*3L: The Southeast Asian Journal of English Language Studies – Vol 23(4): 251 – 264*
http://doi.org/10.17576/3L-2017-2304-19

# Evaluating Intelligibility in Human Translation and Machine Translation

*NORWATI MD YUSOF*
*Universiti Kebangsaan Malaysia*
*norwati@ukm.edu.my*

*SAADIYAH DARUS*
*Universiti Kebangsaan Malaysia*

*MOHD JUZAIDDIN AB AZIZ*
*Universiti Kebangsaan Malaysia*

ABSTRACT

*Research in automated translation mostly aims to develop translation systems to further enhance the transfer of knowledge and information. This need of transfer has brought machine translation (MT) to show major steps in translation software development and encourages further research in various MT related areas. However, there have been no focused investigations of criteria for evaluation particularly evaluation that considers human evaluators and the reconciliation of human translation (HT) and MT. Thus, focusing on two attributes for evaluation, namely Accuracy and Intelligibility, a study was conducted to investigate translation evaluation criteria for content and language transfer through reconciliation of HT and MT evaluation based on human evaluators' perception. The study focused on human evaluators' expectation of range of criteria for HT and MT under the two attributes and the evaluation was tested on a machine system to observe the system's performance in terms of Accuracy and Intelligibility. This paper reports the range of criteria to evaluate translation in terms of Intelligibility as expected by human evaluators in HT and MT in terms of content and language transfer. The study uses a mixed method approach with soft data and hard data collection. The results demonstrate that the range of each criteria identified for content evaluation in HT is expected to be higher than in MT. The implications of the study are described to provide an understanding of evaluation for human and automated translation in terms of Intelligibility.*

*Keywords: criteria; evaluation; intelligibility; human translation; machine translation*

## INTRODUCTION

Today's age of globalization has generated new translation needs in many folds. The translation industry throughout the world has crossed borders particularly to meet demands of information transfer among languages of the world. This scenario requires the practice of translation to go through evaluation in ensuring the quality of translation services given by translation providers.

Translations are evaluated every day and almost everywhere in the world. Translations are evaluated for quality by examiners and educators grading students and trainees or translation job candidates. Translations are also evaluated by employers who decide whether the translations are suitable for use or publication. Translations are even evaluated by the general public who use the translations.

Translation scholars and researchers have been working to develop the various fields in translation, be them for theory or practice. However, translation evaluation has remained the least developed. For many scholars, translation is still perceived as a "probabilistic endeavor" (Arango-Keeth & Koby 2003, p. 117). Recognising the need to analyse the notions and variables that surface in the process of judging the quality of translation, efforts have been made to focus on setting standard procedures to evaluate translations.

## TRANSLATION EVALUATION AND QUALITY ASSESSMENT

Translation evaluation and translation quality assessment has received due attention among scholars in translation studies. Among them are Robinson (1997), Schaffner (1998), House (1997), Nord (1997), Lauscher (2000, pp. 149-168), Brunette (2000, pp. 169-182), Cary and Jumpbelt (1963) and Colina (2003). However, to define Translation Evaluation and Translation Quality Assessment is not an easy task. Reiss (2000a) uses the word evaluation and criticism interchangeably as these usually go together in translator training institutions. Criticism consists of analyzing the translation and "correcting" the errors while evaluation consists of grading the quality of a translation product.

Translation quality evaluation and translation quality assessment have been described by researchers and scholars in various ways depending on the direction of the researches and the discussions. For instance, Fakharzadeh and Mahdavi (2017) reports on the results of a study that investigated the quality of the translation equivalence for mental verbs as guided by monolingual dictionaries providing a general definition of translation quality. Hasuria (1998, pp. 78-96) describes several methods in translation evaluation including proposed models by Caroll (1966, pp. 55-66), House (1981, 2001, pp. 243-257) and Reiss (1977/1989).

Susniene and Virbickaite (2012, pp. 85-90) in their study dealt with the problems of translation and definitions of the two terms. They aim to define the meaning of these terms on the basis of scientific literature, dictionaries and different documents. The meaning of the terms were classified according to their translation found in different resources and the study made the final conclusions and suggestions on their understanding. Based on their study, Susniene and Virbickaite (2012, pp. 85-90) concluded that Assessment mainly refers to "the act of assessing" while Evaluation refers to "the certainty of the value or worth" (Susniene & Virbickaite 2012, p. 89).

Although Straight (2002) indicates that the multidimensionality of the difference between two terms and the variation in each dimension results in a diverse array of examples thus the majority is neither *assessment* nor *evaluation,* he also pointed out the dimensions of the difference between Assessment and Evaluation as in Table 1.

TABLE 1. Difference between Assessment and Evaluation (Straight 2002)

| DIMENSION OF DIFFERENCE | ASSESSMENT | EVALUATION |
|---|---|---|
| Timing | Formative | Summative |
| Focus of Measurement | Process Oriented | Product Oriented |
| Relationship Between Administrator and Recipient | Reflective | Prescriptive |
| Findings | Diagnostic | Judgmental |
| Uses There of | Flexible | Fixed |
| Standard of Measurements | Absolute | Comparative |
| Relation Between objects of A/E | Cooperative | Competitive |

Based on Straight (2002), the current study uses the term Evaluation in reference to the dimensions of Timing, Focus of Measurement, Relationship Between Administrator and Recipient, Findings, Uses Thereof, Standard of Measurements, and Relation Between Objects of A/E. This is because the study develops an evaluation model by analyzing and using data and by grading the translation product of a machine system to make decisions about an evaluation model and about improvements in the selected system.

