English Studies at NBU, 2015
Vol. 1, Issue 1, 18-31

HOW FAR DO WE AGREE ON THE QUALITY OF TRANSLATION?

Maria Kunilovskaya

Tyumen State University, Tyumen, Russia

Abstract

The article aims to describe the inter-rater reliability of translation quality assessment (TQA) in translator training, calculated as a measure of raters' agreement either on the number of points awarded to each translation under a holistic rating scale or the types and number of translation mistakes marked by raters in the same translations. We analyze three different samples of student translations assessed by several different panels of raters who used different methods of assessment and draw conclusions about statistical reliability of real-life TQA results in general and objective trends in this essentially subjective activity in particular. We also try to define the more objective data as regards error-analysis based TQA and suggest an approach to rank error-marked translations which can be used for subsequent relative grading in translator training.

Keywords: TQA, translation mistakes, inter-rater reliability, error-based evaluation, error-annotated corpus, RusLTC

Article history:
Received: 10 April 2014
Accepted: 21 December 2014
Published: 1 February 2015

Maria Kunilovskaya, Ph.D (Tumen State University), is an Associate Professor with the Department of Translation and Translation Studies, Institute of Philology and Journalism, Tyumen State University (Russia). She specializes in translation studies and teaches courses in translation and interpreting. Maria is involved with translation error annotation in the on-line multiple translation corpus Russian Learner Translator Corpus (http://rus-ltc.org). Her research interests involve translation quality evaluation, comparative text linguistics, parallel corpora and NLP.
Email: mkunilovskaya@gmail.com

This research deals with inter-rater reliability of quality assessment in translator training. Our data come from two university translation contests which involved ranging translations based on aggregated grades awarded to translations by several raters and from error-based description of quality as part of routine training. In all cases the translation tasks involve standard, largely informative texts set in non-specific pragmatic and communicative situations, while assessment aims at summative and formative evaluation of the overall translation quality as a measure of general transfer competence (Neubert, 2000, p. 9-15), which is mostly centered around (macro)linguistic issues. The research is limited to translation into the trainees' mother tongue, which to an extent helps concentrate on those components of translation competence which deal with transfer proper as it drives active foreign language skills out of focus (Zwilling, 2009, p. 60).

The subjective nature of TQA is widely recognized in translation studies, yet at the same time it is not seen as a bar to employing statistical measures to gauge it (Zwilling, 2009; Kelly, 2005, p.140; Knyazheva & Pirko 2013; Waddington, 2001, p.24). For both major approaches in TQA – holistic and error-based – there are descriptions and grading scales, which, if successfully acquired by the raters involved, and within a carefully staged experiment, can yield statistically reliable results as shown by Waddington (2001).

This article aims to describe inter-rater reliability of real-life TQA carried out with different methods and assessment criteria and various degrees of their discussion by the assessors. In this research we also put our translation error classification, proposed for error annotation in Russian Learner Translator Corpus (RusLTC[1]), to a reliability test.

In his article on TQA, Williams insists that any assessment model should comply, inter alia, with the requirements of reliability, which is defined as "the extent to which an evaluation produces the same results when administered repeatedly to the same population under the same conditions. Thus a TQA system is reliable if evaluators' decisions are consistent and criteria are stable" (Williams, 2009). In statistics, inter-rater reliability measures show how much agreement there is among raters, by giving a

[1] http://www.rus-ltc.org

score of how much consensus there is in the ratings given by different judges as regards the same object and assessment criterion[2].

Throughout this research we rely on one of the various statistical measures of the inter-rater agreement called Krippendorff's alpha. This coefficient is a statistical approach to generalize several known reliability indices, which (unlike other measures) can be applied to data produced by more than two raters, using any metric or level of measurement; it takes into account chance agreement, can handle incomplete and missing data and allows for algebraic differences between the units of the scale. This coefficient range is $1 \geq \alpha \geq 0$, where $\alpha=1$ means perfect agreement, $\alpha=0$ means that units and the values assigned to them are statistically unrelated, while $\alpha<0$ means that disagreements are systematic and exceed what can be expected by chance (Krippendorff, 2011). This measure originated in content-analysis research and is used in humanities to assess manually coded data (Artstein & Poesio 2008). It is considered more reliable than other reliability measures such as percent agreement or Cohen's kappa. The calculations have been performed using the on-line service developed by Deen Freelon (Freelon, 2010[3]).

