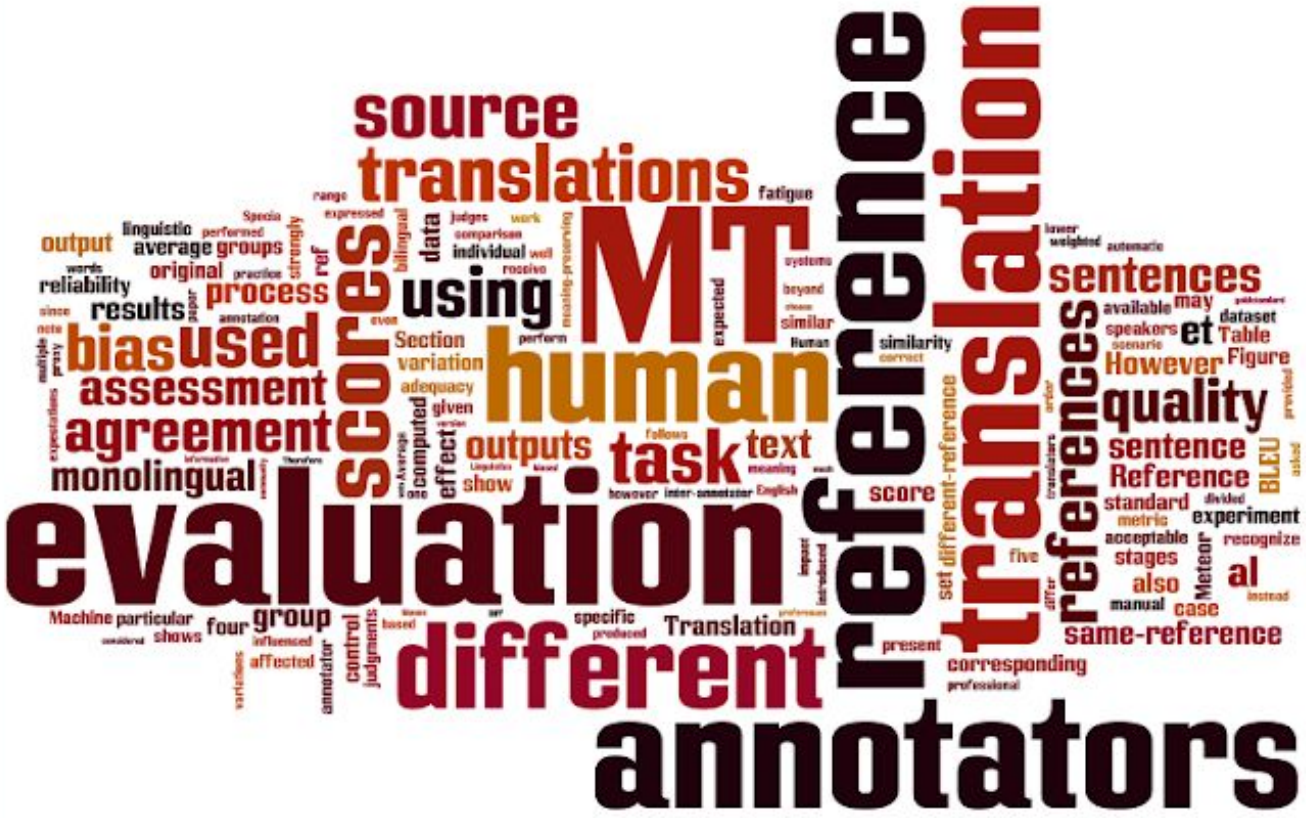
Human Evaluation of Machine Translation

Word Count: 4688



INTRODUCTION

Machine translation, commonly abbreviated MT, is a useful tool which automatically converts phrases in one language to another by a means of technological software systems. Originally introduced in the 1930's as an outlet for foreign government communication, MT started to develop as an alternative to human translators (annotators) after proving itself as undoubtedly faster and more inexpensive than human annotators in several studies (Denkowski 2015). However, the problem lies within evaluating the overall quality of machine-translated versus human-translated outputs; specifically, whether the average consumer can identify any discernible mistakes among either system's interpretation of a phrase (Denkowski 2015). Although professional annotators have already been involved in research queries requiring them to evaluate phrases translated by both humans and machines, all populations must be surveyed in order to identify any significant differences between how people of varied linguistic proficiencies evaluate human-translated versus machine-translated outputs. This study delves further into how high school students of different foreign language fluencies contrastingly assess English phrases translated into Spanish by a means of human and machine translation methods. The research question states: "How does an individual's language proficiency influence their evaluation of machine-translated versus human-translated outputs of English source sentences?"

LITERATURE REVIEW

Monolingual and Bilingual Machine Translation

In order to address the relationship between the translation quality of human annotators and translation softwares, Lucia Specia, a Professor of Language Engineering and a member of

the Natural Language Processing group at the University of Sheffield, and Marina Fomicheva, a graduate student from Pompeu Fabra University, conducted a survey to reveal how professionals with language degrees interpreted the meanings of the two types of translations (Fomicheva 2016). For their survey, Specia and Fomicheva implemented a Chinese-English dataset which was produced by the Linguistic Data Consortium. For the monolingual task, only the judgements of 20 native English-speaking annotators were collected. Of these annotators, all "were either professional translators or researchers with a degree in Computational Linguistics, English, or Translation Studies" (Fomicheva 2016). Additionally, each annotator individually evaluated the same array of 100 MT outputs, taking approximately one hour to complete the task. Furthermore, in the bilingual evaluation task, five annotators, all of whom were native speakers of English and fluent in Chinese, evaluated the same MT outputs as the monolingual annotators. Regardless of whether he or she was part of the monolingual or bilingual evaluation group, each annotator was asked to rank the MT quality on a five-point Likert scale, ranging from 1 ("None") to 5 ("All"), in order to interpret how much of the output's interpretation was expressed in the meaning of the original source sentence (Fomicheva 2016).

Likewise, in a study conducted by Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, computational linguistics researchers at the IBM T.J. Watson Research Center, automatic MT and its accuracy were studied. This method dealt with the Chinese language and involved two groups of human annotators. These annotators were separated into one monolingual group, consisting of ten native speakers of English, and the bilingual group, consisting of ten native speakers of Chinese who had lived in the United States for the past several years. Of these annotators, none possessed any professional degrees in translation or

language studies. For a total of 250 translations to evaluate, each annotator determined whether the source sentence was accurately translated on a scale from 1 ("Very Bad") to 5 ("Very Good"). While the bilingual annotators comprehensively understood the meaning of each source sentence and its five translations, the monolingual annotators determined each translation's accuracy solely by judging its fluency and readability. Interestingly, this translation software most closely generalized its sentence translation predictions with those of the monolingual annotators, meaning it did not quite meet the standards of the bilingual annotators (Al-Onaizan 1999). Additionally, each annotator's understanding of the language was taken into account when evaluating the MT output from the original source sentence. This factor was important in evaluating each annotator's ability to filter any unnecessary words which resulted from some MT outputs that a monolingual annotator may not have been able to distinguish as irrelevant (Al-Onaizan 1999).