Translation evaluation may be performed on the many aspects of translation including evaluating translation materials, which involves a translation product, or evaluating translation processes using Think-Aloud Protocol or evaluating translation training. The focus of the current study is evaluating translation as a product.

The attribute considered in this discussion is Intelligibility which can be defined as the 'clarity' or 'the ease with which a reader can understand the translation' (Hutchins & Somers 1992).

## EVALUATION METHODS

In MT, various methods for evaluation have been employed. However, high quality MT remains a difficult, challenging and complex task not only because it involves many factors but also because measuring translation performance itself is difficult. Despite some 50 years of research in MT, there is still no generally accepted methodology for the evaluation of translation systems (Hutchins & Somers 1992, p. 161), (Hovy, Margaret & Andrei 2002, pp. 43-75).

Although the ALPAC report in 1965 included some evaluations on the systems that exist at that time, it is only since the initial assessments of the Systran system for the European Communities in the late 1970s that the topic had received much attention but most evaluations took place under contract and often under confidentiality agreements. Consequently there is little constructive criticism of methodology and the major deficiency is that many evaluations were undertaken by those with little or no expertise in MT techniques.

Most importantly, Hutchins and Somers (1992) suggest that in view of the misconceptions and misunderstandings concerning nearly all aspects of MT, one role of evaluation must be to introduce realism in public discussions of what MT systems can and cannot do and what they may be able to do in the future. Using the method of using human to evaluate translation product in the current study is in accordance to Hutchins and Somers (1992, p. 163) who stated that human input is required to ensure that a MT system can produce acceptable translations. Thus translator evaluation is important as their instinct is to revise MT output to a quality expected from human translators.

Kuhn and Isabelle (2009) and Melby (2014) describe a variety of evaluation methods that have been used in MT evaluation which is currently a very active field of research, and a hotly debated issue. They proposed ways of evaluating MT quality including asking human annotators to judge, or comparing the similarity of the output of a MT system with translations generated by human translators or considering how much machine-translated output helps people to accomplish a task.

Since manual evaluation uses too broad a standard to measure Correctness, Kuhn and Isabelle (2009) suggest a more common approach using a graded scale when eliciting judgments from human evaluators. The two criteria proposed are Fluency and Adequacy. Evaluation reports on the Accuracy of MT systems have shown relatively increasing interest to judge system quality in relation to the increased number of automated translation systems developed in many language pairs.

Kuhn and Isabelle (2009) in discussing the current research landscape in MT with focus on Statistical Machine Translation (SMT) as opposed to rule-based MT, describe evaluation of MT systems in two methods:

1. Automatic evaluation methods
2. Human evaluation methods.

Various kinds of human evaluations are carried out depending on how the MT system is being used which include fluency assessments and productivity measurements. Since a much higher translation quality is expected from MT systems to ensure user acceptance, evaluation of output by the systems also begins to gain attention.

Studies on output of MT that take some inspiration of human intervention of a MT system can provide an insight into translation studies perspective. They include Post Editing and Eye Tracking approaches which have been used to evaluate the output of a MT system (Doherty & Kenny 2014, pp. 299-315, Carbonell & Tomita 2003, Tomita 1992, Tomita et al. 1993). Despite its significant contribution to MT, the final stage of translation software evaluation which involves the recipients or the end-users is yet to receive attention from the perspectives of scholars and researchers in the translation discipline itself with views of translation models and evaluation attributes. To observe translation evaluation from such a perspective, the current study has attempted to fill this gap.

Based on previous studies, the current research used human evaluators to evaluate a MT output. The specific goals were to find the criteria for evaluation in both HT and MT and then to establish an evaluation model which was next tested on a MT system to observe its performance in terms of Accuracy and Intelligibility. For this paper, the discussion focuses on human evaluators' perception on the requirement of translation criteria and their range in order to evaluate Intelligibility in HT and in MT.

## METHOD OF THE STUDY

According to Creswell (2009, p. 206) mixed methods procedures may conduct data collection at the same time (concurrently or simultaneously) or in phases (sequentially). Adapting from Morse (1991, pp. 120-123), Tashakkori and Teddlie (1998), and Creswell and Plano Clark (2007), Creswell (2009) suggests two forms of data collection. The first form, which is the concurrent or simultaneous data collection, means both quantitative and qualitative data are collected at the same time. The second form, which is the sequential data collection, means one form (e.g. qualitative data) is built on the other (eg: quantitative data).

The current research has two phases where the research first applied the concurrent method. The findings from the concurrent data collection is followed by the sequential data collection as the main frame of the research design which is the second form as in the method proposed by Creswell (1999). Creswell (2009) proposed a set of three types of strategies under Sequential Designs, namely Sequential Explanatory Design, Sequential Exploratory Design and Sequential Transformative Design.