In Section 2 of this article we calculate and describe the inter-rater reliability of TQA in the context of translation competition. This description is based on the scores reached by several raters without prior discussion of either criteria for assessment or evaluation method. It shows how much variance there is in the professional community as far as the opinions on the overall translation quality are concerned. Section 3 contains the description of inter-rater reliability of translation error-analysis based on an agreed error typology. In this case we measured consensus between raters as to the locus of the error and its type, as well as its seriousness. Special emphasis is made on the more subjective and less subjective areas in our implementation of error analysis. In Section 4 there is a description of an experiment which involves both error-analysis and subsequent grading, which helps to establish correlation between number and types of errors and translation ranks in the sample. Section 5 draws comparisons of the results and conclusions of the study and provides an outlook for the error-based TQA research and its classroom applications.

[2] http://en.wikipedia.org/wiki/Inter-rater_agreement
[3] http://dfreelon.org/utils/recalfront

Inter-rater Agreement for Random Panel of Professional Translators and Teachers

The data in our first sample was collected under the following conditions. Six independent raters with different affiliations, including translator trainers and acting translators, assessed 70 student translations produced during a translation competition held by one of the Russian universities. They used a 15-point scale, which was applied as the raters saw fit, i.e. no translation values or assessment criteria were discussed beforehand. All raters worked independently under a reasonable time constraint and the contestants names were encoded, which ruled out any personal bias. The translation brief required a translation of excerpts from a magazine article (352 tokens in size, with full text available) on a general subject aimed for general readership and included no specific communicative challenges.

In order to calculate the alpha coefficient for inter-rater reliability we assumed that the data received from the competition jury of the six people mentioned above are interval by nature. It means that on a translation quality assessment scale from 0 to 15, it is the difference between values that matters. The higher the score, the better the translation, but unlike ratio scales (such as height or weight) the interval scale doesn't have a rational zero point which marks an object with no attribute in question (in our case a translation with no quality of translation). Also we can't say that a 10-point translation is as much better than 8-point one as 14-point translation is better than a 12-point one, which would be the case if the scale belonged to the ratio type.

The inter-rater reliability of data obtained under the conditions described reaches the value of $\alpha=0.569$. The author of the modern mathematical structure of this coefficient specifies that though the minimum acceptable alpha coefficient depends on the importance of the conclusions drawn from the imperfect data, the common threshold is known to be $0.800 > \alpha \geq 0.667$ (Krippendorff, 2004), and in other research $\alpha>0.74$ is described as perfectly reliable (Strijbos & Stahl 2007).

The relatively low degree of agreement suggests that the translation contest jurors either have been very inconsistent in assessing translations, lack the linguistic or subject-field knowledge required, or else they have very different opinions on what is

good in translation, i.e. they have used very different yardsticks for gauging quality. Lack of any quantification or description behind the points awarded by evaluators does not allow any further analysis in this case.

Nonetheless, further analysis of the data shows that Juror 3 can be considered an outlier. This term is used in statistics to refer to observation points that are too distant from other observations. It is possible to exclude these data from the set for statistical analysis. If we exclude data from Juror 3, the agreement between the remaining five jurors jumps to 0.676. The specific approach taken by Juror 3 is confirmed by agreement statistics for any team of five jurors including Juror 3: it amounts to 0.543 without Juror 4, α=0.512 without Juror 1, and α=0.590 without Jurors 5 or 6. Note that all these figures are lower than that for the jury without Juror 3.

