International Journal of Research Studies in Educational Technology
2015 April, Volume 4 Number 1, 51-62

A survey on the subject-verb agreement in Google machine translation

Bozorgian, Mojtaba
Department of Language and Literature, Kerman Science and Research Branch, Islamic Azad University, Kerman, Iran (mbozorg2@gmail.com)
Azadmanesh, Nematollah
Department of Language and Literature, Kerman Science and Research Branch, Islamic Azad University, Kerman, Iran (nematollah.azadmanesh@yahoo.com)

ISSN: 2243-7738
Online ISSN: 2243-7746

OPEN ACCESS

Received: 22 October 2014 **Revised**: 14 December 2014 **Accepted**: 16 December 2014
Available Online: 10 January 2015 **DOI**: 10.5861/ijrset.2015.945

Abstract

To investigate subject-verb agreement of Persian translated sentences in Google machine translation, 100 sentences were taken from the BBC's English web site that 50 sentences were randomly selected to be translated by both Google Machine Translator and four Human translators, to reach a reference translation. Descriptive-statistics method was used to find out; firstly, whether Google translator compared to human translator can apply subject-verb agreement properly. Secondly, whether there is a relationship between human judgment and Machine Translation automated evaluation method. To this end, after reaching a reference translation, the researcher invited four other human translators to judge and score the Google output based on three criteria. A five-point Likert Scale was used for rating the Google translated sentences. The scale (1) was assigned for a completely unacceptable translation and (5) for an excellent one with almost no errors. Then, the researcher evaluated the same 50 Google translated sentences based on three automated metrics of Precision, Recall, and F-measure. Then, the researcher described the findings through his observation of the scores given by four human translators and found out that Google translator in comparison to the human reference translation could not apply the subject-verb agreement properly in all sentences. Then, after analyzing the "human and F-measure" assessment scores by SPSS version 22 and taking the correlation coefficient of the two scores, the researcher figured out that there was a significant relationship between human and F-measure scores.

Keywords: machine translator; Google translator; subject-verb agreement

A survey on the subject-verb agreement in Google machine translation

1. Introduction

Machine translation is the application of computers to the translation of texts from one natural language into another, as stated in Hutchins (1985) and Somers (2011). The term Machine Translation (MT) is the new traditional and standard name for computerized systems responsible for the production of translations from one natural language into another, with or without human assistance (Huchins & Somers, 1992, p. 3). Machine Translation (MT) as a one of the major research area in computational linguistics has taken attention since 1950s. From the time MT was utilized, people thought of an automatic translation of all sorts of texts in all fields with a best output quality equaling the job of a best human translator.

Machine translation neck and neck with human translation (HT) came to the realm of competition. Gradually, MT proved that it could move on faster than HT in finding equivalents for the words given, however, translation is not just switching words between languages. Regardless of speed in suggesting equivalents, MT was lag behind in giving the best choice in many cases. As Robinson (2003) puts in, "translation is not the same sort of activity as tying your shoes or brushing your teeth" (p. 50). According to him, translation is always an intelligent behavior. He further confirms that "translation is a highly complicated process requiring rapid multilayered analyses of semantic fields, syntactic structures, the sociology and psychology of reader- or listener-response, and cultural difference" (p. 50).

All in all, there are some problems regarding translation either by a human or by a machine translator. One field of problems to be concerned about is concord or agreement which itself has some subcategories. Hence, here in this research, the researcher is aimed to consider just the subject-verb agreement in the translation of English to Persian by Google machine translator. Google translate is one of the services that Google provides. Google translate is supporting almost 60 languages. Google translator had many mistakes and errors and still has, but in some cases Google translator improved in translations by the facilities it provided its users, for instance, with editing, substituting, and deleting the wrong choices it gave.

Agreement features are very important and should be carefully applied to ensure the generation of sound sentences in the target language. Since agreement applies to the target language, in this research Persian language, agreement should fulfill the specific requirements of this language. Mistakes and errors in the MT output can either be the result of analysis problems at the source language level, or due to the generation problems at the target language level. The aim of this research is to explore the implications and effects of the subject-verb agreement features in the Google Machine Translation process. The research target is to determine to what extent subject-verb agreement, as a set of features and as a set of rules, is responsible for generating coherent Persian structures in the Google Machine Translation output. Hence, the present study seeks to find out the answers to the two following research questions:

- Which one deals better with subject-verb agreement while translating English sentences into Persian, Google machine translator or human translator?
- How do scores obtained from human judgment correlate with scores obtained from MT evaluation method?