In its larger picture, the study has adopted the Sequential Exploratory Design strategy. This strategy involves the first phase of qualitative data collection and analysis which is then followed by the second phase of quantitative data collection and analysis. The qualitative data collection was carried out by conducting interviews followed by the quantitative data collection through questionnaire survey. This means that the results of the first phase which is the qualitative data collection enabled the study to conduct the second phase data collection which was carried out using a research instrument that was built based on the findings of the first phase.

Figure 1 describes the research design of the study. Interviews and text data collection were carried out concurrently in the first phase followed by questionnaire survey conducted in sequence after the qualitative data collection phase.

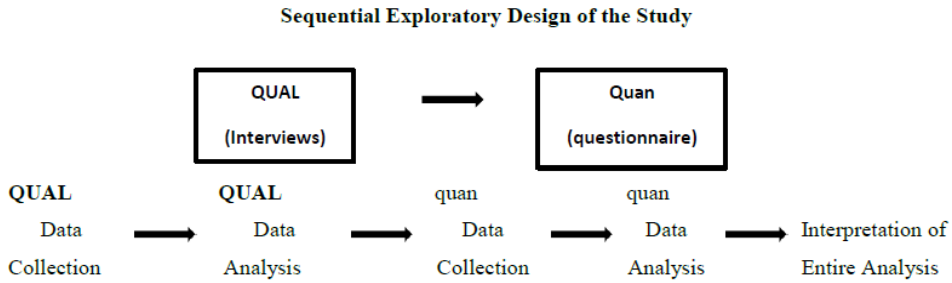


FIGURE 1. Sequential Exploratory Design of the study

However, in the first phase of the sequential design, although qualitative data collection was conducted in the current study, concurrent design was also applied during this phase. Apart from interviews for qualitative data collection, at this phase this study concurrently conducted text data collection to incorporate its result and the results of the interviews into the second phase of the sequential design, where quantitative data was collected.

## DATA

The data in this study were based on: (1) The interviews, (2) Sentence Pair Equivalences from Text Data Corpus, and, (3) The questionnaire survey. The first two types of samplings involved the concurrent phase of the first stage of data collection while the third type of sampling involved the second phase data collection in sequence after the first stage.

## PROCEDURE

The research carried out soft method data collection by conducting interview sessions with 10 professional translators who were educators and translation practitioners. Two were interviewed in the pilot study to test the Interview Protocol. The questions in the pilot study did not consider the aspects of language and content transfer. However, it was found that the two interviewees mentioned about these two aspects of transfer. Thus, in the actual interviews, the aspects of content and language transfer were considered to elicit responses for both aspects.

The Interview Protocol was prepared based on literature review. This was carried out first by identifying and selecting the criteria for evaluation of human and machine translation. They were included as a guideline to the researcher to discuss in the interview sessions to elicit evaluation criteria from the interviewees. The interviews were recorded and then the verbal recordings were transcribed and used as a written data reference and the content of the interview was categorised in several sections according to the result of the interview. The categories were the criteria for evaluation in HT and MT under two attributes; namely, Accuracy and Intelligibility.

The Text Data Corpus was initiated concurrent with the interviews. The corpus enabled the selection of test sentence equivalences or the Sentence Pair Equivalences incorporated into the questionnaire. Some of the articles in the corpus were selected from the Internet while some were from print materials. The print materials were captured and saved as PDF files. The collection of the raw data converted from the pdf files had to be cleaned from errors that occurred during conversion from PDF files to Microsoft Word documents.

The primary data was collected from the respondents using the questionnaire which was used to elicit the range of criteria for evaluation which was analysed and synthesised into a model. The questionnaire was not piloted as the content was developed based on the

interview results, which was already piloted. The questionnaire also contained sentences from the Text Data Corpus. The questionnaire was distributed via emails as well as by hand.

## RESULTS

To investigate Intelligibility in HT and MT, the criteria selected based on data collection, analysis and synthesis of the findings are Comprehensibility, Coherence and Wellformedness. Also based on the result of the analysis from the interview data, both content and language aspects were considered for Comprehensibility and Coherence, but only the content aspect is considered for Wellformedness.

For Comprehensibility, 2 statements were prepared to identify the range of mean for content and they were labelled as Question S2A23 and Question S2A24, as depicted in Table 2 and another 2 statements were prepared for language and labelled S2A25 and S2A26 as depicted in Table 3. For Coherence, 2 statements were prepared for content which were labelled S2A27 and S2A28 as depicted in Table 4 and another 2 statements were prepared for language and labelled S2A29 and S2A30 as depicted in Table 5. Meanwhile, for Wellformedness, 5 statements were prepared in terms of content which were labelled S2A31 until S2A35 as depicted in Table 6.

The first criteria under Intelligibility is Comprehensibility for which the study takes into considerations of both content and language aspects. There were 2 statements concerning this criteria and the result depicted from statement S2A23 is ranged as mean 3.62 expected in HT, while for MT it is ranged 3.26 with the difference of 0.36 in mean value. Statement S2A24 has mean value 3.40 for HT and mean 2.96 with the difference of 0.44 in mean value. Although the range for Comprehensibility is not as high as in Correctness, the results still indicates the range to be above 3.0 for this criteria in HT and only slightly below this point for MT. This indicates that this criteria is also regarded important for evaluation in both human and machine translation for content. If the observation on this criteria in the aspect of language also scores above 3.0 mean point, the figures will enable the discussion to indicate the range of importance of this criteria.