By comparing statistics we have found jurors who agree the most (Jurors 1 and 2 show inter-rater agreement α=0.824, Jurors 1 and 6 – 0.774, and the result for any other possible pair does not exceed α=0.539).

The low reliability of TQA results revealed in this research signals a good deal of subjectivity in assessing translations and disagreement within the professional and educational community when the assessment is exercised in the holistic setting. For the results of this approach to be reliable it requires a much more complicated procedure than the sum total of the points awarded by all raters, which determined the competition winner in our case. As shown by Knyazheva and Pirko (2013), the variety of opinions displayed within a holistic approach to translation assessment can be fairly accounted for on the basis of system analysis methods. It requires formulating criteria and prioritizing them in terms of significance to the overall translation quality as well as meticulously assessing translations according to these criteria.

Error Analysis Reliability

In the second experiment we aim to describe different aspects of reliability of data that come from translation assessment, performed by two translation teachers, who used a pre-defined error typology to mark up mistakes in student translations. The inter-rater agreement is described as agreement between raters as to the error location in translation and as to the type and seriousness of errors marked.

This statistics is supposed to highlight the types of mistakes that are spotted and agreed upon by both independent raters versus those which cause most disagreements and, therefore, can be considered more "subjective". Besides, the results of this analysis and the discussion of its results will help to determine the faults of the proposed error classification and improve it before it is used for RusLTC mark-up.

The sample under analysis included 27 anonymized translations from English into Russian of 6 original newspaper texts which added up to 7874 tokens in size. The number of translations to each original varies from two to seven. All translations were done by students majoring in translation studies and translation.

The two evaluators worked independently on the basis of RusLTC Translation Error Mark-up Manual, which contains the general description of the translation error classification, its principles and examples for each type of mistake along with a commentary. The mark-up was technically performed in the customized version of the text annotation program brat (Stenetorp et al, 2012) installed at RusLTC site. It creates standardized text annotations that can be processed automatically.

Within the scope of the present research we do not analyze agreement in all types of mistakes provided for in the classification, and we will not describe the latter here in detail, limiting ourselves to characterizing it as a three-level hierarchy which includes 30 mistake types equally split between two major categories – content-related and language-related, depending on whether the mistake affects understanding of the source text or expression in the target language respectively[4]. In addition to defining the category and the specific type of mistake, the raters were also asked to evaluate them in terms of seriousness using a three-member scale (critical, major and minor) and considering the effect of the errors on the overall quality of translation.

The figures for the total number of mistakes marked by the raters in the same targets in our sample differ substantially, but the ratio of content-related and language-related mistakes is very similar (see Table 1). It means that the raters differ in the rigor of mistakes analysis, i.e. they show different degree of tolerance for mistakes, especially

[4] The complete classification can be found at the RusLTC site http://www.rus-ltc.org/classification.html .

when it comes to target language accuracy. This conclusion is further confirmed by the striking difference in the number of critical errors.

To determine the inter-rater reliability of these data we have tried to examine how often the raters agree on the mistakes locus in the translation and the mistakes types and seriousness. In the first case we calculated the quantity of mistakes marked by both raters in the same text span. In our sample the raters agreed on the locus of a mistake in the text in 343 cases, including 33 cases of double or overlapping annotations. It makes 54.4% for Rater 1 и 76.6% for Rater 2 (see Table 2).

This means that raters more often agree that translations are faulty in a particular text fragment. It is important to highlight that our data are characterized by the high percent agreement on the category of the mistakes located by both raters – 80.5 %; whilst they only disagree in 67 cases out of 343. The fact that raters tend to agree on the general type of mistake which they both locate in a particular text span confirms the validity of the traditional dichotomy between content-transfer errors and target language errors that are often used as the top-level categories in translation error hierarchical classifications.

If we bear in mind that for each rater target language-related mistakes prevail in our sample, it is no surprise that they are more numerous among the “locus- and type-agreed” mistakes. It is noteworthy that the ratio between language and content mistakes in this part of the data is tilted towards the former – it is 0.747. We can therefore conclude that our raters tend to agree on content-related mistakes a bit more than on target-language related ones.

