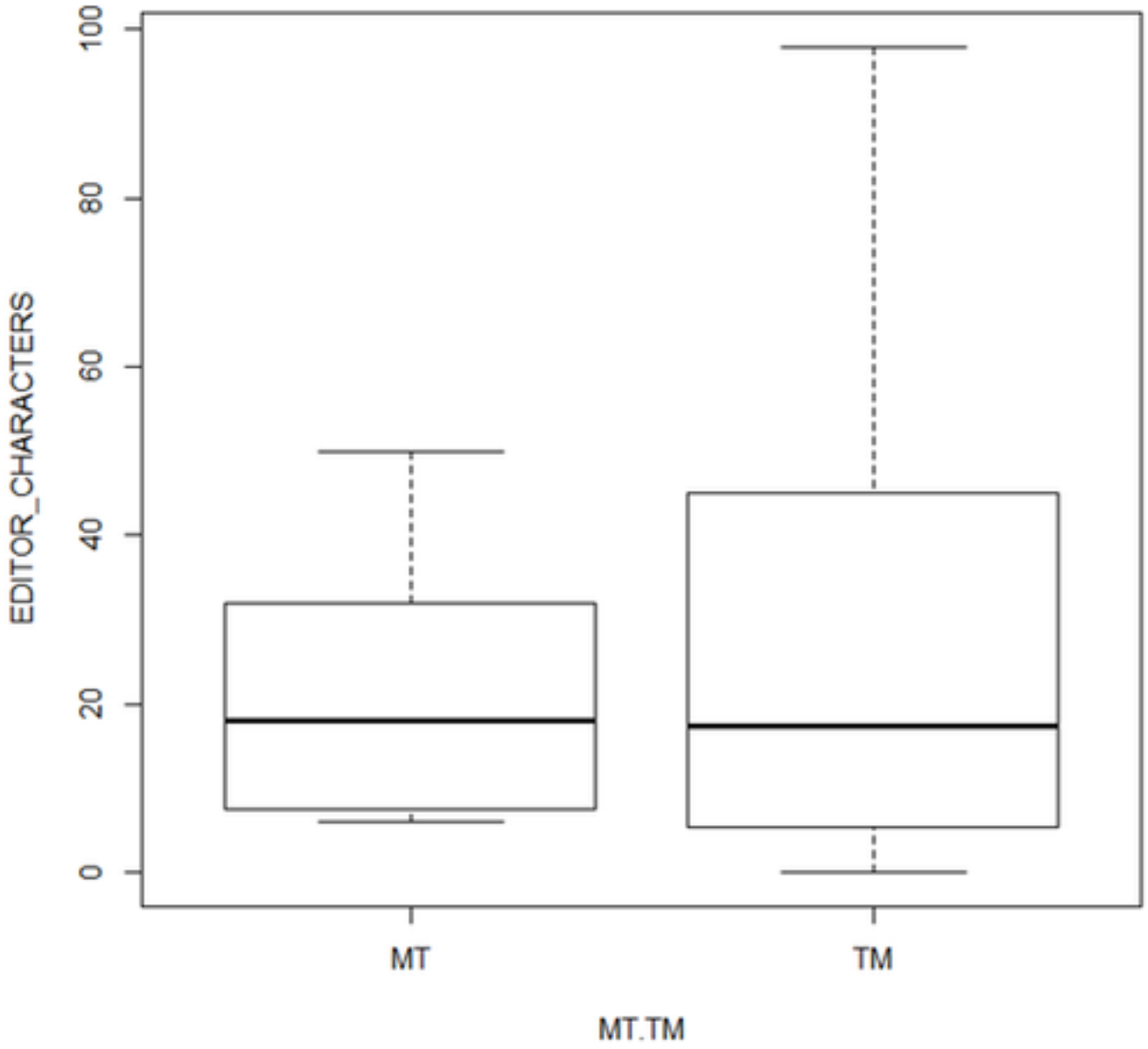
Figure

Click here to access/download;Figure;Figure_16-a.bmp



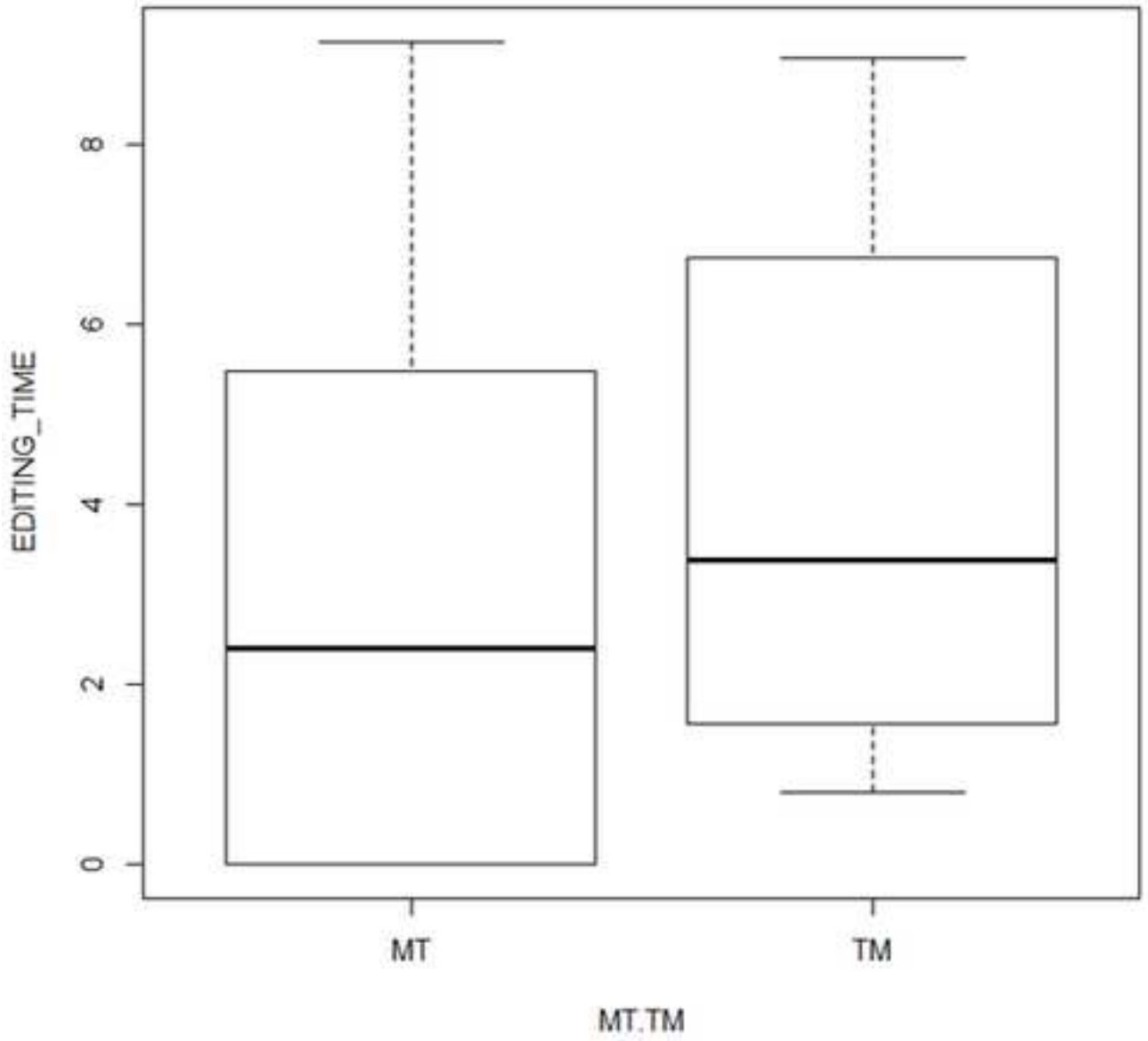
Figure

Click here to access/download;Figure;Figure_16-c.bmp



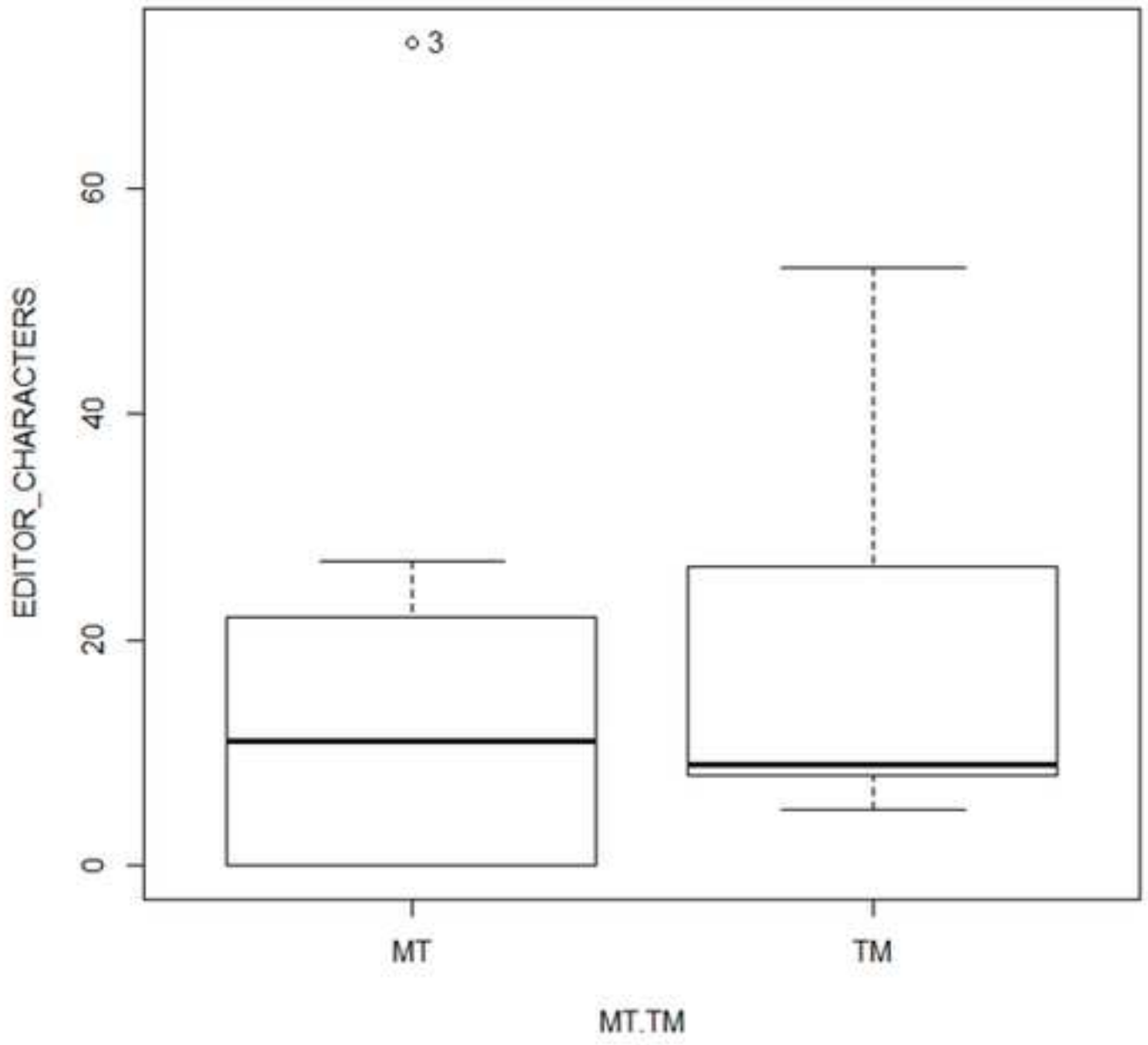
Figure

Click here to access/download;Figure;Figure_16-d.bmp



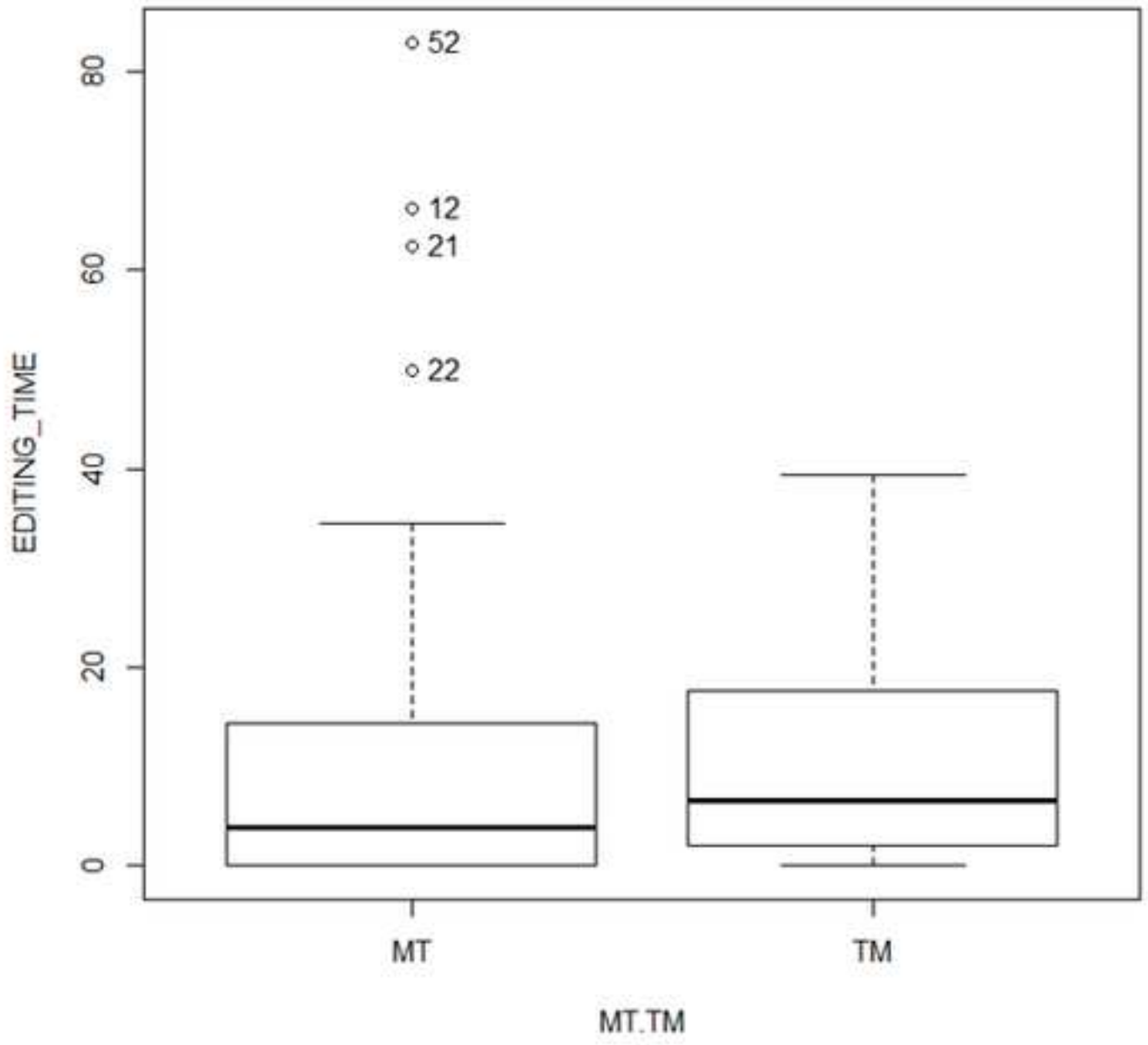
Figure