In both studies, the researchers concluded that the MT was quite accurate in determining the accuracy of each translation in respect to its original source sentence as seen through its high correlation with the judgements of both monolingual and bilingual human annotators.

Translation Quality with Interactive Assistance and Reference Bias

Throughout several evaluations concerning the quality of MT outputs from original source sentences, various language and linguistics researchers implemented reference biases as human annotators formed their interpretations. In a study conducted by Philipp Koehn, a computer scientist and researcher in the field of MT at Johns Hopkins University, and Barry Haddow, a member of the School of Informatics at the University of Edinburgh, sought to uncover whether there were any significant variations among the evaluations of human

annotators when given assistance throughout the translation process (Koehn 2004). In this study, ten human translators, half of whom were "native speakers of French studying in an English-speaking country [and] the other half [of whom were] native speakers of English with university-level French skills" (Koehn 2004) were asked to manually translate French source sentences into English outputs. Throughout the process, each translator evaluated 194 sentences from French to English in blocks of about 40 sentences, each under five different types of assistance: "(1) unassisted, (2) post-editing machine translation output, (3) options from the translation table, (4) prediction (sentence completion), (5) options and predictions" (Koehn 2004). In each of these interactive assistance options, a positive correlation resulted when differentiated from unassisted translations (Koehn 2004). The correlations also measured the amount of pauses each human translator took between deciding the appropriate translation for each source sentence, finding that there were less frequent and shorter pauses when using interactive assistance as opposed to no assistance (Koehn 2004). This finding demonstrates that the provided interactive assistance, though it did not necessarily increase the accuracy of each human translation output, decreased the time it took each evaluator to translate the original source sentence.

Contradictory to the study described above, a separate review of MT outputs also evaluated the accuracy of human translation when aided with computer assistance versus translating independently (Langlais 2000). In this evaluation, the human annotators were asked to manually translate English source sentences into French translations. When asked if the computer-generated translation suggestions aided the human evaluators, most of them said "yes". However, when asked if the suggestions improved their typing speed, most of them replied "no".

Therefore, it can be concluded in this study that the interactive assistance in the translation process, though helpful in determining the translation of each source sentence, ultimately slowed down the overall procedure (Langlais 2000). A possible difference between this result and the results of the study above can be that translators in this study had to manually type each of the translations whereas, in the previous study, translators only had to point and click the translation which they believed was most fitting to the original source sentence.

Likewise, other researchers perceived a specific bias throughout each survey, most notably because of each translator's different backgrounds and perceptions of the language (Fomicheva 2016). Additionally, each human annotator held differing opinions concerning how to "fix" an MT output (Denkowski 2010). As a result, there is an apparent bias between each of the human annotator's interpretation of the source sentences, though likely not significant enough to cause a major disrupt in the judgement of each MT output (Fomicheva 2016). This report can also explain the difference in the two studies listed above and why each of the researchers reached different conclusions. All in all, it can be concluded that human translators generally favored the interactive assistance in the translation process, regardless of whether it was truly beneficial.

Translation Evaluation with Accordance to Time

A significant factor which several sources also took into account when each translator evaluated the source sentences was time. Particularly, surveys took human fatigue into account when evaluating the impact of reference bias (Fomicheva 2016). As a result, the reliability of the experiment was also questioned as the researchers evaluated the translation quality of each MT output (Fomicheva 2016). In order to truly evaluate the factor of human fatigue, each of the

output scores were tested in chronological portions of the data, calculating the standard deviations of the scores in each set (Fomicheva 2016). Not only was translation speed recorded, but also the accuracy of each translation as the study progressed (Koehn 2004). This translation speed was also evaluated with each of the different types of interactive assistance listed above in order to determine whether the tools were truly useful in each of the evaluation techniques. It was determined that most translators were better and faster with all of the assistances offered whereas only a few translators achieved no success with any assistance (Koehn 2004). These factors, in addition to each sentence's length, context, and novelty, all contributed to the time element in distinguishing between the quality of human translations and MT outputs (Al-Onaizan 1999).

Machine Translation Predictions of Human Evaluation

Because most of these studies developed their own translation softwares, each researcher also predicted the accuracy in the correlation between their MT system and the opinions of human translators. These predictions significantly decreased as the phrase (n-gram) length of each source sentence increased (Papineni 2002). For instance, a sentence of 1-gram length measured a precision of 0.8 whereas a sentence of 4-gram length measured a precision of 0.2 (Papineni 2002). These precisions evaluated the distinguishing factors between human and machine translations. Likewise, when directly comparing machine and human translators' measure of precision for each sentence, humans most always ranked a higher precision for each MT output, regardless of n-gram, than the predicted MT software (Papineni 2002). In a different study which also measured this output, each of the different types of interactive assistances used in the translation procedure were evaluated by a human and the same MT software (Fomicheva

2016). Again, humans tended to give a higher rank to each of the interactive assistances than was predicted by the MT software (Fomicheva 2016). Although this study was conducted fourteen years after the original study, it can be determined that MT software is yet to accurately decipher a human annotator's interpretation when translating sentences and the different types of interactive assistances they prefer.

On the other hand, when predicting the score a human annotator would give to each source sentence's MT output, the MT software performed fairly well. As previously described in the relationship between MT outputs and monolingual and bilingual annotators, when predicting the score which each monolingual annotator would give for each source sentence's MT output, the correlation coefficient was 0.99. Likewise, although slightly smaller, the same MT software produced a correlation coefficient of 0.96 when predicting the score which each bilingual annotator would give for each source sentence's MT output (Papineni 2002). Therefore, it can be determined that, though MT software is not yet ready to evaluate the different types of interactive assistance which human annotators prefer, it can fairly accurately predict how each annotator will judge a source sentence's MT output.

Overall, from the present research, it was found that both monolingual and bilingual annotators display favor to the MT outputs when provided with the option of interactive assistance. Additionally, fatigue played a large role in the variation between the annotators' responses. A common assumption within this area of research is that, however much researchers believe MT software has developed, it is not yet at the level of understanding the judgements of individuals with higher levels of language proficiency, specifically those of bilingual annotators. Therefore, the correlation coefficients between MT and bilingual speakers were usually lower

than that of the monolingual speakers; essentially, this finding indicates that the monolingual speakers agreed more with the MT outputs than the bilingual speakers did.