At the same time, we have to admit that in our first inter-rater experiment “subjective” mistakes (those that are accounted for by only one of the raters) make up a considerable part of the data – 45.6% for Rater 1 и 23.4% for Rater 2 (solid sectors in Fig. 1). In the case of Rater 1, “subjective” mistakes together with cases of disagreements about the type only (shown with the dotted background in Fig. 1) account for more than 50% of the annotation data for this sample.

The figures for the second rater, who showed much more tolerance for language mistakes, are less dramatic. The more subjective area of translation mistakes mark-up

extends to include the degree of mistakes gravity. The raters agreed on this attribute of mistake only in 34.4% of cases. At the same time one can notice that "disagreement" sectors (solids and dots) are always smaller in the area of content-related mistakes (darker sectors in Fig. 1), which speaks of higher agreement on the more serious truly translational mistakes, rather than those associated with language competence.

Drawing conclusions for this part of the research we can summarize it as follows. Our research shows that 1) our raters spot a mistake in the same text locus in more than half cases; 2) out of those, they agree on the type of mistake in more than 80% of cases; 3) they tend to agree more about content errors than language errors. On the other hand, they disagree substantially on 1) degree of tolerance to minor mistakes; 2) the nature and number of good solutions, and 3) the way in which to apply the classification, even at the level of mistakes categories, all of which undermines reliability of the error annotation and points at its subjectivity.

Translation Evaluation Based on Error-analysis

To improve the inter-rater reliability determined in the previous experiment we have introduced changes into the classification, discussed results of the research, and before proceeding we developed and discussed a translation of the source text that could be used for reference by the evaluators.

In the second error-analysis experiment we compared error annotations made by three raters, two of whom had already taken part in the previous experiment. The raters error-annotated 17 translations of the same text (EN>RU, source text size - 571 tokens), and then awarded each of them a grade, based on a 20-point scale. It is important to foreground that they did not use any agreed standard to convert number and types of errors into points, but relied on their own understanding of each translation relative worth.

The inter-rater reliability of the three raters' evaluations in points of the interval scale measured with Krippendorff's statistics for this sample is 0.734. For reasons described above it is close to acceptable.

Estimating reliability of TQA data in this research we have found out that raters tend to agree more on the poorer translations than on the better ones. In the first sample, the agreement between the three raters with the highest level of internal consistency of the data provided (Raters 1, 2 and 6) on the bottom ten translations (according to the aggregated score of the contest results) is estimated as 0.425. In the current sample it is 0.607. Krippendorff's alpha for the top ten translations is 0.127 and α=0.265 in the first and the second samples respectively. As it can be seen from these figures the lower subgroup of translations causes less disagreement between the raters than the higher subgroup.

If we compare the data from this error-analysis experiment to that obtained in the previous experiment, we can register certain improvement in agreement between Raters 1 and 2, while Rater 3 has contrasting results (see Table 3). The difference in data can be attributed to different degree of tolerance to target language mistakes which do not affect understanding and to different levels of understanding of the classification itself. Nonetheless, these data show that the annotations are still dominated by target language mistakes, while the concept of "a good translation decision" remains elusive.

As it has been stated in the case of mistakes analysis, we estimate inter-rater agreement as consensus on locus of mistakes in the text and mistakes type. In the sample used for experiment three, there are a total of 109 words or phrases which, according to the three raters, are erroneous translator decisions. These "more objective" mistakes account only for 1/5 or 1/4 of all mistakes marked by each rater. But if we exclude data from (untrained) Rater 3, the agreement between Raters 1 and 2 will jump to over 2/3 (69.35% and 71.59% for Raters 1 and 2 respectively; the number of mistakes located in the same place in the target text is 310, including 242 which are referred to the same type). It is interesting to note that the figures for the total number of mistakes for these raters are very close (in contrast with Rater 3), which means that the raters applied more or less the same rigor when conducting their mistakes analysis, while percent agreement and Krippendorff's alpha for agreement on the type of mistakes is a bit lower (80.5% and 78.1%; α=0.605 and α=0.561 for the data in the 1st and 2nd error analysis experiments respectively). We attribute these differences to the

insignificant statistical variance related to the nature and size of the sample under analysis.