Click here to access/download;Figure;Figure_16-f.bmp



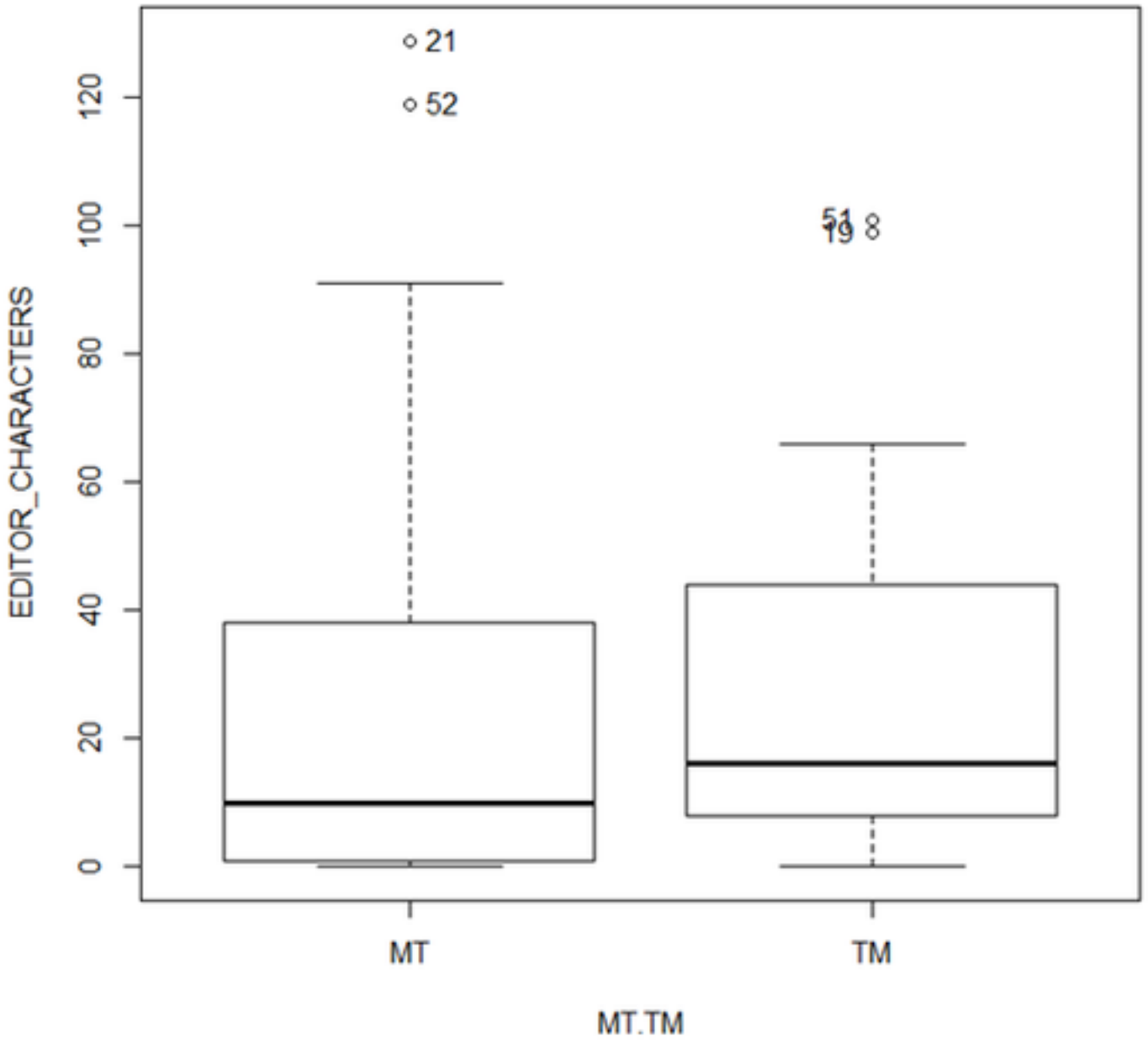
Figure

Click here to access/download;Figure;Figure_17-a.bmp



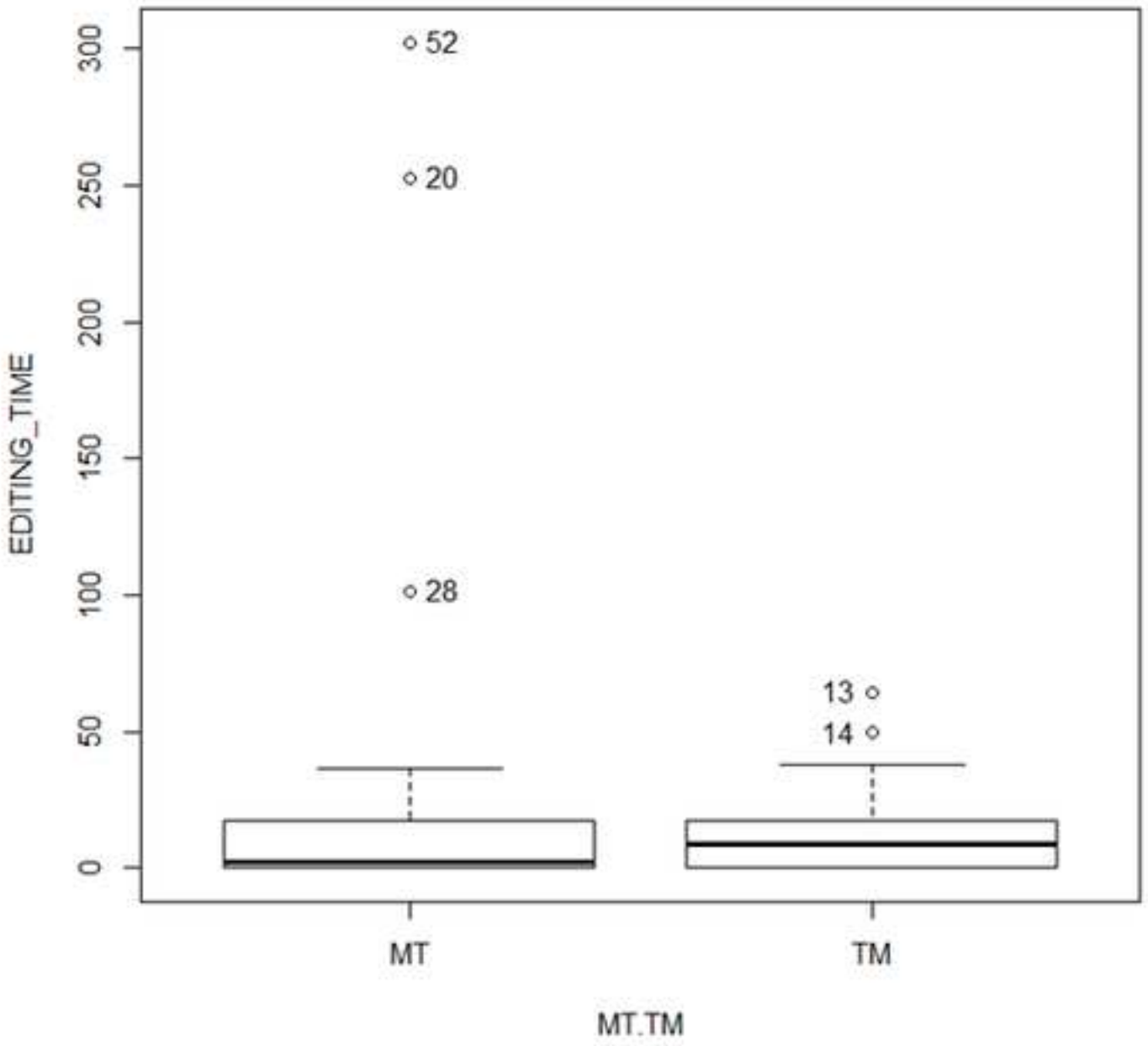
Figure

Click here to access/download;Figure;Figure_17-c.bmp

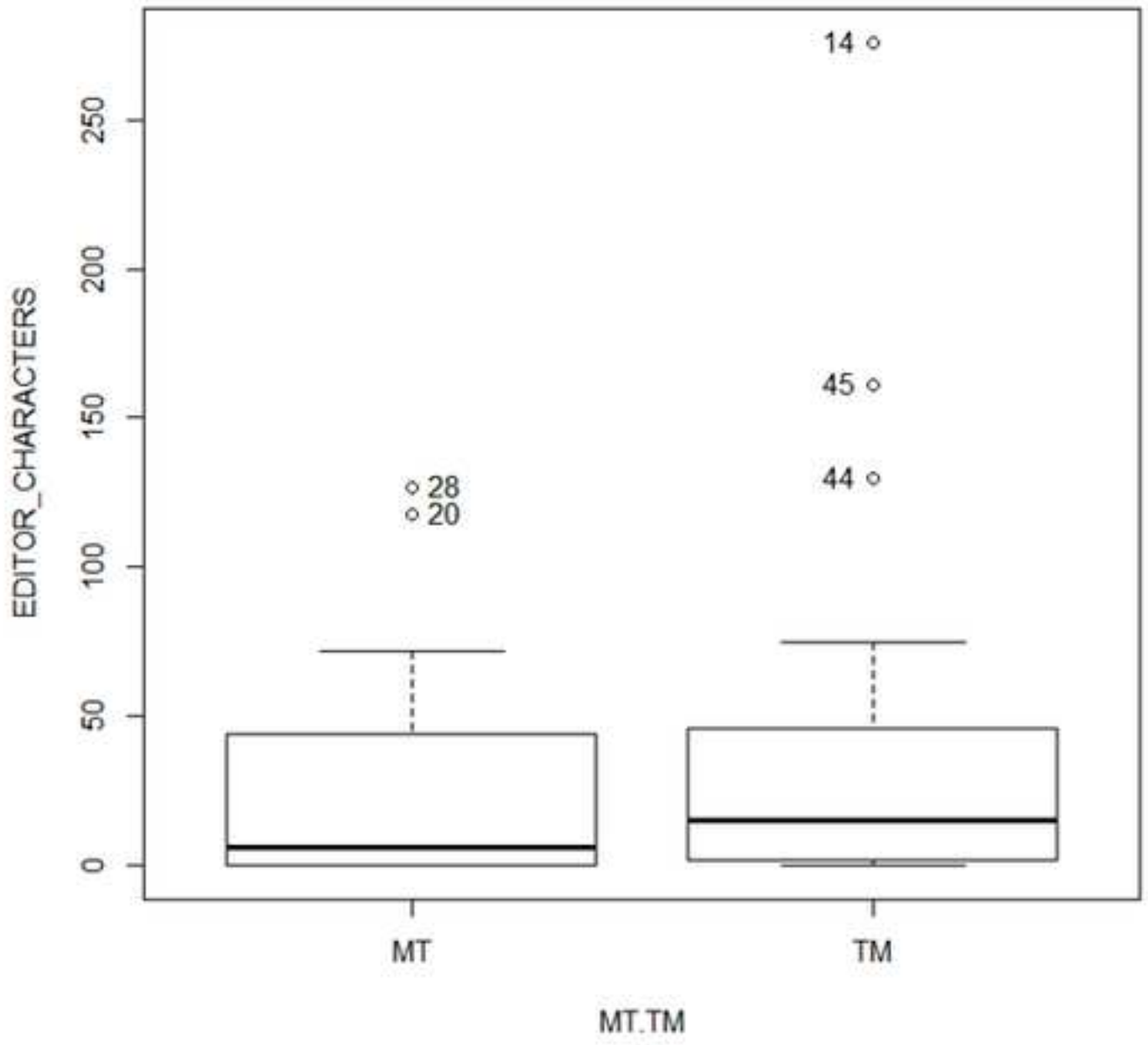


Figure

Click here to access/download;Figure;Figure_17-d.bmp

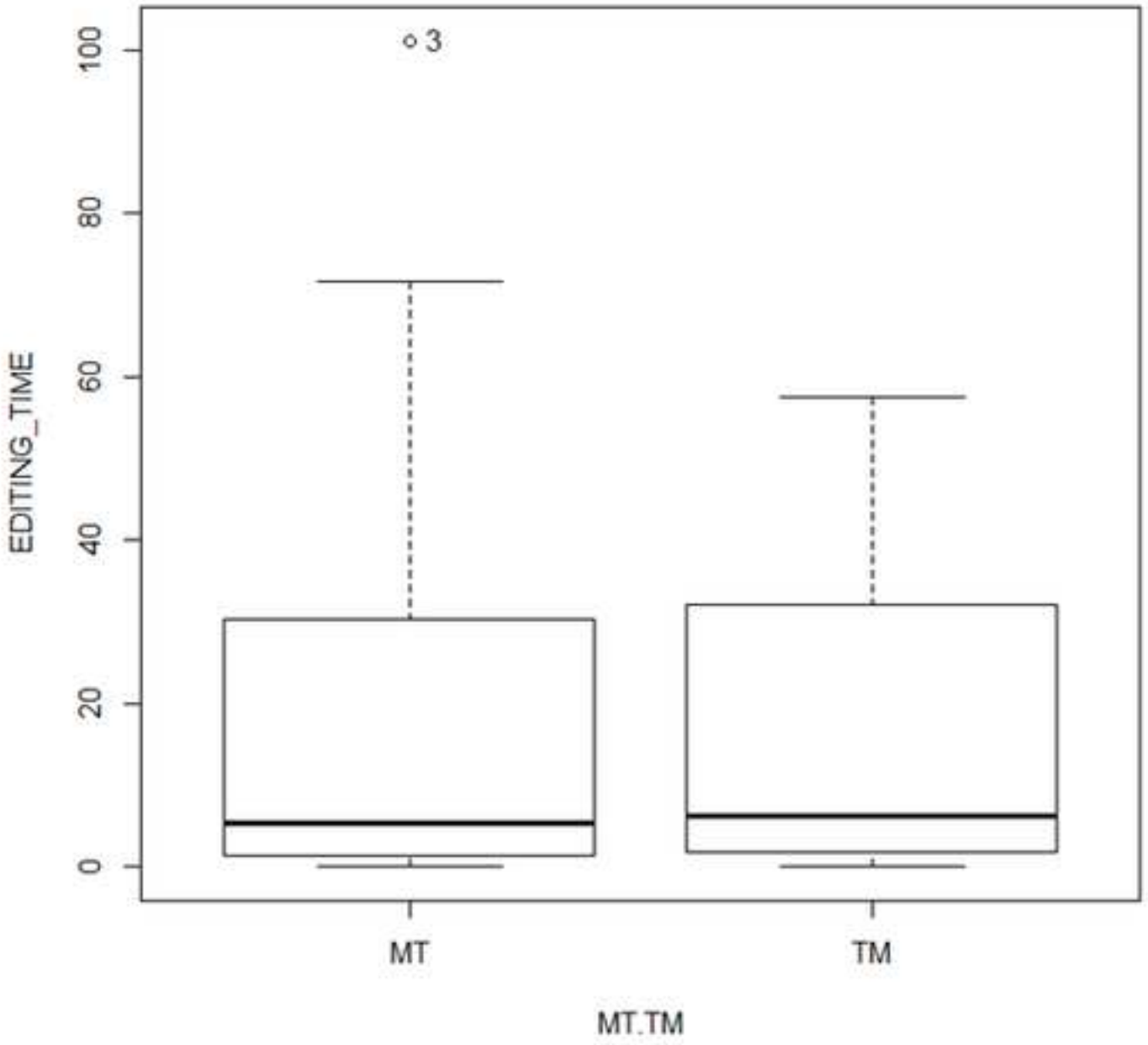