LIMITATIONS

This study was affected by limitations stemmed from the lack of feasible time and availability options. First, the annotators involved in this process were not experts in the language which they were analyzing and processing, unlike the professional annotators referenced throughout the foundational literature. The survey was also conducted by using an English-Spanish dataset in order to collect responses whereas the studies which this survey was based on implemented either a French-English dataset (Koehn 2009) or an English-Chinese dataset (Fomicheva 2016). Therefore, the responses cannot be completely and totally related seeing as the surveys did not process the same language. However, because of the similar survey formats, the findings from this survey may be relatively closely compared with those from the foundational literature. Lastly, and most notably, the survey designs implemented different translation software programs. Both Koehn and Haddow, and Fomicheva and Specia developed their own translation software systems for the sole purpose of their studies. However, because of a lack of resources and general knowledge, the survey for this study incorporated a previously developed and well-known translation software, Google Translate. Accordingly, the relationships of findings for all studies may be slightly skewed as a result of all these factors.

HYPOTHESIS

After careful consideration of the information provided in the current literature and the present limitations, the hypothesis for this study is that individuals with higher levels of language proficiency will associate MT outputs with lower scores and human-translated outputs with higher scores. This hypothesis stems from the findings which claim that individuals with higher language proficiencies are typically able to differentiate between the grammatical errors and formal/casual speech patterns often associated with MT outputs (Denkowski 2015).

METHODOLOGY

The goal of this study is to observe any differences and/or variations among the responses of individuals with lower versus higher levels of language proficiency. This study closely aligns with Marina Fomicheva's and Lucia Specia's study, which examines the relationship between human annotators and their interpretations of machine-translated phrases. Additionally, this study also aligns with Philipp Koehn's and Barry Haddow's study, which observes the relationship between human annotators with different levels of education of a language and their interpretations of machine-translated phrases.

Population

Participants in this study included 101 students (20 in Spanish 1, 22 in Spanish 2, 35 in Spanish 3, 16 in AP Spanish, and 8 in Spanish Speakers 2) who are, or were, enrolled in a Spanish course during some time throughout their high school educational career. These students were selected from a large, public, and linguistically diverse high school, GHC. Specifically, there are seven language classes offered on this school's campus, including Spanish, French,

Italian, Mandarin, Korean, Arabic, and American Sign Language. In order to earn their high school diploma, all GHC-enrolled students are required to take, and pass, two consecutive years of the same language class. Because the majority of students at GHC have taken, or are currently enrolled in, a Spanish class, students in different levels of Spanish classes were selected as the primary focus of the study. Additionally, the convenience of a readily available population and reduced costs to conduct the survey led to an easily accessible data collection process.

Data Collection

A stratified random sample was carried out in order to identify and distribute surveys to students who have taken, or are currently taking, a Spanish class at GHC. This sample selection was implemented in order to ensure that all students associated with a Spanish class had an equally likely chance of being chosen for the study. After additional analysis, students were stratified according to which school gate they entered through every morning, which led to a further stratification of the population into seven gates (Flagpole Gate, Zelzah Teacher Parking Lot Gate, Service Road East Gate, Hiawatha Lot Gate, Service Road West Gate, J Gate, and Kingsbury Gate). Afterwards, the percentages of students who entered these gates every day was calculated and the proportionate number of students found who were associated with a Spanish class were selected from each gate (Figure 1).

Figure 1. GHC Map and Distribution Table (Bui)

Color	Student Entry Points	# of Population SHS Students Entering
	Flagpole	1194
	Zelzah	955
	Service Road East	176
	Hiawatha Lot	261
	Service Road West	496
	J Gate	129
	Kingsbury	991
	Total	**4202**

Every morning, students were randomly selected from each gate and asked whether they were, or are, enrolled in a Spanish course on campus. In order to emphasize randomness, every third student who walked through was selected and, if the selected respondents met the survey's requirements, the survey was distributed to them if they were willing to take it. Each student's ID number was collected using Chromebooks in order to implement ease of distribution of the survey in addition to providing convenience to each student to complete it during his/her own time. Nonrespondents were also recorded. If a student had not completed the survey within one week of the day of distribution, they were considered a nonrespondent. Every respondent will be referred to as a "student annotator" throughout the remainder of this study.

Measures

This study solely collected quantitative responses for a variety of machine-translated and human-translated outputs. The survey was conducted through Google Forms and, because students were aware of the data collection process, the survey format was non-disguised. All student annotators' identification, except information on their highest completed Spanish class level, was kept anonymous. Likewise, in the Koehn and Haddow study, the annotators were required to state their level of education in terms of the language they were processing and analyzing. In order to replicate the design of Fomicheva's and Specia's, and Koehn's, and Haddow's methodology, the student annotators were classified into "monolingual" speakers (Spanish 1-3) and "bilingual" speakers (AP Spanish-Spanish Speakers 2). This classification provides a broader sense of how the monolingual and bilingual annotators contrastingly assessed each translated output.

The English-Spanish human-translated outputs were collected from a dataset cited within the Fomicheva and Specia foundational source and, afterwards, the English source sentences were manually machine-translated using Google's translation software. Each translated-output had an affiliated five-point Likert scale in which student annotators were asked to rank the translation on a scale from 1 ("Worst") to 5 ("Best") in order to measure how much of the meaning of the English source sentence was expressed through each translated output (Fomicheva 2016). A complete collection of the questionnaire and translated-outputs implemented within this survey are provided at the end of this study as "Appendix #1".

In order to prevent the student annotators from distinguishing the human-translated outputs from the machine-translated outputs and deliberately scoring one above the other, the order of translations was randomized (Figure 2).