TABLE 2. Comprehensibility in Terms of Content between HT and MT

| Question | Statement | Criteria | Mean | | Difference in Mean Value |
|---|---|---|---|---|---|
| | | | Human | Machine | |
| S2A23 | The content of a translation as a whole should be easy to understand. | 4A | 3.62 | 3.26 | 0.36 |
| S2A24 | In a translation, valid information and inferences should be able to be drawn from the target text. | 4A | 3.40 | 2.96 | 0.44 |

To further observe the result on Comprehensibility in terms of content as depicted in Table 2, the means are indicated in the form of line graph. This enables the discussion to further compare the range expected in HT and in MT. Figure 2 depicts the range of Comprehensibility in terms of content to compare between expectation for this criteria in HT and MT. The two points for range of expected Comprehensibility range in human translation are connected by the line in blue while the range expected for this criteria in MT is represented in red in Figure 2.

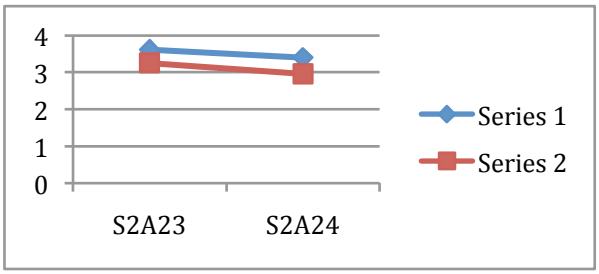


FIGURE 2. Range of Means for **Comprehensibility in Content** between HT and MT

Figure 2 reveals 2 points of mean values for HT and MT respectively. It shows that for both HT and MT, Comprehensibility criteria in HT is expected to be higher than in MT. This is indicated by the blue line being above the red line apart from the blue line which represents HT is above the red line that represents MT.

The graph also depicts that the mean values for this criteria in HT are above 3.0 thus indicating that this criteria is very strong and is important for evaluation of HT. Although the point representing question S2A24 in the red line is below this point, the value is 2.96 which is only slightly below 3.0 mean value for machine translation. Thus, the range for MT evaluation is considerably high referring to the importance of Comprehensibility as a criteria in content aspect.

For Comprehensibility, the aspect of language is also a focus. The results show that question S2A25 is ranged very high at 3.72 in mean value for HT, while the mean value for MT is ranged at 3.11 with the difference of 0.60 in mean value. Question S2A26 has mean value of 3.77 for HT and mean 3.15 for MT with the difference of 0.62 in mean value. The result indicates that the high mean value range for Comprehensibility in terms of language is very high for both human and machine translation. In fact, this result is interesting as it also shows that the range is even higher than the expectation of similar criteria under content. With this value, it shows that this criteria is also strong and indicates an important criteria for evaluation of both human and machine translation.

TABLE 3. Comprehensibility in Terms of Language between HT and MT

| Question | Statement | Criteria | Mean | | Difference in Mean Value |
|---|---|---|---|---|---|
| | | | Human | Machine | |
| S2A25 | A translation should be clear and not confusing to the target readers. | 4B | 3.72 | 3.11 | 0.61 |
| S2A26 | A translation should contain sentences that read naturally in the target language. | 4B | 3.77 | 3.15 | 0.62 |

To further compare the range of Comprehensibility expected in HT and in MT in terms of language, the result is depicted in a line graph. Established from the means of all the results, the points in the line graph were used to indicate the range of the criteria for both human and machine translation. Figure 3 depicts the range of Comprehensibility to compare between expectation of these criteria in HT and MT. The points for range of expected acceptability in HT are connected by the line in blue while the range in MT expected for this criteria is represented in red.

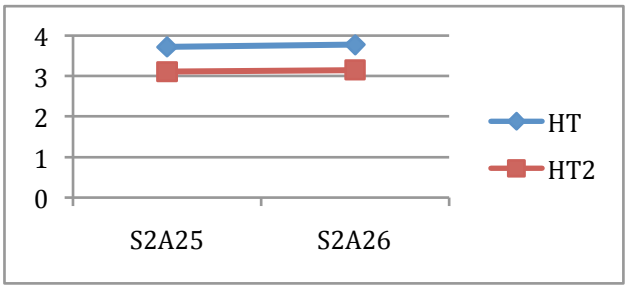


FIGURE 3. Range of Means for Comprehensibility in Language Between HT and MT

Similar to previous results, this reveals 2 points of mean values for HT and MT respectively. It shows very high mean values for HT which are above 3.5 and although the line in red that represents MT is below it, indicating that expectation of MT is lower, the mean points are still above 3.0 mean value. With the strong mean values, the result also indicates that this criteria is very important for evaluation of both HT and MT.

The next criteria under Intelligibility is Coherence. The result is shown in Table 4 for the range of mean values for Coherence in terms of content. Question S2A27 and question S2A28 indicate the result for expectation in HT with mean values of 3.62 and 3.57 respectively while for expectation in MT the mean values are 2.98 and 2.91 respectively. The difference in mean is 0.64 for the first question and 0.66 for the second question.