Figure



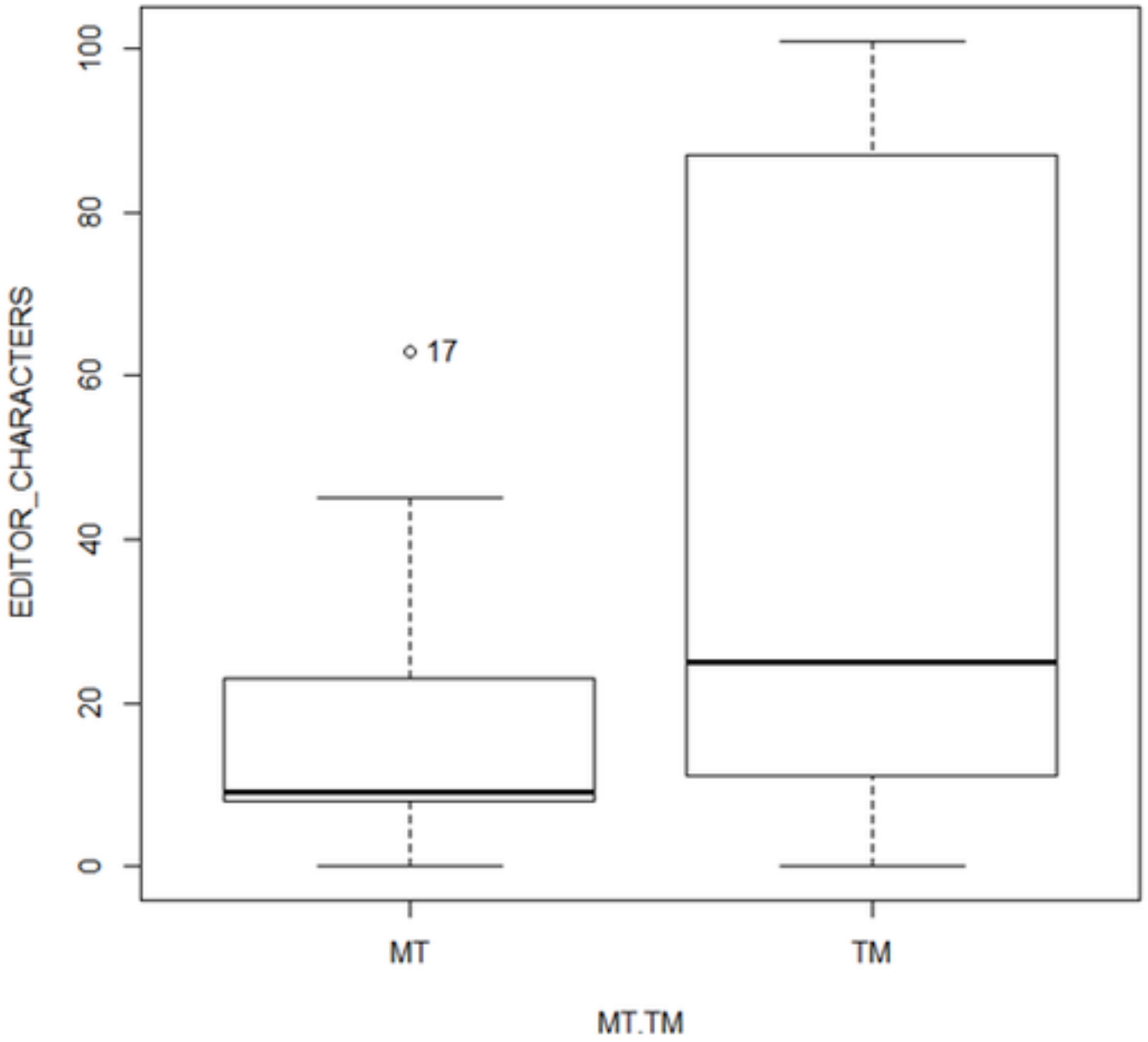
Figure

Click here to access/download;Figure;Figure_18-a.bmp



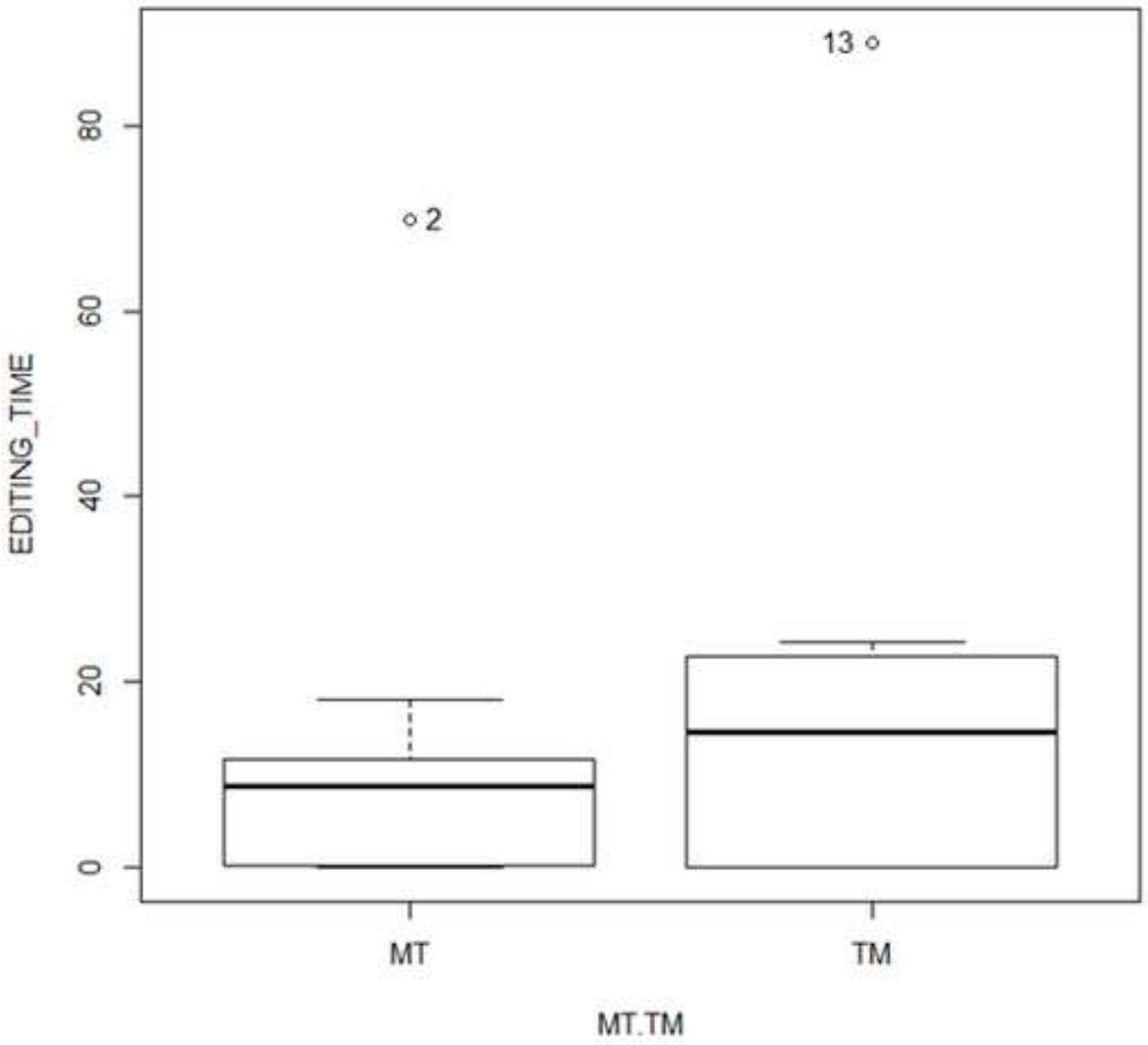
Figure

Click here to access/download;Figure;Figure_18-c.bmp



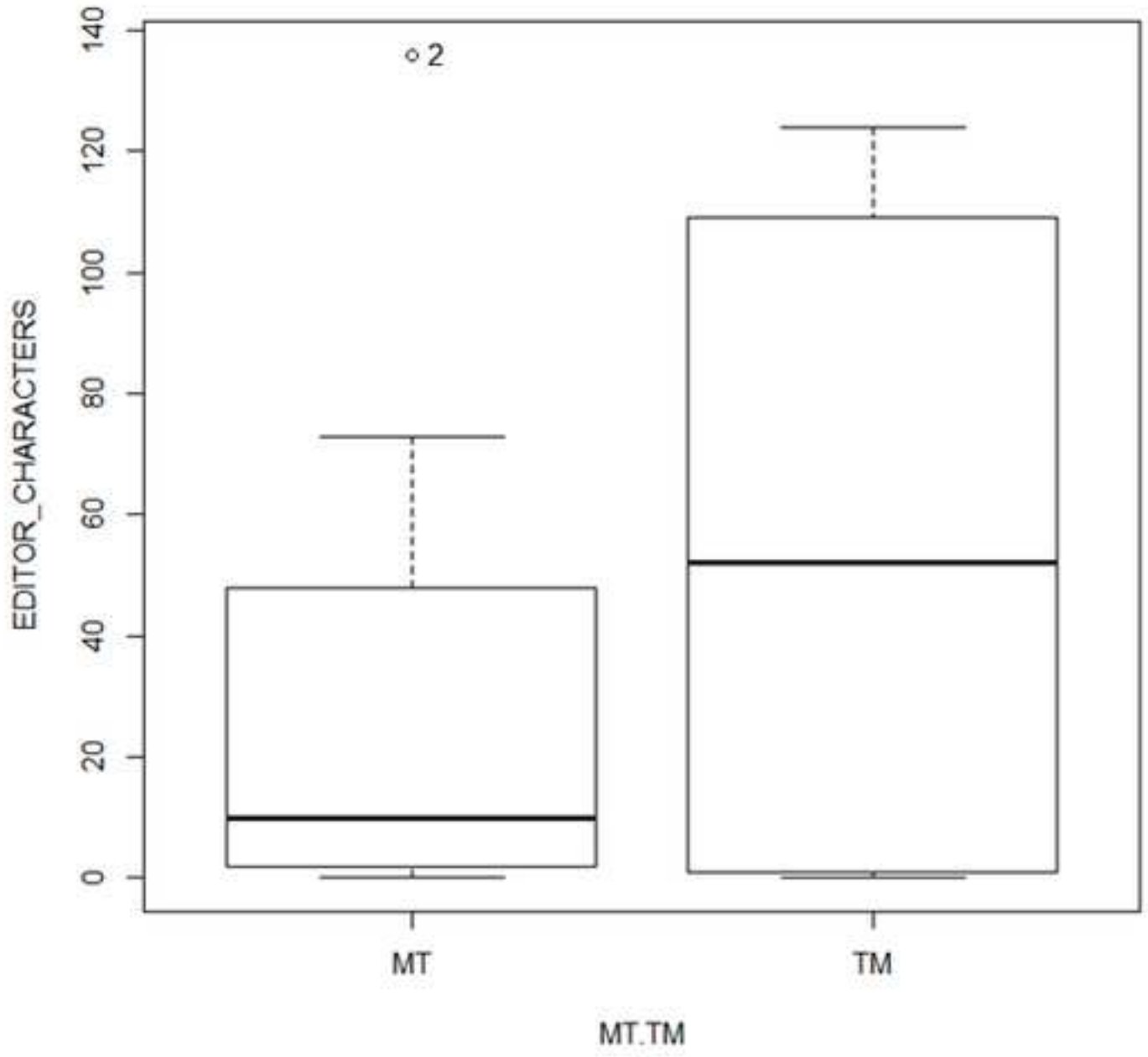
Figure

Click here to access/download;Figure;Figure_18-d.bmp



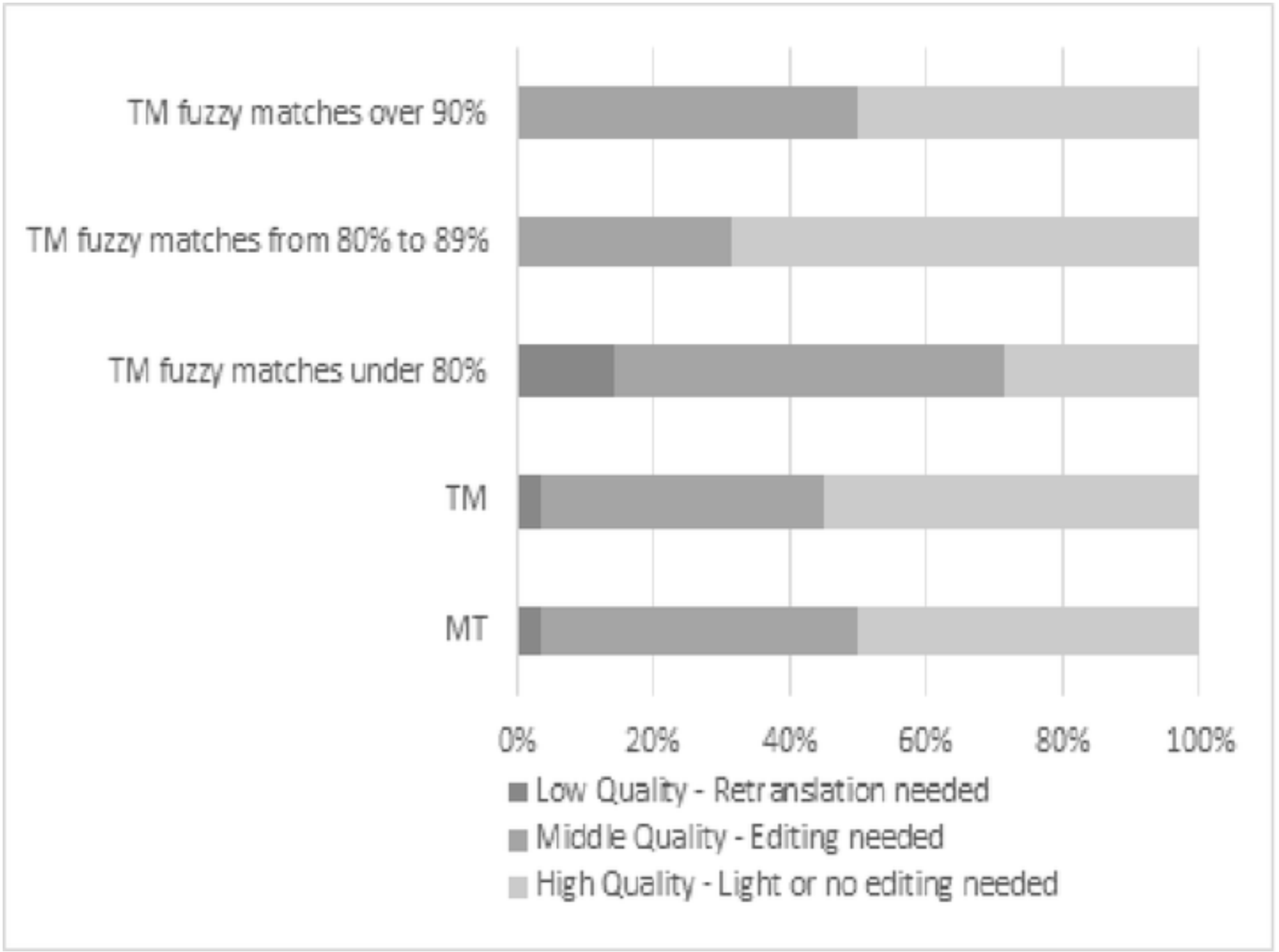
Figure