Figure 2. Sample Screen Capture of GHC Survey

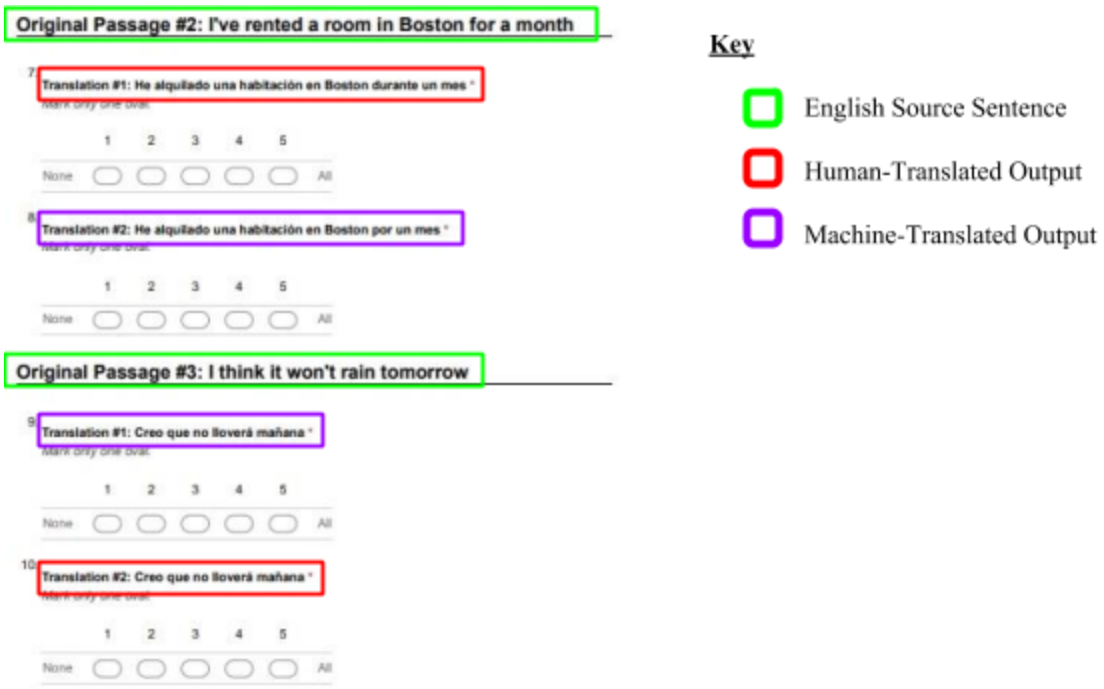


Figure 2 displays two English source sentences implemented within this survey design which have been translated twice via a human and machine. The "Key" designates the randomized order of the translated outputs displayed below the English source sentence. With this randomized order, any bias which may have been present in the annotators' responses was reduced.

Data Analysis

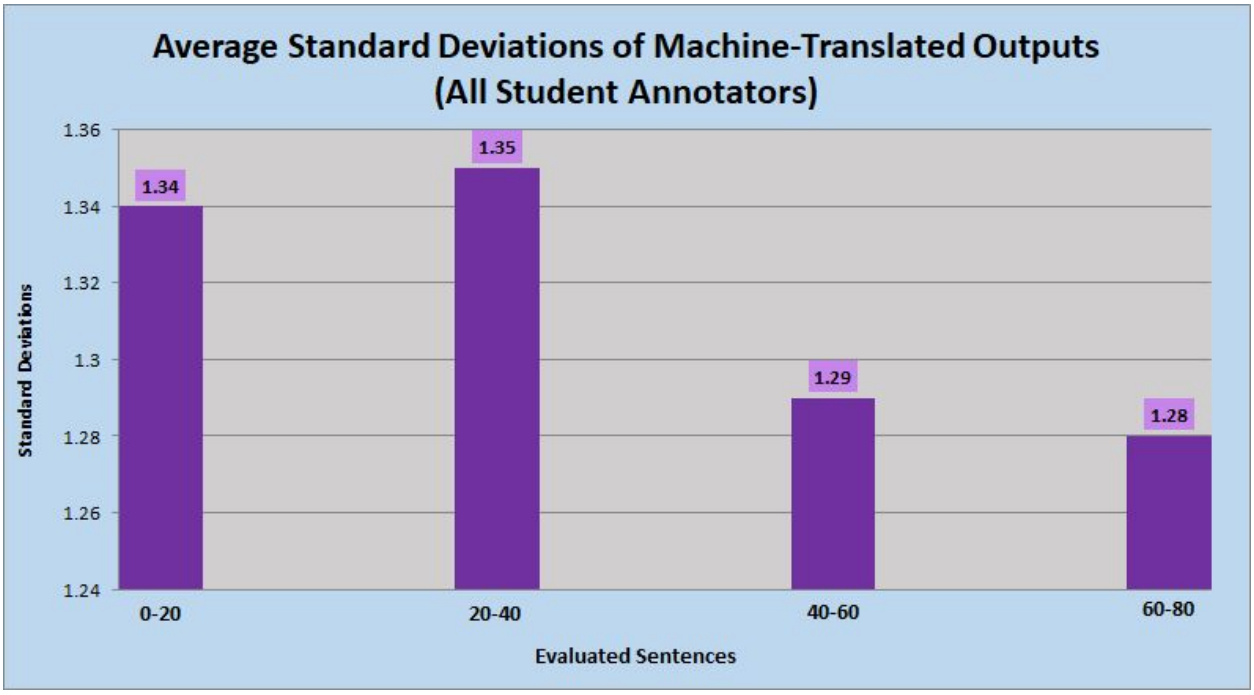
After all of the data was collected, the data was transferred from Google Forms into a Microsoft Excel Spreadsheet. The student annotators' responses were then grouped in

accordance to their level of Spanish proficiency through the use of the Excel Statistics Tool Pack. Descriptive Statistics were then implemented for every question in order to evaluate the mean scores and standard deviations for each translated output's response. Two-sample t-tests were conducted for select questions to identify any significant differences in responses between the monolingual and bilingual student annotators. Using the data, histograms were produced in order to best represent any differences between the mean responses and standard deviations for both human-and machine-translated outputs for all student annotator groups.

FINDINGS AND ANALYSIS

Standard Deviations

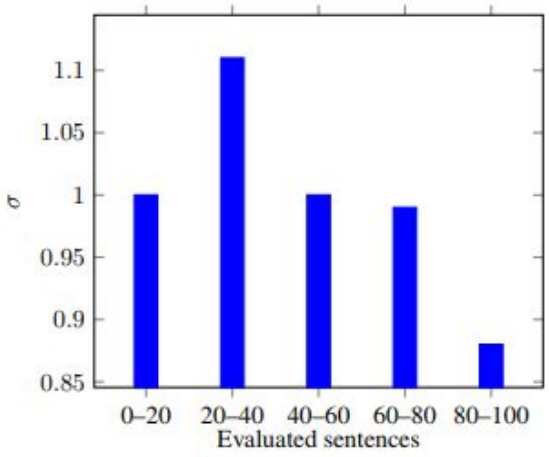
Figure 3. GHC Study - All Student Annotators' Average Standard Deviations of Machine-Translated Outputs



Overall Analysis. Figure 3 displays the average standard deviations collected from all student annotators' rankings of each English source sentence's machine-translated output. The standard deviations ranged from 1.28 (60-80 sentence group) to 1.35 (20-40 sentence group).