These data prove that additional training helps to achieve more reliable data, at least in terms of total number of mistakes. However, agreement on the type and seriousness of the mistakes does not improve much, which means that evaluators tend to agree that a particular phrase is not an adequate translation solution, but they disagree on how to describe it in the categories of error classification and on how to value the seriousness of the error. The data from the second experiment confirm our previous conclusion that the raters tend to see more target language mistakes than content-related ones (the percent of the former varies from 60 to 68%, in Fig. 1 the light-green sector is always larger), but when it comes to the agreement on type of mistake it is the content-related mistakes that cause less disagreement (in Fig. 1 solid dark-red sectors are disproportionally smaller than solid light-green sectors). The agreement on seriousness of mistakes improved, too, from 34.8% to 59.6%, but these data are far from reliable as regards critical mistakes.

However subjective the translation mistakes annotations are, we hypothesized that there is a correlation between the number and types of mistakes and the number of points awarded to each translation according to a certain scale and reflecting the overall relative quality of students' production. To find this correlation we sorted tables containing results of error analysis (such as total number of mistakes, number of target language mistakes, number of content mistakes, number of critical mistakes, number of good translation solutions) and the evaluation in points produced by each rater. It turns out that the most reasonable way to range translations is to take into account the number of critical errors, the number of content-related errors and the total number of mistakes in this consecutive order as this ranging better reflects, in our opinion, their relative quality and can be used for grading translations. In each individual situation a teacher can determine the baselines between different quality groups (grades) depending on the text difficulty, time constraints or other conditions of translation. The grades can be further adjusted to accommodate the number of "good translation solutions" marked-up in translations.

Conclusion

This research has shown that TQA, although rather subjective, does have objective trends that can be used to produce reliable data for further analysis. For both samples of translations that were assessed according to different scales inter-rater reliability of TQA results amounts to α=0.784 and α=0.734. Generally, raters tend to agree more on bad translations than on good ones, probably because bad translations tend to be more homogeneous, while good translations contain more creative and non-standard decisions which may cause disputes.

Error-analysis based TQA can lack reliability if the raters stick to different principles of language use and evaluation of mistakes seriousness. The general trend in error annotation is towards greater number of language errors, although their ratio to content-related mistakes seems to be constant – 0.6. The raters more often agree than disagree on whether a certain translation variant can be described as an error (the agreement averages at two-thirds of all mistakes annotations).

Provided that raters are previously trained, the use of error classification seems to provide more reliable data than holistic approaches to translation evaluation. The more reliable ("objective") data from translation error annotation are the total number of mistakes and the number of content-related mistakes, while the important qualification of mistakes seriousness in terms of the overall text quality or loss/unwanted change of communicative effect raises disagreements and is, therefore, found "more subjective" in this research.

These quantitative translation quality characteristics can be used to range translations of a group of students working on the same target under the same conditions to arrive at a fair and understandable marking grade. The approach suggested on the basis of our analysis is to range translations consecutively on the number of critical errors, number of content errors and total number of mistakes. We do not suggest definitions for any quality baselines, because they should be individual for each text, group of students and translation situation.

Apart from reliability, the application of error analysis has other important benefits. It provides a clear justification of the grade reached by the teacher which is

appreciated by most students. Even if it is not fully reliable, it raises issues for discussion in class. The results of error-analysis, if produced in a machine-readable format allow all sorts of automatic processing, useful in all aspects of translator training (from assessment to teaching material and curriculum design), as well as in translation studies research.

References

Artstein, R. & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596.