TABLE 4. Coherence in Terms of Content

| Question | Statement | Criteria | Mean | | Difference in Mean Value |
|---|---|---|---|---|---|
| | | | Human | Machine | |
| S2A27 | The words used in a target text should naturally connect with one another. | 5A | 3.62 | 2.98 | 0.64 |
| S2A28 | The ideas in a target text should naturally connect with one another. | 5A | 3.57 | 2.91 | 0.66 |

The results is further used to establish a line graph as shown in Figure 4 to enable observations in comparing between the expectations of Coherence in terms of content.

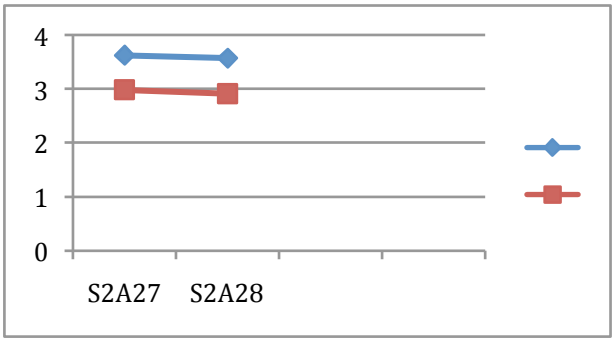


FIGURE 4. Range of Means for Coherence in Content between HT and MT

Figure 3 reveals the mean values of expected Coherence as the criteria for evaluation in HT and MT. For both HT and MT, apart from the blue line being above the red line, they also do not intersect. This indicates that the expectation in HT is higher than that of MT.

Coherence is also investigated in terms of language in this study. The result in Table 5 reveals that the mean for the questions that represent the evaluation for this criteria are at 3.60 and 3.47 for HT and quite low for MT at the range of 2.87 and 2.94 with mean difference of

above 0.5 compared to the results for MT which are at 2.87 and 2.94. The differences in mean values between human and machine translation are considerably high.

TABLE 5. Coherence in Terms of Language

| Question | Statement | Criteria | Mean | | Difference in Mean Value |
|---|---|---|---|---|---|
| | | | Human | Machine | |
| S2A29 | A translation should read fluently in the target text language. | 5B | 3.60 | 2.87 | 0.73 |
| S2A30 | A translation should contain appropriate connecting units in the target language. | 5B | 3.47 | 2.94 | 0.53 |

Figure 5 shows mean values for HT and MT respectively. The means for HT range above 3.0 of mean value and for MT the range is quite high although the range is slightly below 3.0 which again indicates that the expectation of HT is higher than that of the MT.

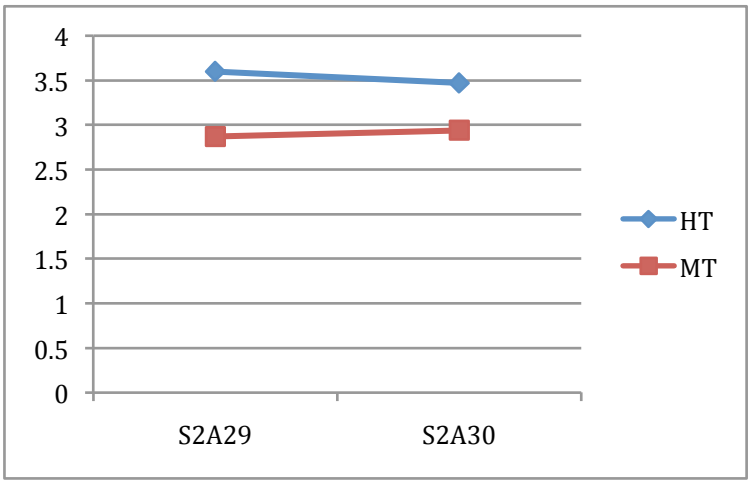


FIGURE 5. Range of Means for **Coherence In Content** Between HT and MT

Wellformedness is another criteria under Intelligibility to evaluate translation and for this criteria only the aspect of content is observed as Wellformedness. According to the result of the interviews Wellformedness is not discussed in terms of language. Table 6 depicts the result of the expectation of human evaluators on Wellformedness as a criteria for evaluation. The result from the table is also used to further establish the line graph to show the comparison between the range of the criteria expected in HT against MT.