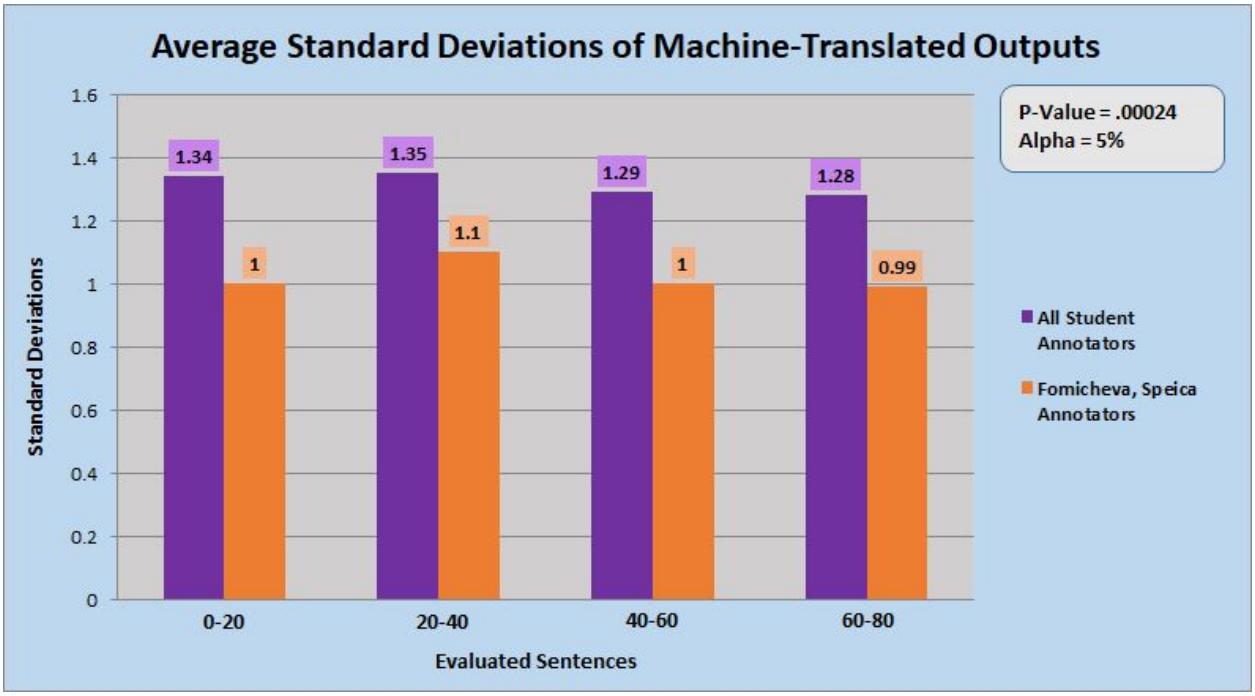
These standard deviations indicate the variation in the student annotators' responses in correspondence to the evaluated sentence group. As shown, the variation in responses has a general decreasing trend, with the exception in the increase from the 0-20 sentence group to the 20-40 sentence group. This pattern closely aligns with the standard deviations found in Fomicheva's and Specia's study, in which the standard deviations displayed a relative decreasing trend with the exception in the 20-40 "Evaluated Sentences" group (Figure 4). However, the variations in Fomicheva's and Specia's study among the professional annotators were more drastic than those found among the student annotators. These results provide evidence that both groups of annotators display relatively similar patterns in their evaluations of machine-translated outputs of English source sentences.

Figure 4. Fomicheva, Specia Study - Average Standard Deviations of Machine-Translated Outputs



Overall Analysis. Figure 4 demonstrates the average standard deviations for all machine-translated outputs in the Fomicheva and Specia study. As previously stated, the standard deviations for each sentence group gradually decrease over the interval, for the exception of the 20-40 sentence group. This finding is consistent with the machine-translated output standard deviations presented in the histogram in Figure 3.

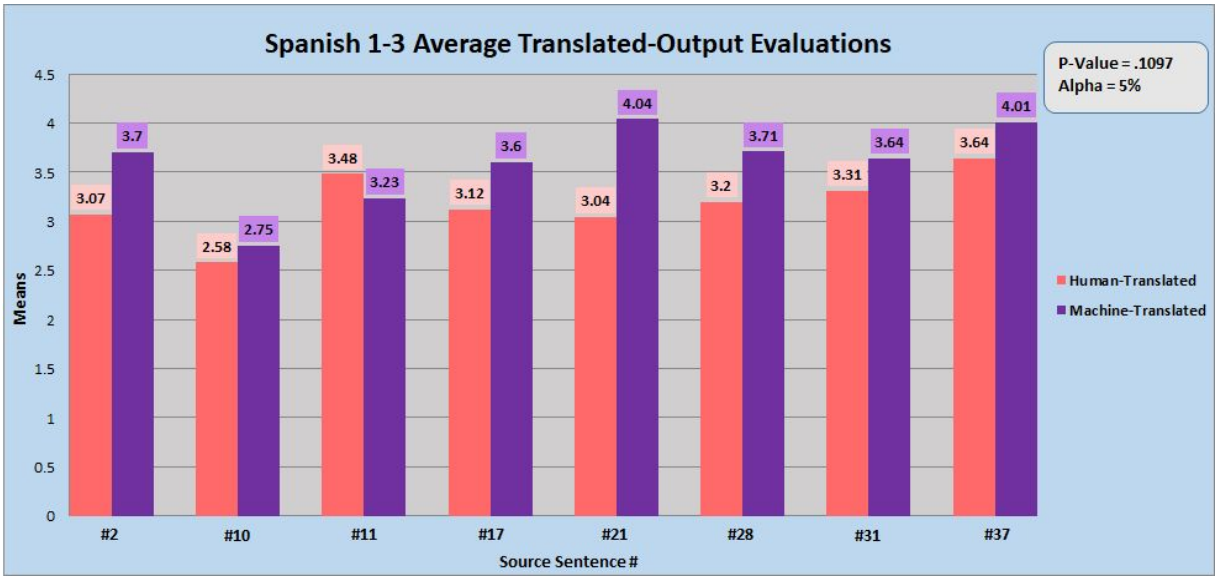
Figure 5. GHC Study vs. Fomicheva, Specia Study - Average Standard Deviations of Machine-Translated Outputs