Accordingly, the researcher devised the following hypothesis:

- Google Machine Translation deals better with subject-verb agreement while translating English sentences into Persian compared to human translator.
- There is no correlation between scores obtained from human judgment and scores obtained from MT

automated evaluation method.

The findings of this study will be beneficial to the students of MA in translation studies. First, the Google machine translated sentences can be given to them to compare and contrast them with human reference translation based on subject-verb agreement rules. Then, students of MA in translation studies will learn how to evaluate translated sentences based on both manual evaluation criteria and automatic evaluation metrics.

Machine translation technology like human translation was brought to Iran due the rapid growth of needs to know about modern life and technologies. Therefore, many scholars and computational linguists in Iran and other countries devised English to Persian machine translation systems. They have already constructed several MT systems for the English/Persian language pair to meet peoples' requirement. Most of those MT systems were purely rule-based. Accordingly, there were lots of researches done by different scholars both in Iran and other countries regarding machine translation and agreement structures of translated sentences in machine translation. Consequently, different researchers voiced different opinions regarding machine translation.

Some people believe that studies of any kind around MT is useless and a futile attempt because machine translator is not capable of translating different literary tasks like those of Shakespeare, Dickens, Saadi, Hafiz and so on. However, translating literary text is not within the scope of MT, as it does not have the qualification as that of a human. Vitek (2000), as stated by Taghvaipour (2004, p, 34), criticized MT by saying, "to try to reduce human language, which is as complex as human thinking, to a series of zeroes and ones, is clearly an exercise in futility." Alternatively, "Machine translation, in spite of all noises about it and billions of yens, marks and dollars sunk into it, will never really amount to anything but a tool that can be used basically only by translators."

On the other hand, some researchers proposed some solutions and supported the act of machine translation regardless of its mistakes and errors. First, Abdel-Aal Attia (2002) in his thesis, *Implications of the Agreement Features in Machine Translation*, showed that many shortcomings in the output of MT are due to either faulty analysis of the source language text or faulty generation of the target language text. He referred to the English into Arabic MT system Al-Mutarjim Al-Arabey by ATA Software as testing ground in his research. In some cases, the MT system was successful in making the correct agreement; in other cases, it was not successful due to lack of requisite information with respect to agreement features. He also believed that enhancement to the output can only be done by formalizing linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. In his research, he further concluded that English has only ten agreement features compared to twenty-four agreement features fully utilized in Arabic.

Next, Lotfi (2006) in his article, *"Agreement in Persian*", "intended to shed light on the yet unexplored complexity of agreement phenomena in Persian" (p.124). He took number value and plurality marking mechanisms in Persian into consideration. Therefore, he proved that "the Animacy Hierarchy formulated by Forchheimer (1953), Smith-Stark (1974), Silverstein (1976), and Corbett (2000) is in need of revision in order to capture the evidence from number marking in Persian" Lotfi (2006, p. 124). In addition, he claims that without such a revision the two features of "autonomous" and "non-autonomous" were simply neglected in favor of one single term - inanimate. He further argued that it depends on the speaker what to think of a plural in animated noun, whether they perceive it as a singular or plural noun. As Lotfi (2006) concludes, "Whether the noun is interpreted collectively or not is conceptually related to this feature of autonomy, and iconically realized as SG/PL in verb morphology" (p. 124). Furthermore, Shahabi (2009) in her paper, *"An Evaluation of Output Quality of Machine Translation Program",* compared two systems of Pars and Padideh based on the criteria of "accuracy" and "intelligibility" in the translation of sentences from English into Persian. Her findings are as follow. First, Padideh translator produced the best output. Then, these two systems had problems in the areas of morphology, complex sentences, syntactic ambiguity, semantic analysis, generation of Persian, and long sentences.

In addition, Mirzaeian (2010), in his study, *"Challenges of Machine Translation in Persian, Using Three MT*

Systems", compared the translation of three machine translations namely, Pars, Padideh, and Google translator. He compared them according to their performances facing issues like Nouns (common, proper, collective), Pronouns (personal, relative, demonstrative, etc.), Verbs (intransitive, transitive, linking), Tenses, Passives, Verbal, etc. He confirmed that each of these systems had problems translating into Persian regarding these structures. He did not introduce any of them better than the others did. Similarly, Mirzaeian in his paper in 2011, "*Improving the Translation of Idioms by Google Translate"*, put Google machine translator under investigation through the translation of idioms. He randomly selected 500 idioms from Yarahmady et al. (2011) and tested them with Google translate to see how well they were translated. He further concluded that Google translator was totally unable to detect and translate English idioms into Persian.