Click here to access/download;Figure;Figure_18-g.bmp



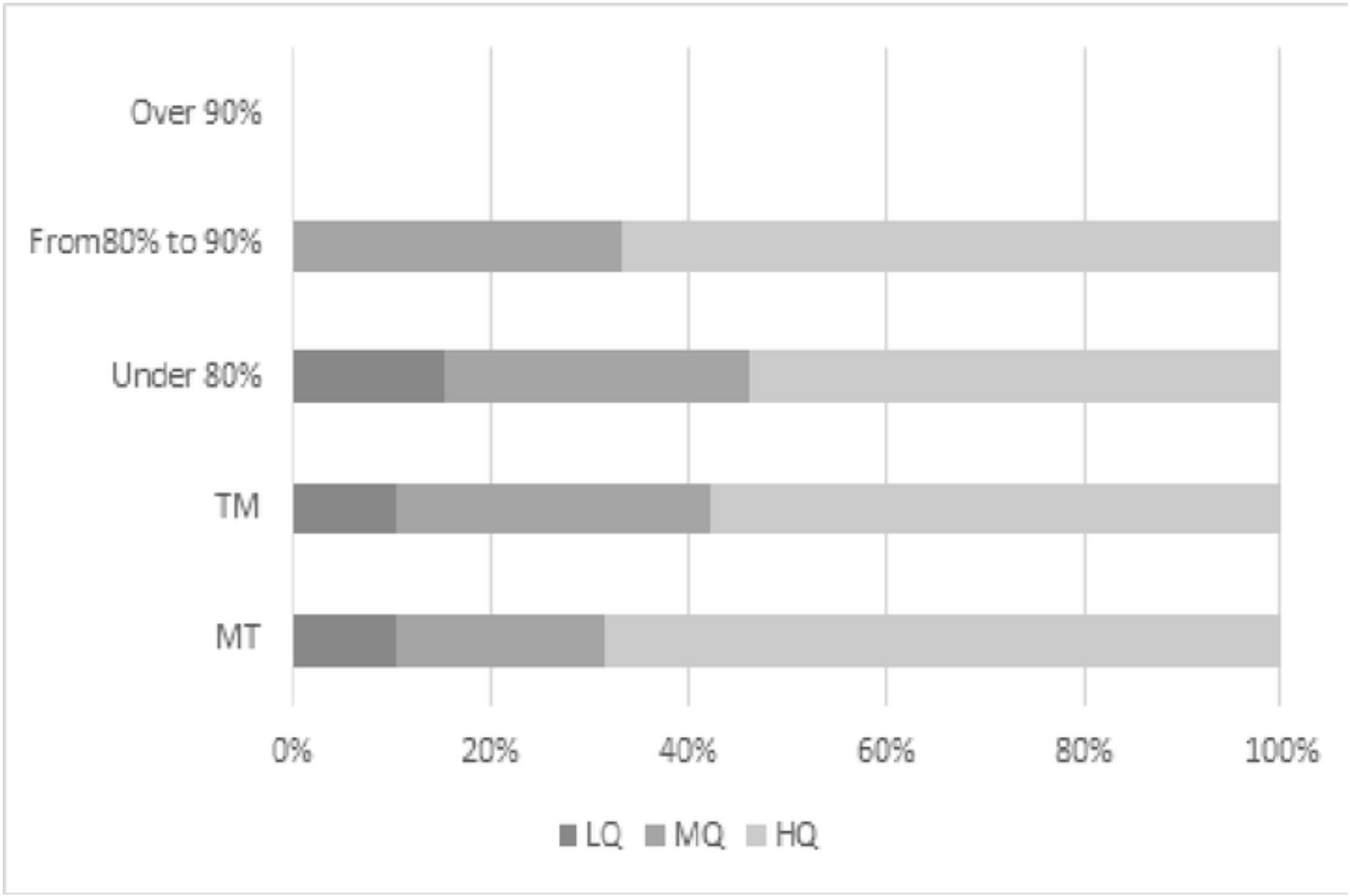
Figure

Click here to access/download;Figure;Figure_01.png



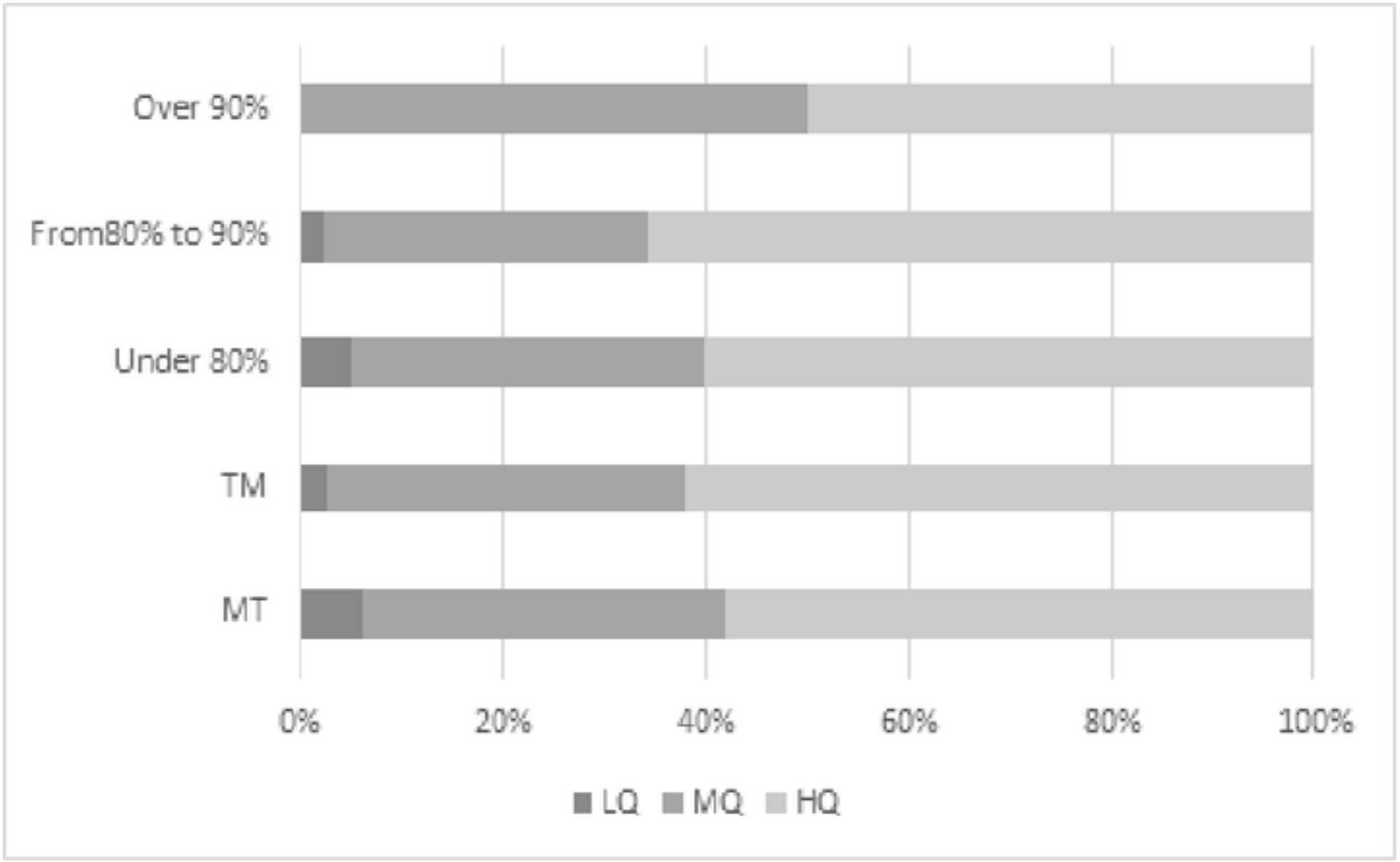
Figure

Click here to access/download;Figure;Figure_02.png



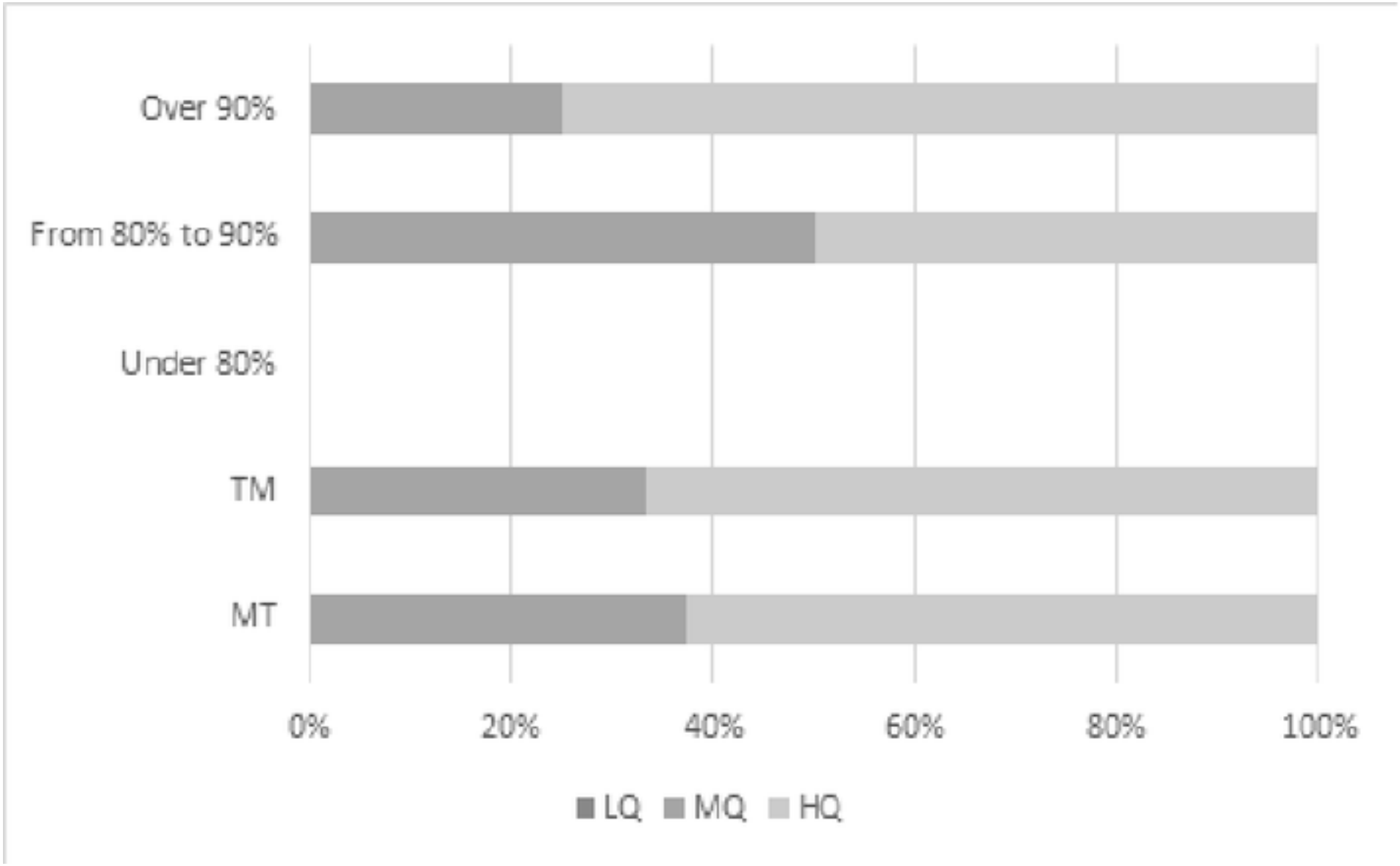
Figure

Click here to access/download;Figure;Figure_03.png



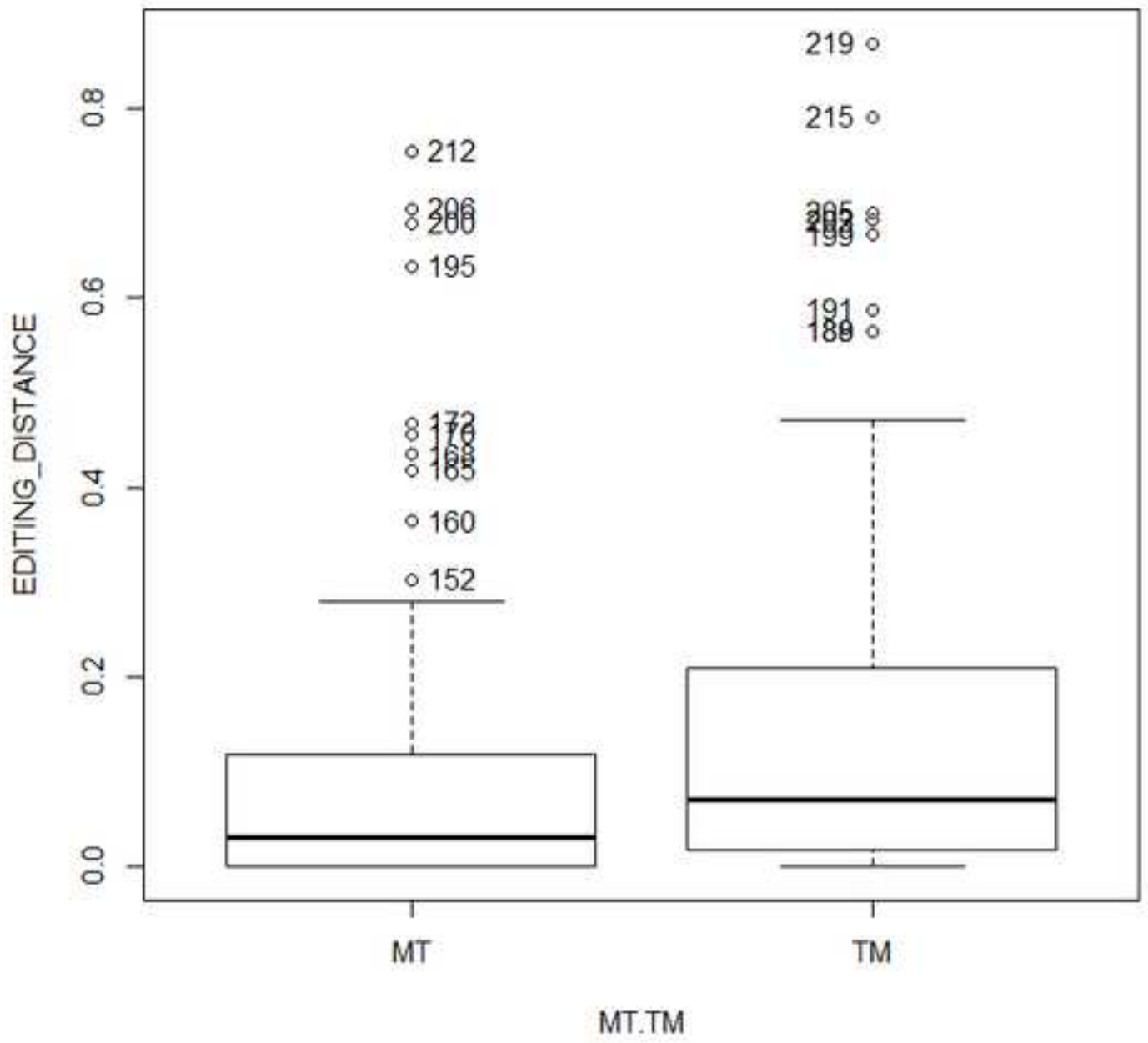