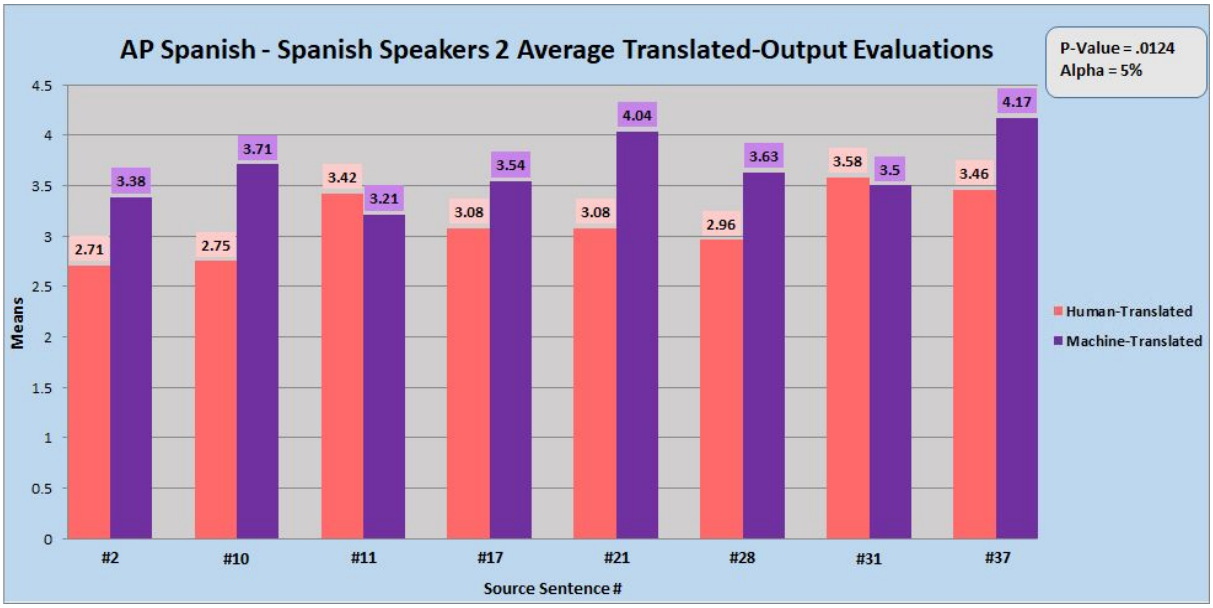


Overall Analysis. Figure 5 represents a graphical expression of the two histograms displayed in Figures 3 and 4. As shown, the two histograms individually represent a similar pattern in their annotators' average standard deviations in association to each sentence group. However, the p-value of 0.00024 indicates that these standard deviations could have occurred .024% of the time just by chance, meaning that there is enough convincing evidence to assume

that the student annotators will approximately always have greater variations between their responses than the professional annotators.

Figure 6. GHC Average Means for Monolingual and Bilingual Student Annotators in Correspondence to Source Sentence

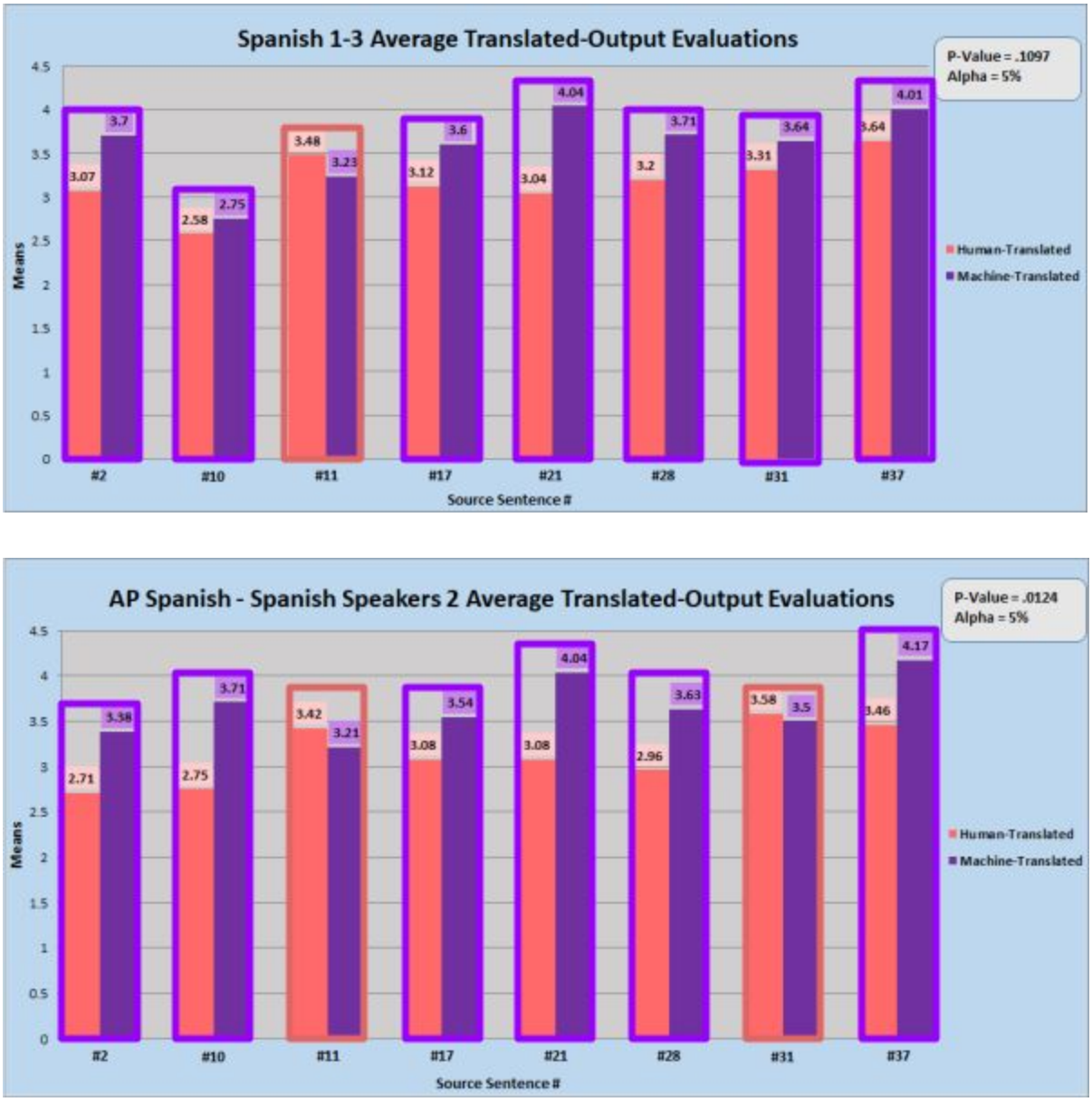




Overall Analysis. The two histograms displayed in Figure 6 indicate the mean responses of the evaluated translated-outputs for both the monolingual annotators (Spanish 1-3) and the bilingual annotators (AP Spanish - Spanish Speakers 2). Two translated outputs from each "Evaluated Sentences" group with the greatest variation in responses were chosen for evaluation for both the monolingual and bilingual speakers. These evaluations concerned both the human-translated and machine-translated outputs for each corresponding English source sentence.

As shown, the monolingual speakers' responses resulted in a p-value of .1097; essentially, this value indicates that the differences in mean responses for both the human-translated outputs and the machine-translated outputs could have occurred 10.97% of the time just by chance, revealing that this p-value is not statistically significant at the 5% significance level. On the other hand, the same test was conducted on the bilingual speakers' responses, only that this group's p-value of .0124 is statistically significant at the 5% significance level, revealing that it is highly unlikely that the variation within these responses was due to randomness.