On the other hand, Ghoreyshi and Aminzadeh (2011) in their study, *The Evaluation of Translations of Three Persian Systems of Machine Translation, Based on Catford's Shifts*, investigated three machine translators of Pars, Padideh, and Google based on Catford's (1965) shifts (structure shift, unit shift, class shift, intra-system shift) in the Persian translation of a text. They evaluated the translation of these three machine translators with respect to a translation of a human translator of the same text to see which of them translates similar to human translator regarding Catford's shift. Ghoreyshi and Aminzadeh concluded that their findings are to some extent similar to Mirzaeian's (2010) findings and their work is in confirmation of Mirzaeian's. They further attested that none of the three machine translators could translate without mistakes. They believed that the three systems have still long distance to human translation regarding grammatical and lexical accuracy.

Ahangar, Jahangiri, and Mohammadpour (2012) in their research, *"A Lexical-Functional Model for Machine Translation of English Zero-place Predicators into Persian*", utilized lexical-functional grammar for a machine translation system, which was designed for the translation of some English zero-place predicators. This lexical-Functional Grammar is able to translate English zero-place predicators into Persian as one-place predicators and with consideration of Persian word order that is more natural for people whose mother tongue is Persian (p. 1). Overall, all the above-mentioned researches were done on Machine translation, especially on statistical machine translation, but they were not merely considering the subject-verb agreement structure in the output of Google machine translator. Therefore, this study aims to focus on the problems of subject-verb agreement occurred in the translation of sentences from English to Persian in Google machine translator. In addition, the study will conduct a kind of comparison between the human judgment and an automatic machine translation evaluation with respect to F-measure scores to find a correlation coefficient between the two evaluation methods.

Before considering the research aim, it would suffice to notice some points like agreement and evaluation of machine translation. Agreement, sometimes referred to as concord, as Matthews in 1981 defined, is "a relation between words that share a morphosyntactic feature" (p. 246). A more elaborate definition from the American Heritage Dictionary of the English Language is the "correspondence in gender, number, case or person between words" (Heritage, 1996). Both human judges and automatic metrics evaluate machine translation output. The researcher devised a set of scoring procedure for human evaluation based on three criteria of Lexical Accuracy, Semantic Accuracy (Convey the same meaning as the original), and Syntactic Accuracy (subject-verb agreement). In addition, the researcher used three automatic metrics of F-measure, Precision, and Recall to score Google translated sentences. Precision is the percentage of generated words that are actually correct.

$$\text{precision} = \frac{\text{correct}}{\text{output-length}}$$

Recall stands for the percentage of words that are generated and that are actually found in the reference translation.

$$\text{recall} = \frac{\text{correct}}{\text{reference-length}}$$

"F-measure is the harmonic mean of recall and precision" (Kohen, 2010).

$$\text{f-measure} = \frac{(\text{precision})(\text{recall})}{(\text{precision} + \text{recall})/2}$$

However, the researcher ends to investigate the F-measure scores of Google translation with respect to one human reference translation.

2. Method

This is a descriptive-statistic research. The researcher is to demonstrate a comparison between human translator and Google machine translator with respect to subject-verb agreement in 50 sentences taken from the BBC's site. Then, the researcher is to determine that Google Translator is different from Human Translator concerning subject-verb agreement in Persian-translated sentences. Finally, the study will attest that there is a relationship between Human evaluators and F-measure MT evaluation method in Persian-translated sentences. Finally, it must be mentioned that the unit of translation in this study is a sentence. To do the research, the researcher was provided with some instruments including a lap top computer, an internet connection, and 50 sentences taken from BBC English site. The source language (SL) is English and the target language (TL) is Persian in this research.

In addition, eight translators have been invited to participate in the research that four of them were asked to do the job of translation and reach a reference translation for further investigations. Then, the other four translators were asked to judge the Google translated sentences based on human reference translation with respect to three assessment criteria. It is important to mention that these raters do not know the reference translators and even the raters themselves. The researcher did this intentionally to prevent raters being biased when judging the Google translation output. In addition, the researcher did not devise any instructions for human raters to judge the Google translated sentences. Furthermore, according to the fact that they are translators, it is believed that they all know about subject-verb agreement structures in both Persian and English language. Hence, the researcher left them be free in judging sentences with respect to subject-verb agreement. Nevertheless, a brief note of subject-verb agreement was given to the four raters as a reminder in this respect. This reminder is taken from Lazard (1992, pp. 178–181) table1.