TABLE 6. Wellformedness in Terms of Language

| Question | Statement | Criteria | Mean | | Difference in Mean Value |
|---|---|---|---|---|---|
| | | | Human | Machine | |
| S2A31 | A translation should respect the grammar rules of the target language. | 6A | 3.74 | 3.17 | 0.57 |
| S2A32 | A translation should use correct word order of the target language. | 6A | 3.66 | 3.13 | 0.53 |
| S2A33 | A translation should not contain punctuations errors which affect the meaning of the target text. | 6A | 3.53 | 3.13 | 0.40 |
| S2A34 | A translation should respect the target language in terms of sentence structures | 6B | 3.62 | 3.02 | 0.66 |
| S2A35 | A translation should render appropriate writing style of the target text. | 6B | 3.66 | 2.94 | 0.72 |

Five statements were used for this section, one each for the questions labelled Questions S2A31 to S2A35. Analysis of the data based on Table 6 shows that the result for the question labelled as S2A31 is ranged as mean 3.74 expected in HT, compared to the result for MT which is ranged as mean 3.17 with the difference of 0.57 in mean value. Meanwhile, question S2A32 has a mean of 3.66 for human translation expectation and for machine translation the mean is 3.13 with the difference of 0.53 in mean value. For question S2A33 the mean value for HT is 3.53 and 3.13 for MT with 0.40 mean value difference. Question S2A34 has a mean value of 3.62 for HT compared to 3.02 for MT with a difference of 0.72 of mean value.

To see the comparison of the range of criteria between HT and MT, Figure 6 depicts the differences in all five statements tabulated in Table 6. As in previous results, the range of criteria for HT which is indicated by the blue line is also higher in HT with the highest mean value of 3.74 and lowest is 3.53. For MT, one of the means indicates lower than 3.0, however, the value is just slightly below this point where the value is at 2.94 indicating that Wellformedness in the form of content is an important criteria required for evaluation.

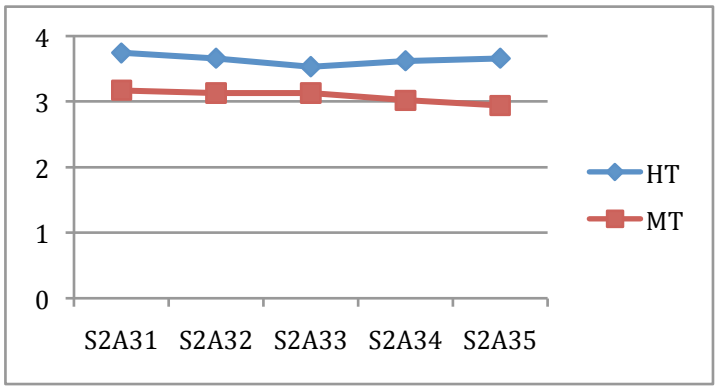


FIGURE 6. Range of Means for Wellformedness in Content Between HT and MT

## DISCUSSIONS

The discussion describes the investigation of the study on the range of criteria for evaluation in human and machine translation under Intelligibility. The range of the three criteria namely Comprehensibility, Coherence and Wellformedness will be analysed in terms of Content and Language aspects. Table 7 depicts the analysis on the average of the total means of Comprehensibility for both HT and MT in terms of content.

TABLE 7. The average of the total means of Comprehensibility in terms of Content for both HT and MT

| Question | Criteria: COMPREHENSIBILITY in terms of content | Mean | | Difference in Average of Mean Values |
|---|---|---|---|---|
| | | Human | Machine | |
| S2A23 | 4A | 3.62 | 3.26 | |
| S2A24 | 4A | 3.40 | 2.96 | |
| TOTAL OF MEANS | | 7.02 | 6.22 | |
| AVERAGE OF MEANS | | 3.51 | 3.11 | 0.40 |

The total of means under Comprehensibility for HT is 7.02 while the total for MT is 6.22 in value. Thus the average of means for Comprehensibility expected in HT is 3.51 as a shown in Table 7 while for MT the average of means is 3.11 in value. In comparing between human and machine translation the difference is 0.40 average mean value.

Similar to the analysis for Comprehensibility in terms of content, data was also tabulated to observe the average of means for Comprehensibility in the aspect of language. This is tabulated in Table 8.

TABLE 8. The average of the total means of Comprehensibility in terms of Language for both HT and MT

| Question | Criteria: COMPREHENSIBILITY in terms of language | Mean | | Difference in Average of Mean Values |
|---|---|---|---|---|
| | | Human | Machine | |
| S2A25 | 4B | 3.72 | 3.11 | |
| S2A26 | 4B | 3.77 | 3.15 | |
| TOTAL OF MEANS | | 7.49 | 6.26 | |
| AVERAGE OF MEANS | | 3.74 | 3.13 | 0.61 |

In terms of language, HT has a mean total of 7.49 mean value with an average mean of 3.74 while for MT, the total mean is 6.26 with an average mean at 3.13 in value. Although the average mean for both human and machine translation are high where both are above 3.0 value, the difference of the average is 0.60 which is considerably higher compared to previous results.

The next criteria under Intelligibility is Coherence. Table 9 shows the average of the total means of Coherence for both HT and MT in terms of content.

TABLE 9. The average of the total means of Coherence in terms of Content for both HT and MT

| Question | Criteria: COHERENCE In terms of content | Mean | | Difference in Average of Mean Values |
|---|---|---|---|---|
| | | Human | Machine | |
| S2A27 | 5A | 3.62 | 2.98 | |
| S2A28 | 5A | 3.57 | 2.91 | |
| TOTAL OF MEANS | | 7.19 | 5.89 | |
| AVERAGE OF MEANS | | 3.60 | 2.95 | 0.65 |

The total of the mean tabulated for Coherence in terms of Content expected in HT is 7.19 with an average mean of 3.60 and in MT the total is 5.89 in value. With the average mean for HT at 3.60 and for MT at 2.95, the difference in value is 0.65 which is above 0.5 value.