Figure 7. General Preference of Monolingual vs. Bilingual Student Annotators



Overall Analysis. Figure 7 displays the two same histograms shown in Figure 6, only with the outlined colored bars. The pairs of bars outlined in purple indicate the source sentence in which the machine-translated output is preferred whereas the pairs of bars outline in red indicate the source sentence in which the human-translated output is preferred. As indicated, it is evident that the

machine-translated outputs are generally preferred among both the monolingual and bilingual student annotators.

DISCUSSION

Review of Findings

This study provided information which suggested that individuals with higher levels of language proficiency, as opposed to individuals with lower levels of language proficiency, contrastingly evaluate machine-translated and human-translated outputs of English source sentences. However, contrary to this study's hypothesis, both the monolingual and bilingual student annotators almost always preferred the machine-translated outputs over the human-translated outputs. The bilingual student annotators displayed a statistically significant variation in their responses between the human-translated and machine-translated outputs whereas the monolingual student annotators did not. This finding reveals that the variation in the group with higher levels of language proficiency are likely not an account of randomness and occurred as a result of a true difference between the average response rates.

On a broader scale, the average standard deviations between the responses of all student annotators and the professional annotators from Fomicheva's and Specia's research study for only machine-translated outputs were also compared using a two-sample t-test. This finding evaluated whether there were any significant differences within the variations of the average responses between the student annotators and the professional annotators. Resulting in a p-value of approximately 0, it is evident that the variation in the two groups' responses did not occur

solely by chance, but rather likely established that the student annotators will have an overall greater variation in responses than the professional annotators throughout repeated sampling.

Overall, through the findings between the monolingual and bilingual student annotators, in addition to the findings between all student annotators and professional annotators, it was shown that individuals with higher levels of language proficiency tend to significantly differentiate - while maintaining a lower variation in - their responses to human-translated versus machine-translated outputs. However, regardless of level of language proficiency, all student annotators were also found to have preferred the machine-translated outputs over 75% of the time as compared to the human-translated outputs.

Bias

Since the general methods of the foundational literature were used within the GHC study, the bias present in this study are the same as those from the foundational studies, which have undergone extensive peer review. Additionally, nonresponse bias was present within this study because the surveys were distributed only through email and many students did not respond within the allotted time period, or at all. Lack of general language proficiency and fluency may have also diverted students from completing the survey, specifically among potential monolingual annotators.

However, because every student eligible for the survey on campus has a student-affiliated email, students were able to conveniently access the survey and submit their responses at their own leisure. The survey also provided a disclaimer that all responses would be kept anonymous in order to encourage students to participate in the survey, despite if they were wary of their

decisions. Students were notified that only their highest completed level of Spanish class would be taken into account for the survey.

CONCLUSION

Significance

Despite the limitations, this study explores the importance of recognizing and establishing the differences in responses among annotators of varied language proficiencies. While researchers like Fomicheva, Specia, Koehn, and Haddow have already surveyed and identified the variations in the responses of professional annotators for machine-translated outputs, this study focused on these same factors in regards to a less-educated population consisting of high school students with different levels of the Spanish language proficiency. Additionally, this study also focused on not only the evaluation of machine-translated outputs, but also human-translated outputs of the same English source sentences in order to identify any significant differences in the annotators' responses of the translations. This study attempted to address the gap in the field of knowledge through the identification of variations among the responses of linguistically diverse annotators.

Call for Future Research

Because this survey involved only the short term collection of responses among the student annotators, a long term study should be carried out. This long term study should be carried out over the course of a monolingual annotator's progression into a bilingual annotator or a bilingual annotator's progression into a professional annotator in order to compare how one's evaluation changes over time as their language proficiency increases. This method of data

collection, though timely, would more accurately assess the differences and variations between the responses of annotators of lower and higher language proficiencies as they would be coming from the same people.

Additionally, this study should be replicated in either French or Chinese while using the English source sentences in the foundational literature in order to more closely align its findings with those presented in the foundational studies. This method could create a better, and more thorough, understanding of the present findings as they could be directly correlated with the findings in the foundational studies.

Although there are several limiting factors which were not incorporated within this study, the findings only result in further questions which concern the prospect of the implementation of machine translation in the future.

WORKS CITED

Al-Onaizan, Yaser, et al. "Statistical machine translation." *Final Report, JHU Summer Workshop*. Vol. 30. 1999.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Barrachina, Sergio, et al. "Statistical approaches to computer-assisted translation." *Computational Linguistics* 35.1 (2009): 3-28.

Bojar, Ondrej, et al. "Findings of the 2014 workshop on statistical machine translation." *Proceedings of the ninth workshop on statistical machine translation*. 2014.

Brown, Peter F., et al. "The mathematics of statistical machine translation: Parameter estimation." *Computational linguistics*19.2 (1993): 263-311.

Bui, Diane. “GHC Map and Distribution Table.” GHC Map and Distribution Table, 2017.

Callison-Burch, Chris, et al. "Findings of the 2011 workshop on statistical machine translation." *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011.

Carl, Michael, Arnt Lykke Jakobsen, and Kristian TH Jensen. "Studying Human Translation Behavior with User-activity Data." *NLPCS*. 2008.

Cohn, Trevor, and Lucia Specia. "Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2013.

Denkowski, Michael, and Alon Lavie. "Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks." AMTA, 2010.

Denkowski, Michael. *Machine translation for human translators*. Diss. Ph. D. thesis, Carnegie Mellon University, 2015.

Denkowski, Michael, and Alon Lavie. "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems." *Proceedings of the sixth workshop on statistical machine translation*. Association for Computational Linguistics, 2011.

Fomicheva, Marina, and Lucia Specia. "Reference bias in monolingual machine translation evaluation." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2016.

Germann, Ulrich, et al. "Fast and optimal decoding for machine translation." *Artificial Intelligence* 154.1-2 (2004): 127-143.

Isabelle, Pierre, et al. "Translation analysis and translation automation." *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*. IBM Press, 1993.

Karakos, Damianos, et al. "Machine translation system combination using ITG-based alignments." *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008.

Koehn, Philipp, and Barry Haddow. "Interactive assistance to human translators using statistical machine translation methods." *MT Summit XII* (2009).

Koehn, Philipp, and Christof Monz. "Manual and automatic evaluation of machine translation between european languages." *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2006.

Koehn, Philipp. "Statistical significance tests for machine translation evaluation." *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.

Langlais, Philippe, George Foster, and Guy Lapalme. "TransType: a computer-aided translation typing system." *Proceedings of the 2000 NAACL-ANLP Workshop on Embedded machine translation systems-Volume 5*. Association for Computational Linguistics, 2000.

Llorà, Xavier, et al. "Combating user fatigue in iGAs: partial ordering, support vector machines, and synthetic fitness." *Proceedings of the 7th annual conference on Genetic and evolutionary computation*. ACM, 2005.