Freelon, D. G. (2010). ReCal: Intercoder Reliability Calculation as a Web Service. *International Journal of Internet Science*, *5*(1), 20–33. Retrieved from http://www.ijis.net/ijis5_1/ijis5_1_freelon.pdf

Kelly, D. (2005). *A Handbook for Translator Trainers. A Guide to Reflective Practice*. Manchester: St. Jerome Publishing.

Knyazheva, E & Pirko, E. (2013). Otsenka katchestva perevoda v rusle metodologii sistemnogo analiza [TQA and Systems Analysis Methodology]. *Journal of Voronezh State University. Linguistics and Intercultural Communication Series, 1,* 145-151.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology.* Sage Publications.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

Neubert, A. (2000). Competence in Language, in Languages, and in Translation. In Schäffner, C. & Adab, B. (Eds.). *Developing Translation Competence*. Amsterdam/Philadelphia: John Benjamins Publishing Company (pp. 3–17). Retrieved from http://www.benjamins.com/cgi-bin/t_bookview.cgi?bookid=BTL%2038

Strijbos, J.-W. & Stahl, G. (2007). Methodological Issues in Developing a Multi-dimensional Coding Procedure for Small-group Chat Communication. *Learning and Instruction*, *17*(4), 394-404.

Waddington, Ch. (2001) Should Translations be Assessed Holistically or through error analysis?. *Hermes*, *26,* 15-37. Retrieved from http://download2.hermes.asb.dk/archive/download/H26_03.pdf

Williams, M. (2009). Translation Quality Assessment. *Mutatis Mutandis*, *2*(1), 3–23.

Zwilling, M. (2009). O kriteriiakh otsenki perevoda [On Translation Quality Assessment Criteria]. In Zwilling, M. (Ed.), *O perevode i perevodtchikakh* [On Translation and Translators] (pp. 56–63). Moskva: Vostotchnaia kniga.

Appendix

Table 1. General mistakes statistics in student translations as marked by two independent raters

	Rater 1	**Rater 2**
Total number of mistakes	630	448
inc. content-related	247	165
inc. language-related	383	283
inc. marked as critical	102	30
Content- and language-related mistakes ratio	0.645	0.583
Percent of language-related mistakes to the total number	61%	63%
Number of translators' decisions marked as particularly good	4	9

Table 2. Two raters: Inter-rater agreement statistics as to the locus, type and seriousness of mistakes

	Absolute figures	**Agreement measures**
Number of mistakes marked in the same text span ("locus agreement")	343	54.4% (of Rater 1 total) 76.6% (of Rater 2 total)
inc. mistakes which were referred to the same category	276	80.5% α=0.605 (based on coded nominal data)
inc. content-related	118	42.8%
inc. language-related	158	57.2%
inc. mistakes with the same seriousness for both raters	96	34.8%
inc. crititcal	19	5.5%
inc. minor	23	6.7%

Table 3. Three raters: general statistics on translation mistakes analysis

	Rater 1	**Rater 2**	**Rater3**
Total number of mistakes	447	433	262
inc. content-related	173	172	83
inc. language-related	274	261	179
inc. marked as critical	39	50	No data
Content- and language-related mistakes ratio	0.631	0.659	0.464
Percent of language-related mistakes to the total number	61%	60%	68%
Number of translators' decisions marked as particularly good	17	30	18

Table 4. Three raters: Inter-rater agreement statistics as to the locus, type and seriousness of mistakes

	Absolute figures	**Agreement measures**		
		Rater 1	**Rater 2**	**Rater 3**
Number of mistakes marked in the same text span ("locus agreement")	109	24.38%	25.17%	41.60%
inc. mistakes which were referred to the same category	72	76.758% (average percent agreement) α=0.535 (based on coded nominal data)		
inc. content-related (of the content-related mistakes total for each expert)	38	21.96%	22.09%	45.78%
inc. language-related	34	12.40	13.02	18.99
inc. mistakes with the same seriousness for both raters	185 of 310 (for two raters)	59.68%		No data
inc. critical	20	6.5 (of all mistakes located in the same place for two raters)		No data
inc. minor	81	26.13 (of all mistakes located in the same place for two raters)		No data