Figure

Click here to access/download;Figure;Figure_04.png



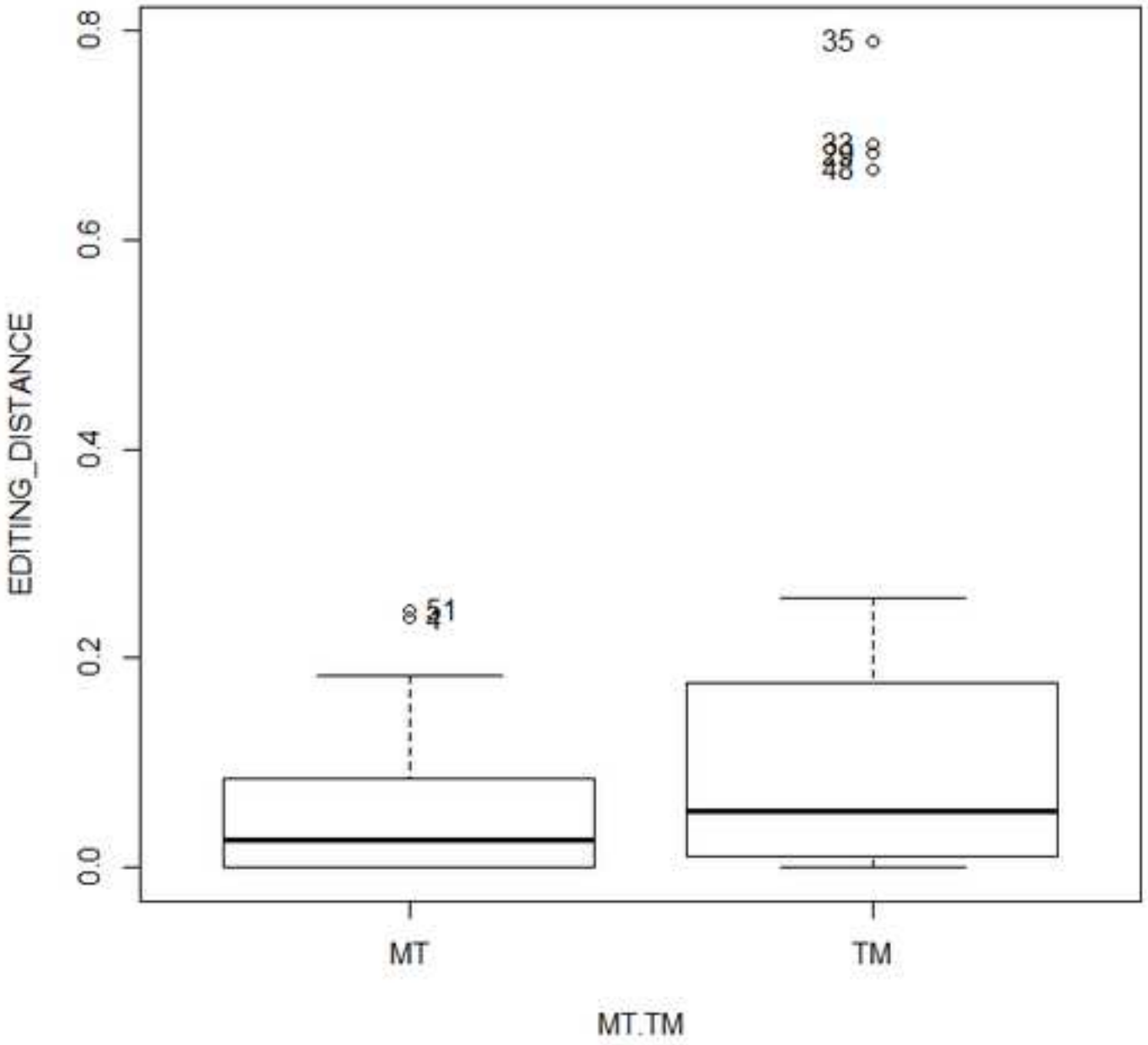
Figure

Click here to access/download;Figure;Figure_6.bmp



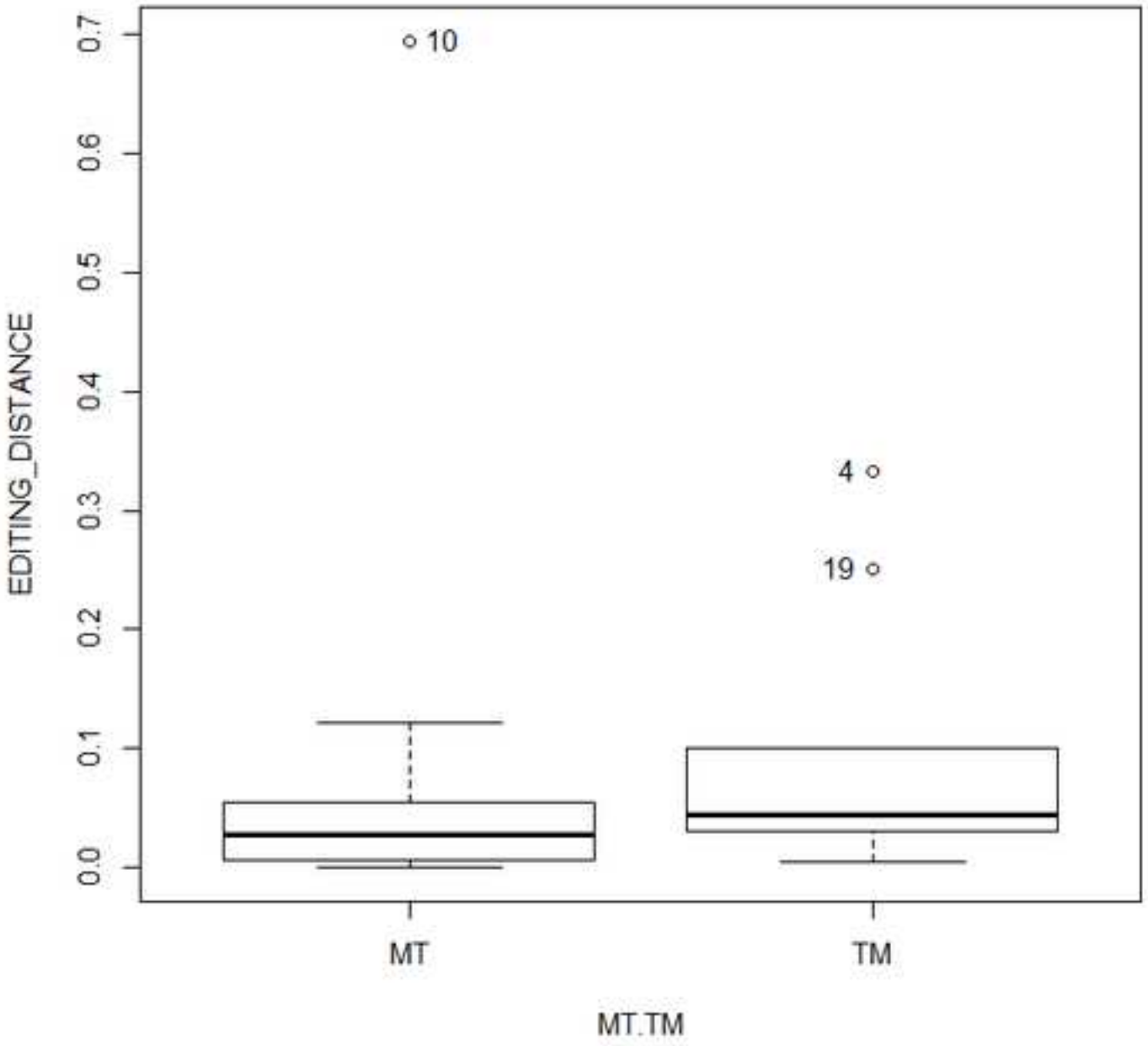
Figure

Click here to access/download;Figure;Figure_17e.bmp



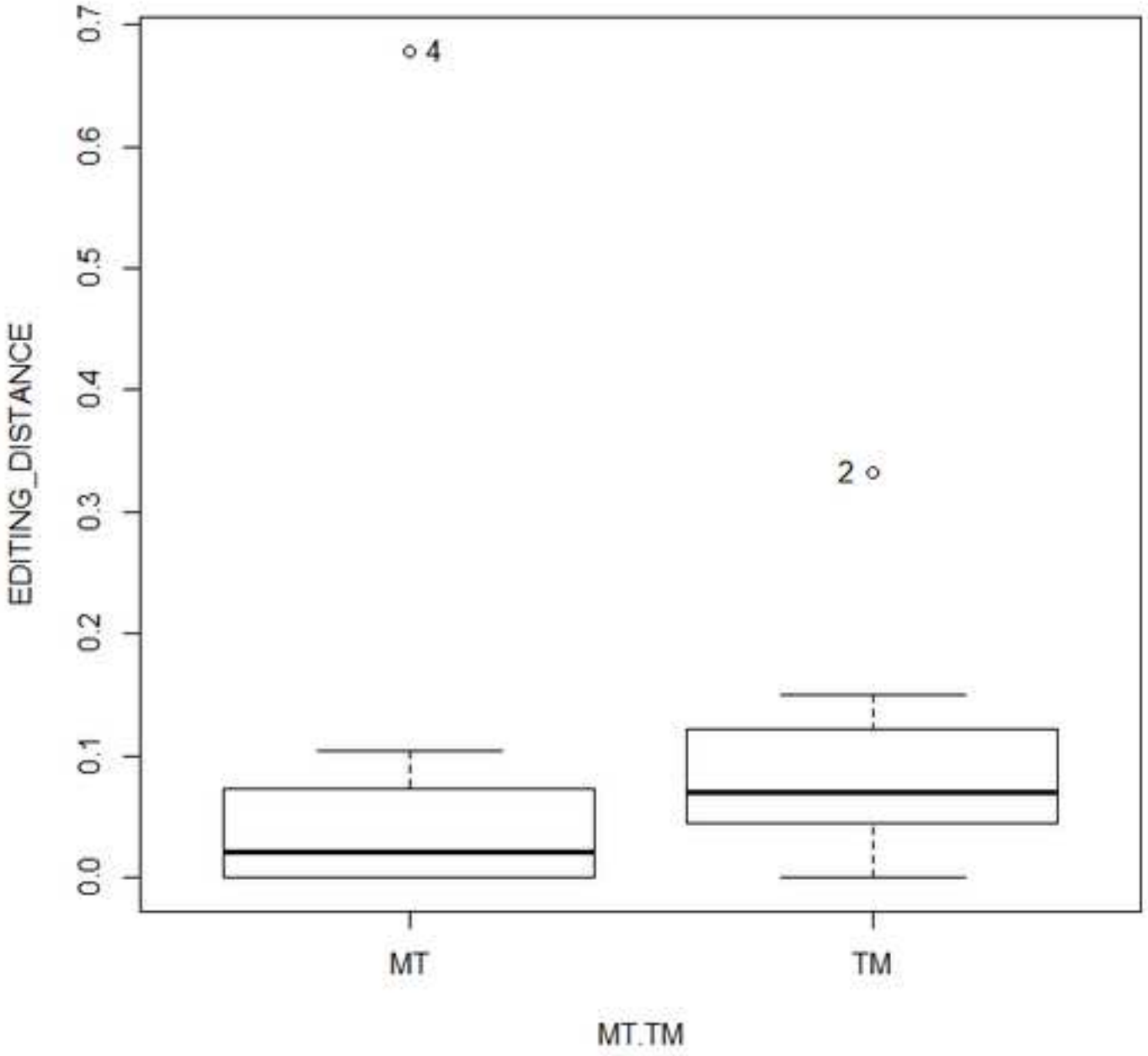
Figure

Click here to access/download;Figure;Figure_18b.bmp