Pal, Santanu, et al. "A neural network based approach to automatic post-editing." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2016.

Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.

Slocum, Jonathan. "A survey of machine translation: its history, current status, and future prospects." *Computational linguistics* 11.1 (1985): 1-17.

Snover, Matthew, et al. "A study of translation edit rate with targeted human annotation." *Proceedings of association for machine translation in the Americas*. Vol. 200. No. 6. 2006.

Turian, Joseph P., Luke Shea, and I. Dan Melamed. *Evaluation of machine translation and its evaluation*. NEW YORK UNIV NY, 2006.

Xia, Fei, and Michael McCord. "Improving a statistical MT system with automatically learned rewrite patterns." *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

Appendix #1
Inventory of Survey Questions

H = HUMAN-TRANSLATED OUTPUT
M = MACHINE-TRANSLATED OUTPUT
(Order of displayed translated outputs aligns with order displayed within distributed survey)

Annotators were asked to rank each translated output on a scale from 1 (“Worst”) to 5 (“Best”)

English Source Sentence	**Spanish Output**
1. Drunk driving is a serious problem	● H: Conducir ebrio es un problema serio ● M: La conducción borracha es un problema serio
2. I've rented a room in Boston for a month	● H: He alquilado una habitación en Boston durante un mes ● M: He alquilado una habitación en Boston por un mes
3. I think it won't rain tomorrow	● M: Creo que no lloverá mañana ● H: Creo que no lloverá mañana
4. Tom is the captain of the football team	● H: Tom es el capitán del equipo de fútbol ● M: Tom es el capitán del equipo de fútbol
5. Tom fell asleep while reading a book	● M: Tom se quedó dormido mientras leía un libro ● H: Tom se quedó dormido mientras leía
6. I usually have dinner at seven	● M: Normalmente ceno a las siete ● H: Habitualmente ceno a las siete
7. A book came for you in the mail today	● H: Hoy ha llegado en el correo un libro para usted ● M: Hoy le llegó un libro por correo
8. It was the most popular sport in this country	● M: Fue el deporte más popular en este país ● H: Era el deporte más popular de este país
9. Almost everyone in the class voted in favor of having a thank-you party for the teachers	● M: Casi todos en la clase votaron a favor de tener una fiesta de

	agradecimiento para los maestros ● H: Casi todos en la clase votaron a favor de tener una fiesta de agradecimiento a los profesores
10. The taxi arrived late	● H: El taxi atrasado llegó ● M: El taxi llegó tarde
11. Tom forgot to feed his dog	● M: Tom se olvidó de alimentar a su perro ● H: Tom olvidó alimentar a su perro
12. I think birthdays are important	● H: Creo que los cumpleaños son importantes ● M: Creo que los cumpleaños son importantes
13. The city was alarmed by the earthquake	● H: La ciudad estaba alarmada por el terremoto ● M: La ciudad se alarmó por el terremoto
14. Can I use my laptop in the bath?	● H: ¿Puedo usar mi portátil en la bañera? ● M: ¿Puedo usar mi laptop en el baño?
15. I wish I hadn't spent so much money	● M: Ojalá no hubiera gastado tanto dinero ● H: No debí haber gastado tanto dinero
16. You have to be there by 2:30	● H: Tienes que estar allí para las dos y media ● M: Tienes que estar allí a las 2:30
17. I go to a driving school	● M: Voy a una escuela de manejo ● H: Voy a una autoescuela
18. Tom read a poem to Mary	● M: Tom leyó un poema a María ● H: Tom le leyó un poema a Mary
19. How long can you hold your breath?	● M: ¿Cuánto tiempo puede aguantar la respiración? ● H: ¿Durante cuánto tiempo puedes aguantar la respiración?
20. I have to get up early tomorrow. Can you	● H: Tengo que levantarme pronto

give me a call at six?	mañana, ¿me llamas a las seis? ● M: Tengo que levantarme temprano mañana. ¿Me puedes llamar a las seis?
21. Tom was 13 at the time	● H: Tom tenía 13 en ese momento ● M: Tom tenía 13 años en ese momento
22. He left Japan never to come back	● M: Dejó Japón para no volver nunca más ● H: Él dejó Japón para nunca volver
23. Let's measure how tall you are	● H: Veamos cuánto mides ● M: Vamos a medir qué tan alto eres
24. Her brother is a good driver	● H: Su hermano es un buen conductor ● M: Su hermano es un buen conductor
25. An ugly man knocked on my door	● M: Un hombre feo llamó a mi puerta ● H: Un hombre feo llamó a mi puerta
26. When was the last time you took a shower?	● H: ¿Cuándo fue la última vez que te diste una ducha? ● M: ¿Cuándo fue la última vez que te duchabas?
27. I will have him repair this watch	● H: Voy a hacer que él arregle mi reloj ● M: Lo haré reparar este reloj
28. What do Japanese students usually eat for lunch?	● M: ¿Qué suelen comer los estudiantes japoneses durante el almuerzo? ● H: ¿Qué comen de almuerzo comúnmente los estudiantes japoneses?
29. Your hair smells wonderful	● M: Tu cabello huele maravilloso ● H: Tu pelo huele de maravilla
30. A cookie is under the table	● H: Hay una galleta debajo de la mesa ● M: Una galleta está debajo de la mesa
31. He is like a father to me	● M: El es como un padre para mi ● H: Él es como un padre para mí
32. He did not like to travel	● M: No le gustaba viajar ● H: A él no le gustaba viajar

33. There's a math test tomorrow	● M: Hay una prueba de matemáticas mañana ● H: Mañana hay prueba de matemáticas
34. Tom answered all the questions that Mary asked him	● M: Tom respondió a todas las preguntas que Mary le hizo ● H: Tom respondió a todas las preguntas que le hizo Mary
35. Tom can't drive a car so he always rides a bicycle	● H: Tom no puede conducir un auto, así que siempre anda en una bicicleta ● M: Tom no puede conducir un automóvil, así que siempre monta una bicicleta
36. Where are you going to eat lunch?	● M: ¿Dónde vas a almorzar? ● H: ¿A dónde vas a ir a almorzar?
37. The small boy slowly made some new friends	● H: El niño poco a poco hizo algunos nuevos amigos ● M: El niño pequeño hizo algunos nuevos amigos lentamente
38. He had no money and so could not buy any food	● H: Él no tenía dinero, y por eso no podía comprar comida ● M: No tenía dinero y no podía comprar ningún alimento
39. Spring is just around the corner	● M: La primavera está a la vuelta de la esquina ● H: La primavera llegará pronto
40. Tom spends too much time on the computer	● H: Tom pasa demasiado tiempo en el computador ● M: Tom pasa mucho tiempo en la computadora