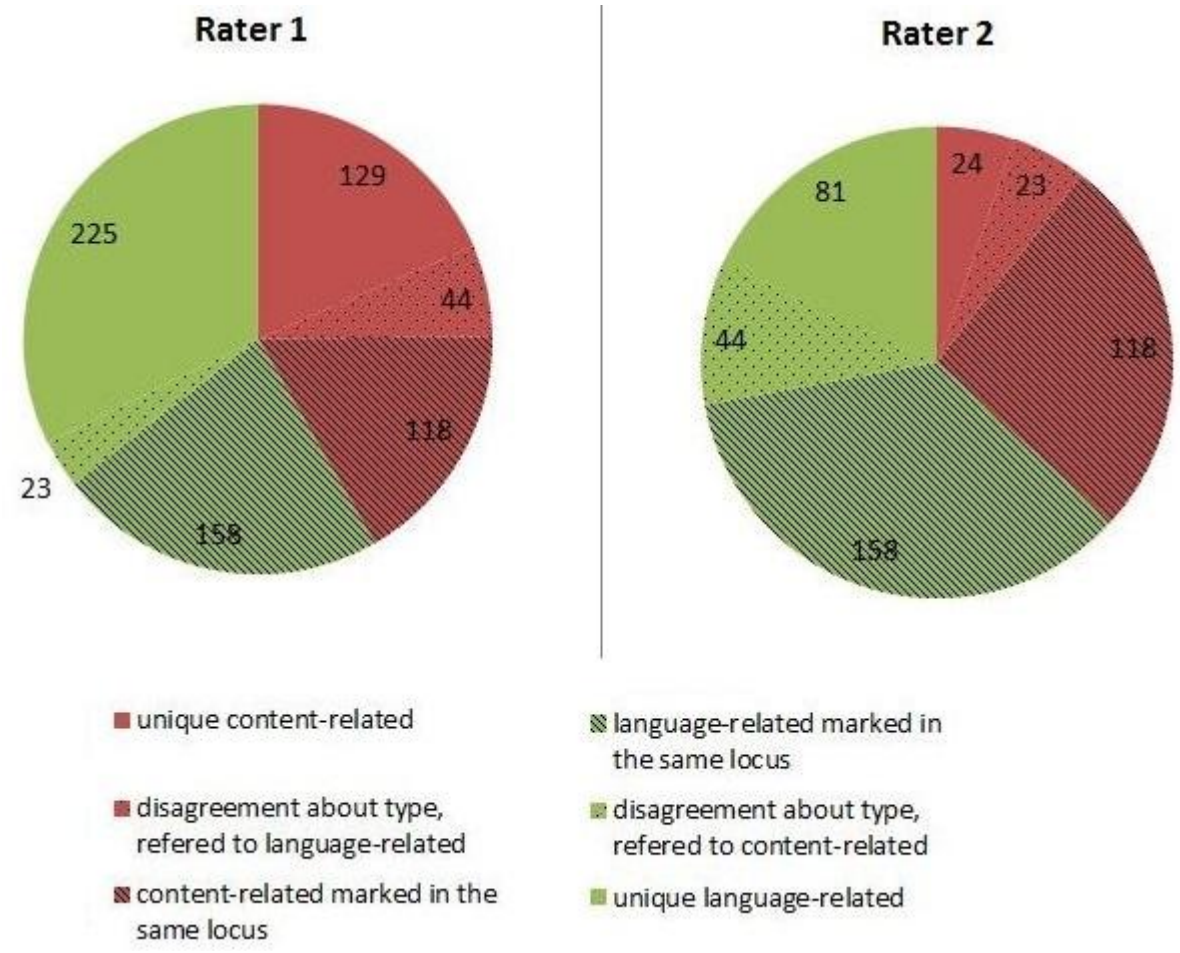

Figure 1. Raters 1 and 2: Ratio of different mistakes, including inter-rater agreement groups