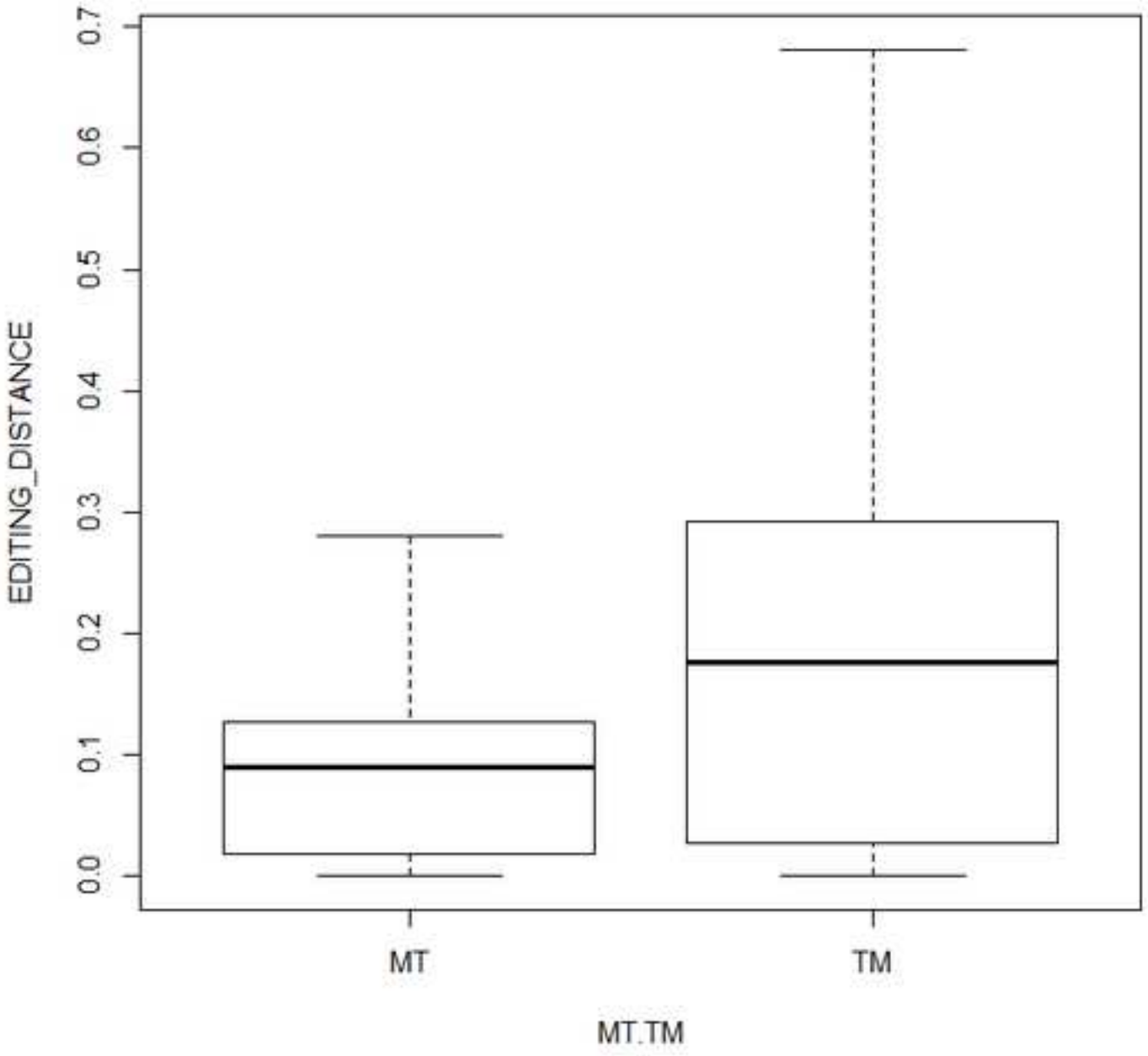
Figure

Click here to access/download;Figure;Figure_18e.bmp



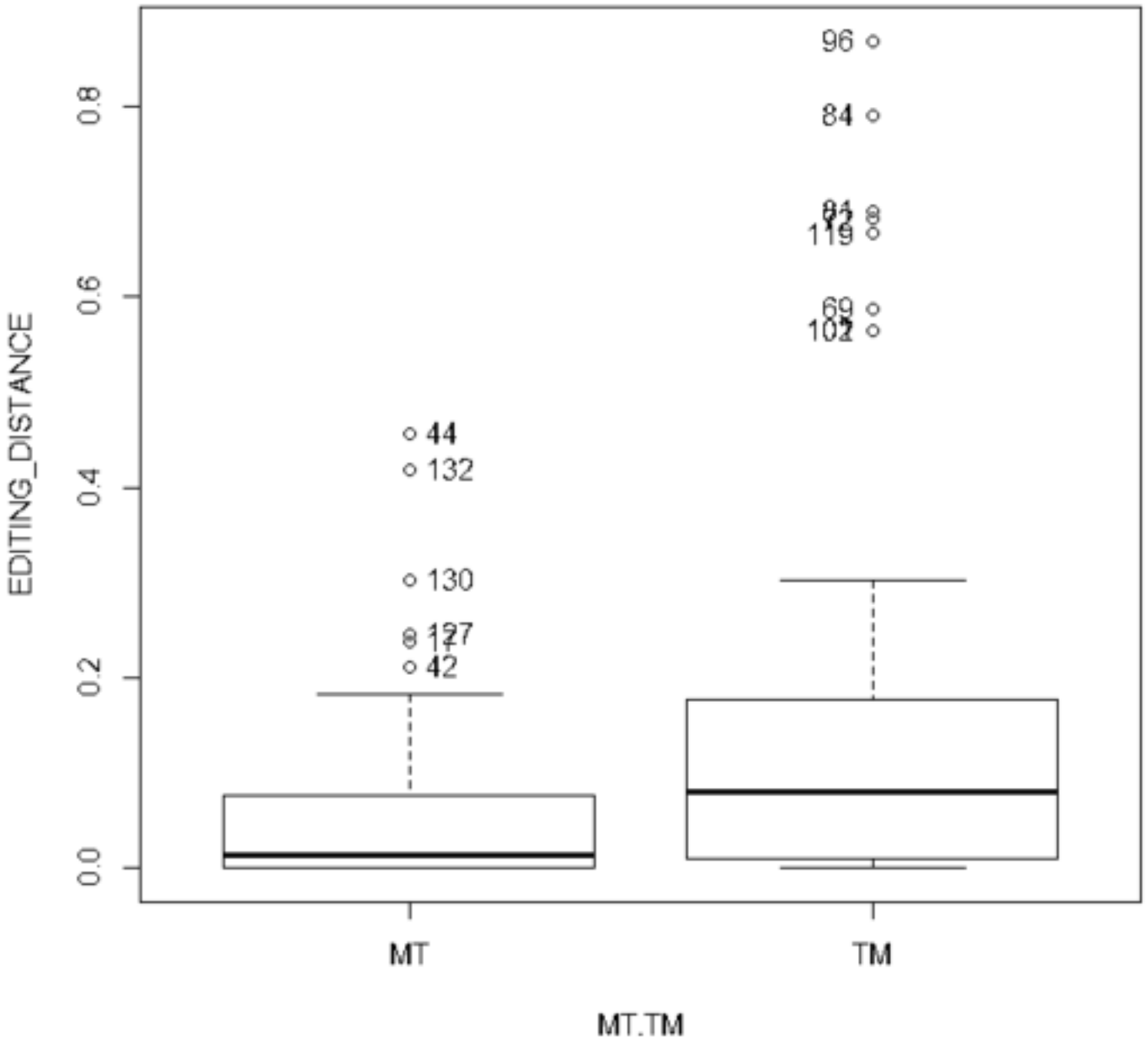
Figure

Click here to access/download;Figure;Figure_7a.bmp



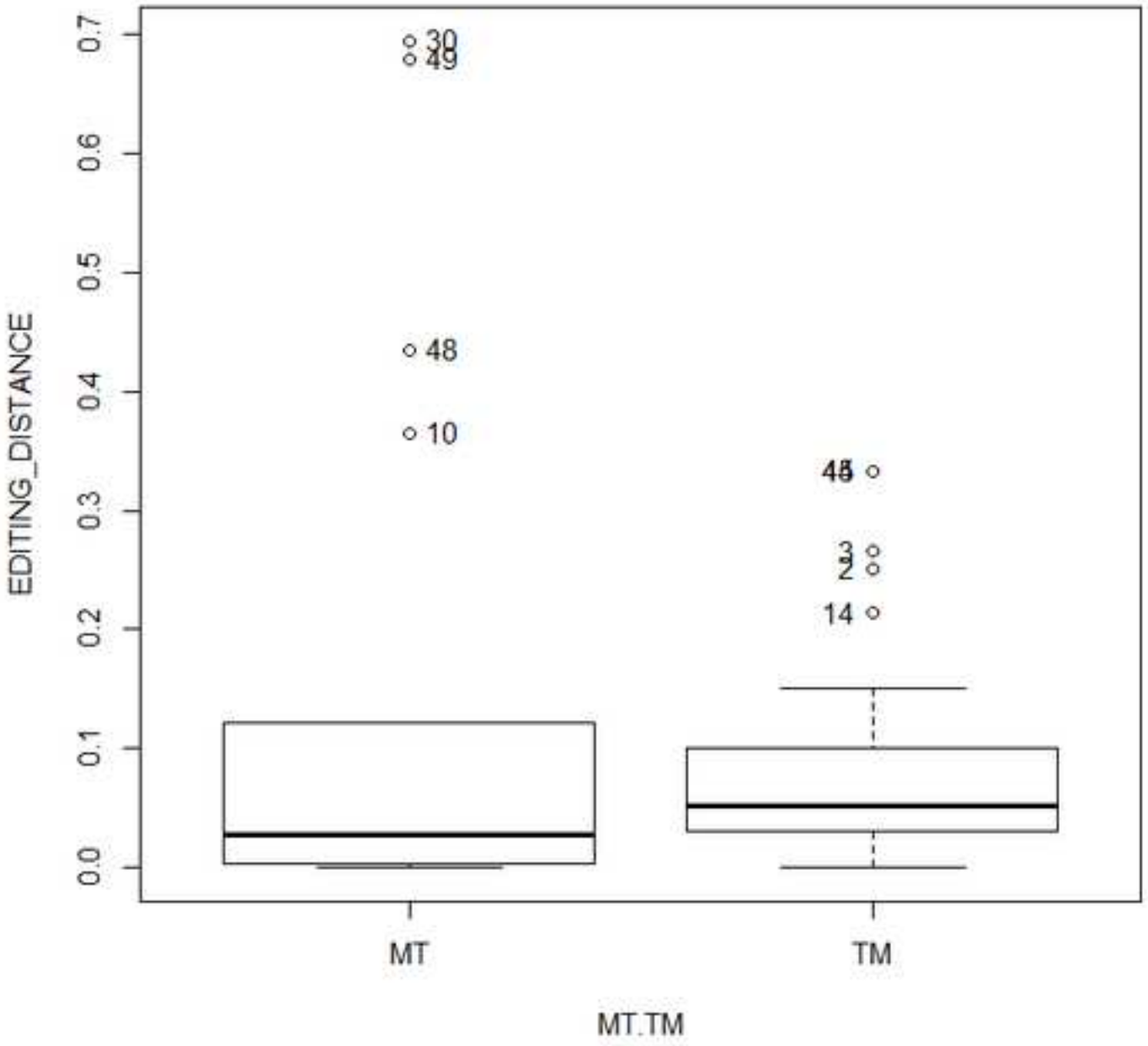
Figure

Click here to access/download;Figure;Figure_7b.bmp



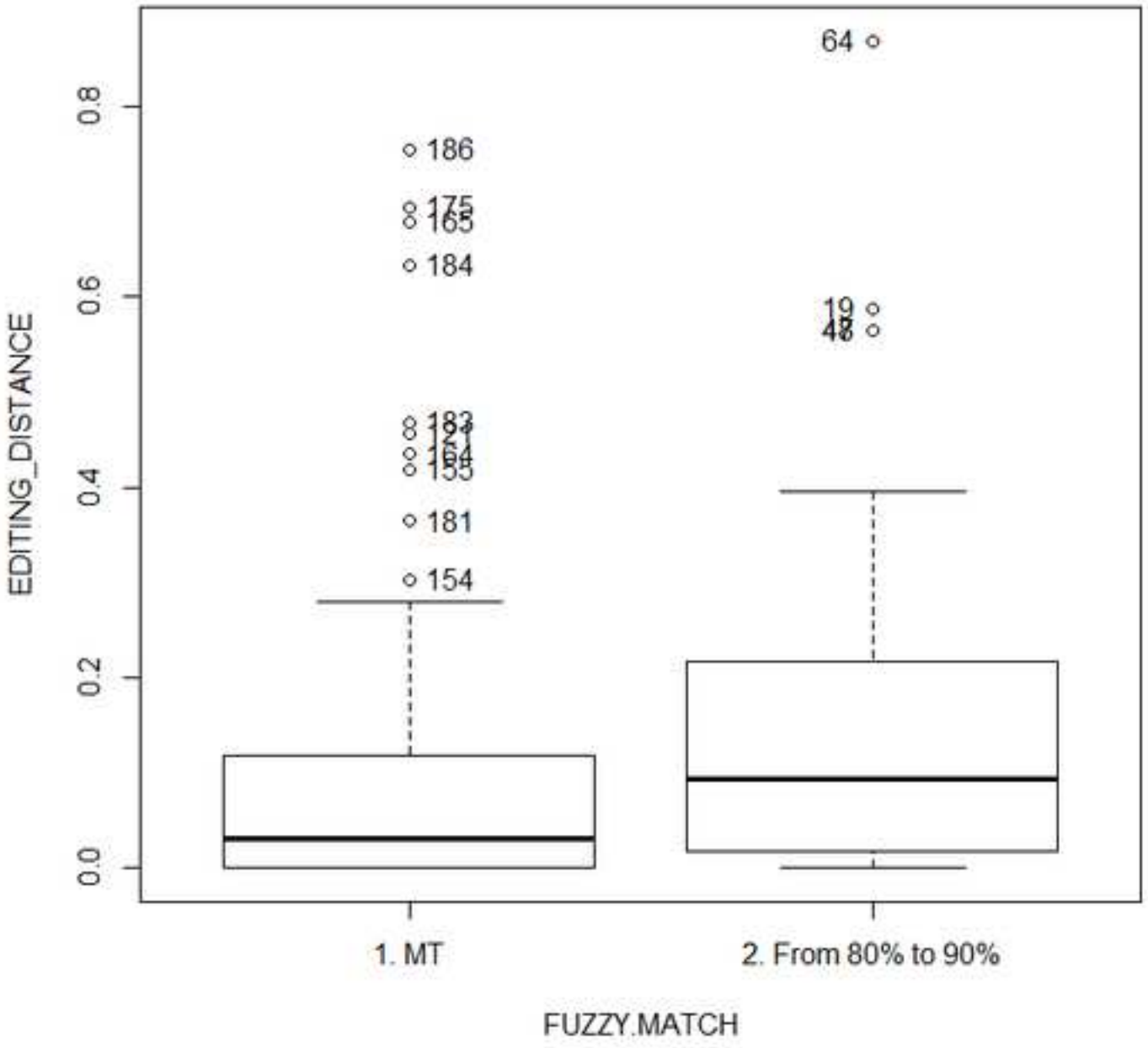
Figure