Analysis on the average of means for Coherence in the aspect of language is shown in Table 10.

TABLE 10. The average of the total means of Coherence in terms of Language for both HT and MT

| Question | Criteria: COHERENCE In terms of language | Mean | | Difference in Average of Mean Values |
|---|---|---|---|---|
| | | Human | Machine | |
| S2A29 | 5B | 3.60 | 2.87 | |
| S2A30 | 5B | 3.47 | 2.94 | |
| TOTAL OF MEANS | | 7.07 | 5.81 | |
| AVERAGE OF MEANS | | 3.54 | 2.90 | 0.64 |

Similar to the result for the average total means of Coherence in terms of Content, the results for language also has a high value of difference between human and machine translation. The difference is of 0.64 of value. With a total of mean at 7.07 value, HT has an average of means at 3.54 value while HT is 2.90 value.

The final criteria under Intelligibility is Wellformedness. Since data from the interviews does not indicate result in terms of language, the following discussion for this criterion only describes the average total means of Wellformedness in terms of Content in order to observe the difference between the results for human and machine translation.

TABLE 11. The average of the total means of Wellformedness in terms of Content for both HT and MT

| Question | Criteria: WELLFORMEDNESS In terms of content | Mean Human | Mean Machine | Difference in Average of Mean Values |
|---|---|---|---|---|
| S2A31 | 6A | 3.74 | 3.17 | |
| S2A32 | 6A | 3.66 | 3.13 | |
| S2A33 | 6A | 3.53 | 3.13 | |
| S2A34 | 6B | 3.62 | 3.02 | |
| S2A35 | 6B | 3.66 | 2.94 | |
| TOTAL OF MEANS | | 18.21 | 15.39 | |
| AVERAGE OF MEANS | | 3.64 | 3.08 | 0.56 |

Table 11 shows the average of the total means of the criteria Wellformedness for both HT and MT in terms of content. The total of means under this criteria for HT is 18.21 while the total for MT is 15.39 in mean value. Thus, the average of means for Wellformedness expected in HT is 3.64 as shown in Table 11 while for MT the average of means is 3.08 with a difference of 0.56 average mean value for Wellformedness.

Based on the result discussed in the above, it can be concluded that among the 3 content transfer criteria of Intelligibility, for HT, Wellformedness is the highest criteria expected to achieve in HT, followed by Coherence and Comprehensibility. However, for MT, Comprehensibility is the highest expected while the ranking for the other two criteria are similar in MT as in HT. In terms of language, for HT, Comprehensibility is found to be a more important criteria than Coherence, and interestingly, similar results is found in MT where Coherence in translation is regarded as secondary to Comprehensibility.

## IMPLICATIONS OF THE STUDY

The study has identified the criteria for Intelligibility in evaluation of human and machine translation. The range of criteria for evaluation in terms of content and language transfer has been described to see the similarities and differences in the expectation of human evaluators in evaluating HT and MT.

Several implications can be concluded from the findings of this study. First, the result provides three criteria for evaluation in terms of content and language transfer in human and machine translation. Studies conducted by other researchers may suggest different criteria and the criteria selected in the current study will provide a different perspective to include Comprehensibility, Coherence and Wellformedness in evaluating translation and MT.

Secondly, the range of criteria according to this study also depicts the similarities and differences in the range required for human and MT. It compares the range of criteria evaluation between HT and MT. Future studies to compare between human and machine translation may consider the range proposed in the current study as a guideline and reference, along with other studies to strengthen the methodology and theoretical framework in studying evaluation in either HT or MT or both.

## CONCLUSION

In the present globalized world, translation plays a very important role as a medium of communication. The advancing sophistication of cyber communication for knowledge and information transfer at all levels of society demands translation to work at a more vigorous rate which can be provided through automation. The quality of translation is evaluated every day and almost everywhere in the world. Although translation evaluation has remained the

least developed, efforts to set standard procedures to evaluate translations has gained attention among researchers and scholars.

The study proposes further investigations in several aspects. The most important aspect is providing definitions and details of each evaluation criteria already established in the current study. An investigation taking off from the current study on the evaluation criteria should venture further into the details and definitions of the six criteria established in the study.

Another recommendation is to collect a larger data by involving more number of participants in answering the questionnaire. The limitation faced by the current research occurred where only 50 out of 200 invited participants responded to return the questionnaires and only 47 were used as valid data. The researcher suspects several reasons for the limited participation. First is the involvement of evaluation for English to Malay text where some of the participants claimed that they were not professionals in evaluating the Malay language. Similar reason was given when potential participants claimed that they were not familiar with machine translation.

Another reason why the current study is suspected to have only a small number of respondents is because of the elaborative nature of Section 2 Part B of the questionnaire. In this section, the respondents have to evaluate 28 sentences from machine output, and each sentence is evaluated based on 6 criteria which defines the evaluation as actually having to evaluate the sentences 168 times. Apart from this, the evaluation for each sentence relies on reading 3 other sentences in English. This aspect will have to be fine-tuned in similar future researches.