Click here to access/download;Figure;Figure_7c.bmp



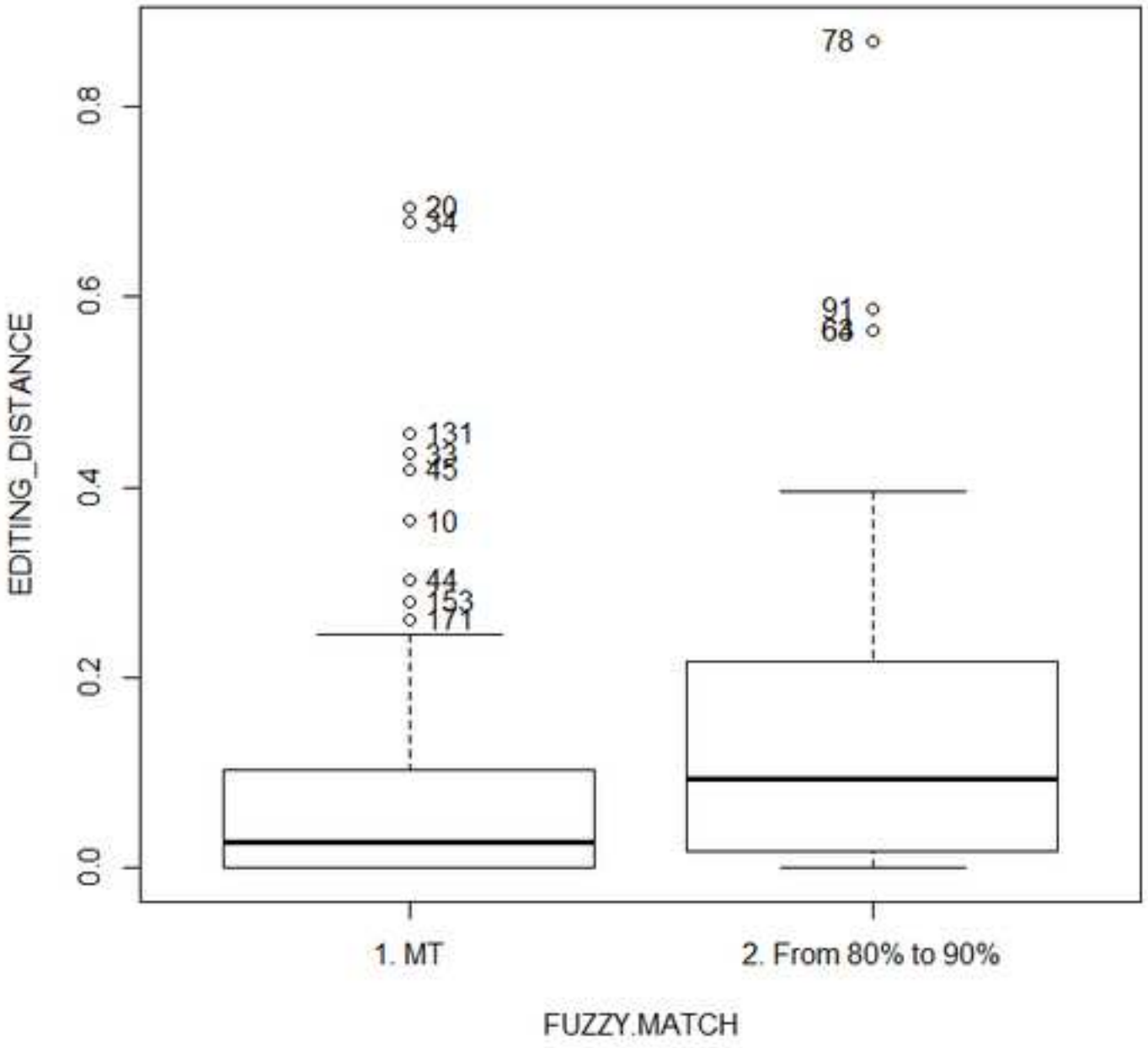
Figure

Click here to access/download;Figure;Figure_8.bmp



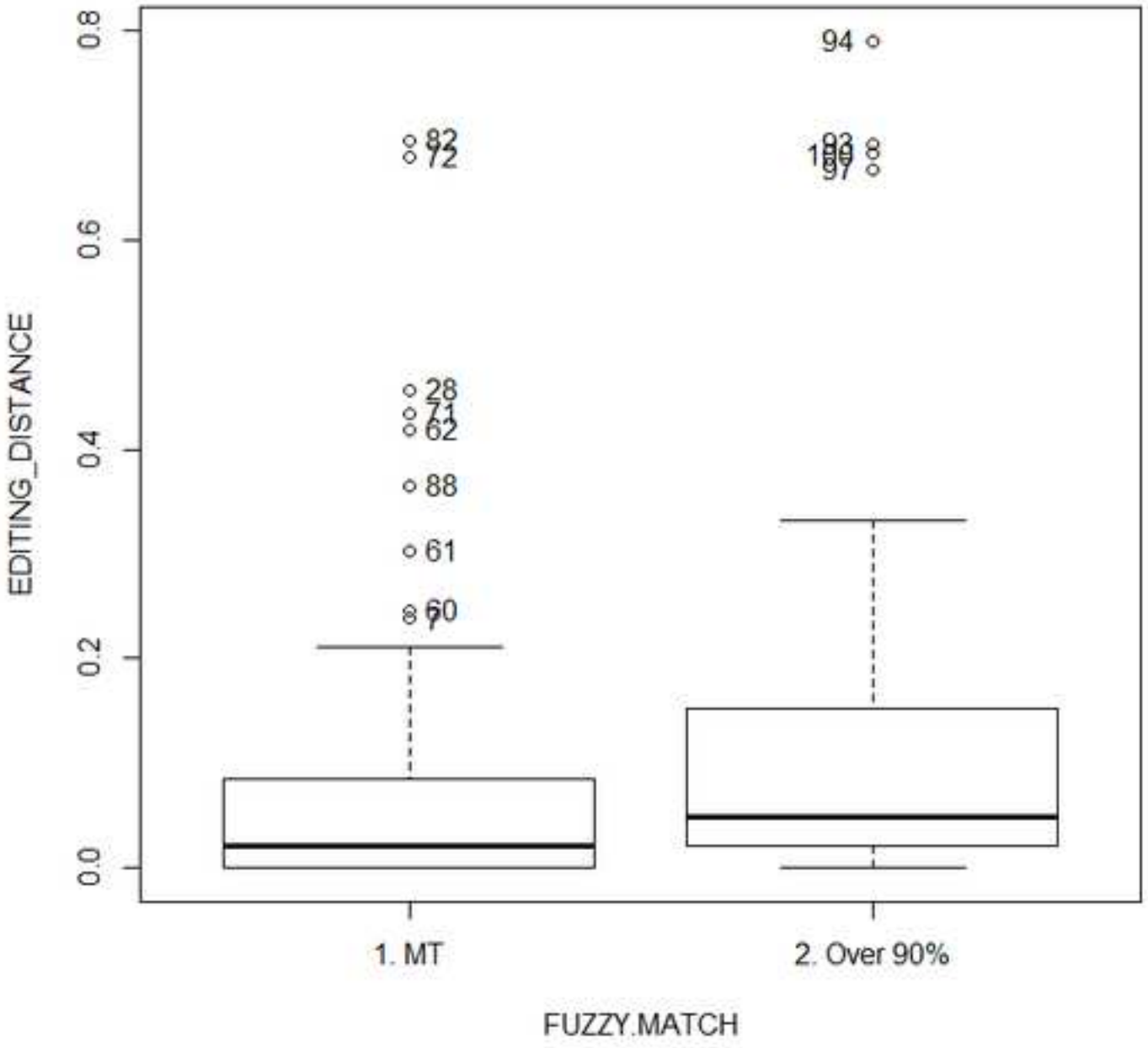
Figure

Click here to access/download;Figure;Figure_9.bmp



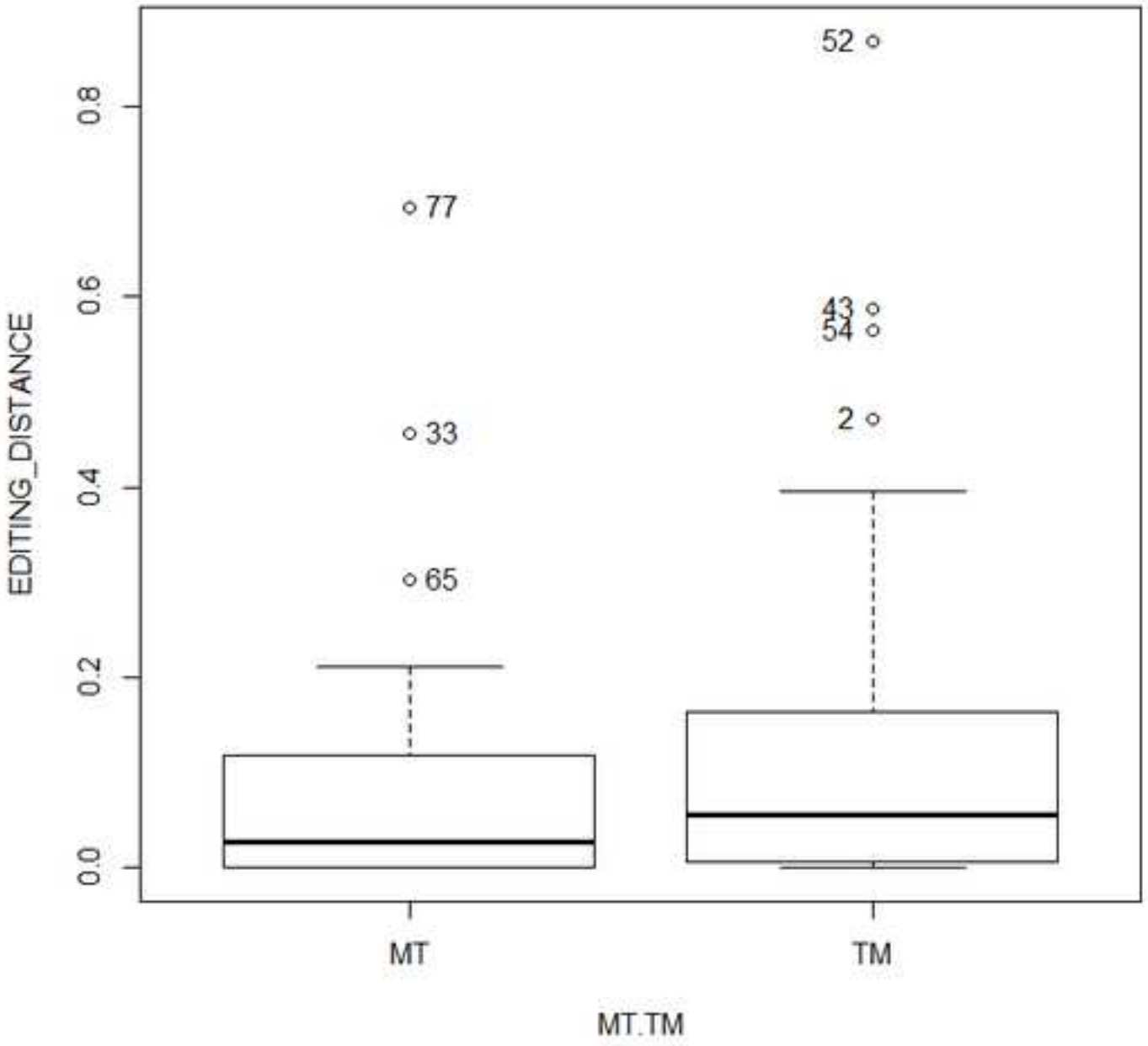
Figure

Click here to access/download;Figure;Figure_10.bmp



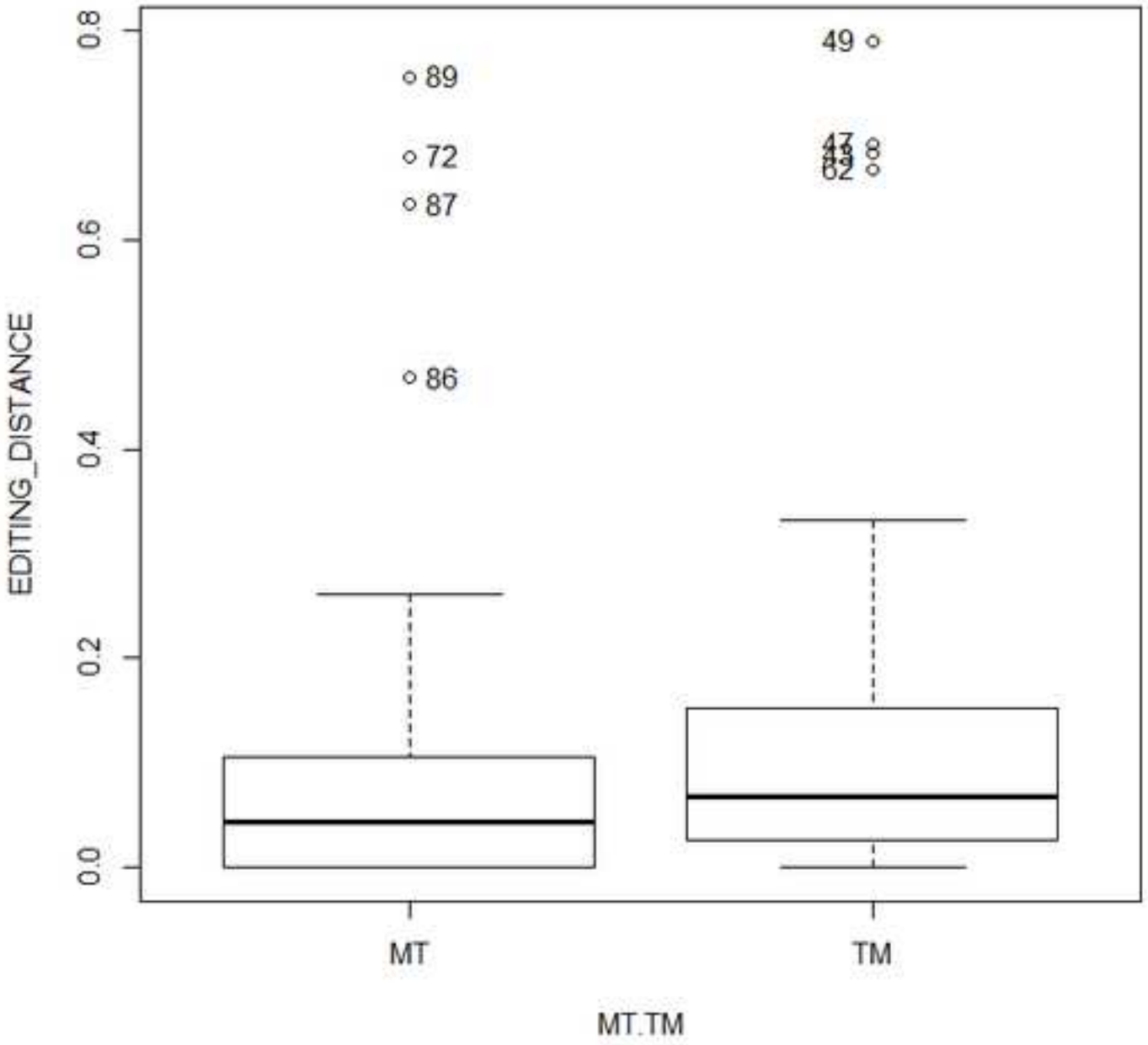
Figure

Click here to access/download;Figure;Figure_14b.bmp



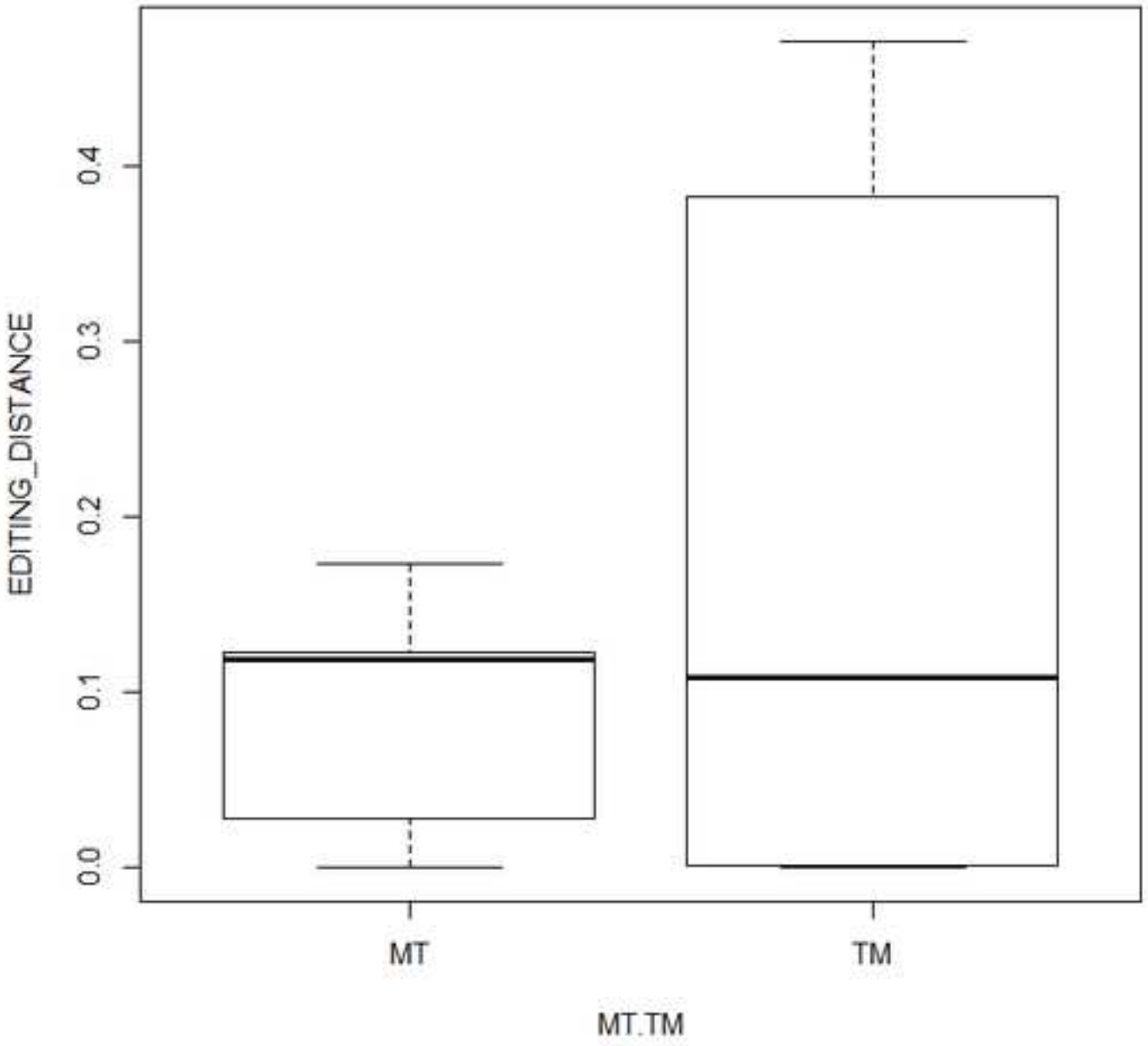
Figure

Click here to access/download;Figure;Figure_15b.bmp



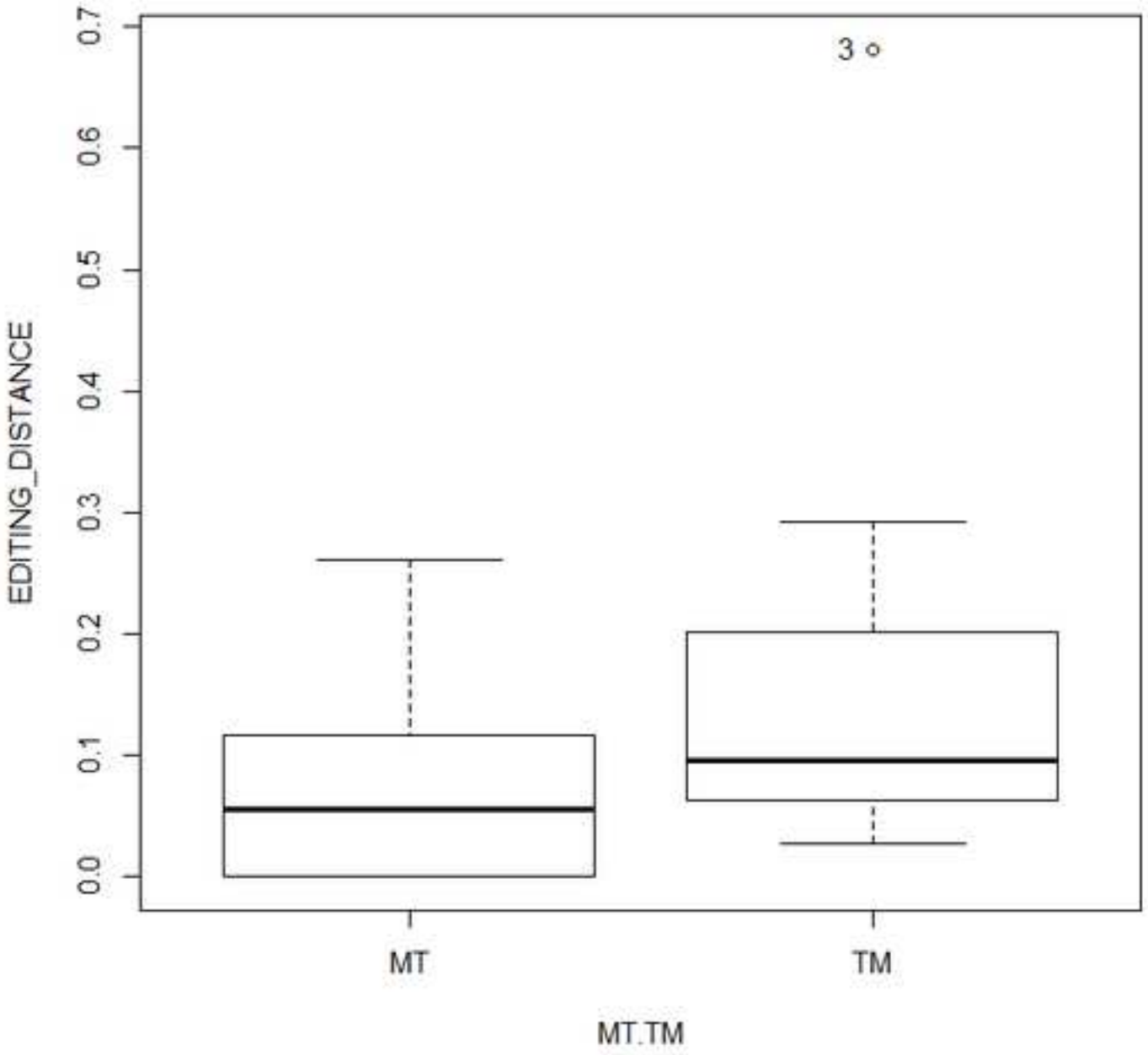
Figure

Click here to access/download;Figure;Figure_16b.bmp



Figure

Click here to access/download;Figure;Figure_16e.bmp



Figure

Click here to access/download;Figure;Figure_17b.bmp