Another recommendation is to investigate the strengths and the weaknesses of the evaluation model developed from the study. With appropriate application of theoretical and methodological strategies, such extended investigation will enable the current established model to be upgraded. It would also be interesting if the model is extended to test on another translation system or even to test on different language pairs other than English to Malay.

Lastly, the research suggests that for future research, more criteria can be included to test MT systems. Reconciled criteria under both HT and MT such as speed and cost may extend the current study to a higher level in the attempt to reconcile translation studies and machine translation.

## REFERENCES

Arango-Keith, F. & Koby, G. S. (2003). Translator Training evaluation and the needs of industry quality assessment. In Brian James Baer & Geoffrey S. Koby (Eds.). *Beyond the Ivory Tower: Rethinking Translation Pedagogy. Scholarly Monograph Series. Volume XII.* American Translators Association.

Brunette, L. (2000). Towards a Terminology for Translation Quality Assessment. *The Translator. Vol. 6*(2), 169-182.

Carbonell J. G. & Tomita. M. (2003). New Approaches in Machine Translation. *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* Colgate University, Hamilton, New York, August 14-16, 2003

Carroll, J. B. (1966). An Experiment in Evaluating the Quality of Translation. *Mechanical Translation and Computational Linguistics. Vol. 9*(3-4), 55-66.

Cary, E. & Jumpbelt, R.W. (Eds.) (1963). *Quality in Translation*. U.K: Pergamon Press.

Colina, S. (2003). *Translation Teaching: From Research to the classroom. A handbook for Teachers*. Boston: McGraw Hill.

Creswell J. W. (2009). *Research Designing and Conducting Mixed Methods Research: Qualitative, Quantitative, and Mixed Methods Approaches.* London / Thousand Oaks, CA: Sage Publication.

Creswell J.W. & Plano, C.V. (2007). *Research Design and Conducting Mixed Methods Research*. London/ Thousand Oaks, CA: Sage Publication.

Doherty, S. & Kenny, D. (2014). The Design and Evaluation of a Statistical Machine Translation Syllabus for Translation Students. *The Interpreter and Translator Trainer. Vol. 8*(2), 299-315.

Fakharzadeh, M. & Mahdavi, B. (2017). Full Sentence vs. Substitutable Defining Formats: A Study of Translation Equivalents. *3L: The Southeast Asian Journal of English Language Studies. Vol.* 23(2), 167-179.

Hasuria Che Omar. (1998). Model Cadangan untuk Menganalsis dan Menilai Terjemahan. *Jurnal Dewan Bahasa.* 78-96. Kuala Lumpur: Dewan Bahasa dan Pustaka.

House, J. (1997). *Translation Quality assessment: A Model Revisited*. Tubingen: Gunter Narr.

House, J. (1981). A Model for Translation Quality Assessment. 2nd Ed (1 ed. 1997) Tubingen: Narr. House

Hovy, E.H, Margaret, K. & Andrei P. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation. Vol. 17*(1), 43-75.

Kuhn, R. & Isabelle (2009). *MT: The Current Research Landscape.* Institute for Information Technology National Research Council, Canada.

Lauscher, S. (2000). Translation Quality of Assessment: Where Can Theory and Practice Meet? *The Translator. Vol. 6*(2), 149-168.

Melby, K. A. (2014). Can Translation Quality be Defined? Yes, But Not Absolutely. Mans vs. Machine-Poceeding of the XX FIT World Congress, Berlin.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research*. *Vol. 10*(1), 120-123.

Nord, C. (1997). "A Functional Typology of Translation". In Trosborg, Anna (Ed.). *Text Typology and Translation*. Amsterdam/Philadelphia: John Benjamin.

Reiss, K. (2000). "*Type, Kind and Individuality of Text: Decision Making In Translation" Venuti, Lawrence (Ed.). The Translation Studies Reader*. London/New York: Routledge.

Reiss, K. (1977/ 1989). Text- types, Translation types and Translation Assessment. In A. Chesterman (Ed.). *Reading in Translation Theory* (pp. 105 – 115). Helsinki: Finn Lectura.

Schaffner, C. (1998). *Skopos theory. Baker. M. (Ed.), Routledge Encyclopedia of Translation Studies*. London: Routledge.

Straight, H. S. (2002). The *Difference Between Assessment and Evaluation.* Retreive December 2010, from http://www2.binghamton.edu/academics/provost/document/assessment-eveluation-straight.ppt.

Susniene, D. & Virbickaite, R. (2012). Translation and Definition of Term Evaluation and Assessment. *Studies About Languages*. *Vol. 20*, 85-90.

Tashakkari, A. & Teddlie, C. (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Thousand Oaks. CA: Sage.

Tomita, M. (1992). Application of the TOEFL Test to the evaluation of Japanese- English MT. *Proceeding of the AMTA Workshop on MT Evaluation*. San Diego. CA.

Tomita, M., Masako, S., Tsutsumi, J., Matsumura, M. & Yoshikawa, Y. (1993). Evaluation of MT Systems by TOEFL. *Proceedings of the 5th International Conference on Theoretical and Methodology Issues in Machine Translation: MT in the Next Generation.* (TMI-93) 14-16 July 1993. Kyoto. Japan. 252-265.