Table 1

Options with subject–verb agreement in Persian

Subject	Verb
Plural animate beings (having will or feeling)	Plural
Inanimate beings (or things considered as inanimate)	Singular
Things that are conceived as endowed with a certain activity, or such that there is cause to insist on their plurality and the individuality of each item.	Plural
Animate beings which are not conceived of as the agents of the process or as affected by it.	Singular
Collective noun (human collectivity)	Most often plural
Numerals and expressions of quantity	Plural, but rarely the singular is found
Certain plurals with a collective value	Singular
Distributive expression (har kas(-i), har yeki 'each', etc.)	Singular/plural

Furthermore, it must be mentioned that the four raters' judgment and scoring result will not be compared with each other in this study. In addition, the raters' sex and ages are not of any importance in this study.

A five-point Likert Scale was used in human judgment of Google translated sentences. The scoring

procedure is as follow:

Table 2

Human Assessment Criteria

Sentence No.					
Criteria	Assessment				
Lexical Accuracy	No error	1-3 error	4-6 error	7-9 error	More
	5	4	3	2	1
Convey the same meaning of the original	Excellent	Very good	Good	Rather weak	Weak
	5	4	3	2	1
Syntactic Accuracy(Subject-verb Agreement)		Yes		No	
		5		1	
Total score					

It must be notified that the word 'No' does not contain any negative meanings. It just means that one of the subject or verb lacks the quality of having agreement with the other part. Next, the 50 output sentences of Google will be measured on three automated scales of 'Precision, Recall, and F-measure'.

2.1 Procedure of the Research

- 100 sentences are taken from BBC's web site out of which 50 sentences were randomly chosen on a random sampler site on the internet. These sentences are in English language.
- The 50 sentences are given to Google machine translator to translate them from English (SL) to Persian (TL).
- The 50 sentences are also given to four University Professors to translate them from English to Persian. Then, reaching a unique translation of the four human-translated sentences, these sentences will be used as a reference translation for the comparison between evaluation of human translators and Google machine translator.
- After reaching a reference translation, four other University Professors holding master's degree in Translation Studies are invited as raters to judge the Google translation. It must be mentioned that these raters did not do the job of translation; therefore, they remained unbiased when judging the Google translated sentences according to the human reference translation. The reason that translators were chosen to judge the translation is that translators have proficiency in both the source (English) language and the target (Persian) language.
- The researcher devises three criteria based on which four University Professors will score Google translated sentences. The four University Professors holding master's degree in Translation Studies are asked to score these 50 Google-translated sentences from 5 to 1 Likert-based scale on three criteria of Lexical accuracy, Convey the same meaning of the original text, and Syntactic accuracy regarding subject-verb agreement.
- The output sentences of Google machine translation is evaluated with three scales of 'Precision, Recall, and F-measure'.
- The correlation coefficient of two scores, F-measure and average score of professors, are analyzed by SPSS software.

2.2 Variables

The research investigates the kind of relationship between the following variables:

- F-measure
- The mean score obtained from the group of human evaluators.

The first score is the predictor variable and the second score is criterion variable which is used to verify how accurate the first score is.

2.3 Data analysis

- At first, based on the human evaluation, the researcher counts different times that human judges agreed there are subject-verb agreement. Then, we will decide on how many sentences Google could correctly translate sentences with respect to subject-verb agreement among 50 sentences as a whole. After that, the researcher counts the sentences that each human judge has agreed that the sentence has subject-verb agreement. At last, these data will be set to compare and contrast the result. The result will be described in details.
- F-measure scores and the mean scores of human evaluation are submitted to the SPSS software to estimate the correlation coefficient between these two sets of scores.

3. Results

Based on these research questions, the following main hypotheses have been formed:

- Google Machine Translation deals better with subject-verb agreement while translating English sentences into Persian compared to human translator.
- There is no correlation between scores obtained from human judgment and scores obtained from MT automated evaluation method.

Therefore, the null hypotheses, which are going to be rejected by the researcher, are: "Google Machine Translation deals better with subject-verb agreement while translating English sentences into Persian compared to human translator." In addition, "There is no correlation between scores obtained from human judgment and scores obtained from MT automated evaluation method." Having applied these three criteria for each sentence, the four human evaluators reached the different mean scores out of 15. This can be observed in table 3 below.

Table 3

Scores of Four Human Raters out of 15

	N	Sum	Mean
Rater1	50	366.00	7.3200
Rater2	50	358.00	7.1600
Rater3	50	322.00	6.4400
Rater4	50	318.00	6.3600
Valid N (list wise)	50		

According to table 4 below, the mean score out of all four raters is 6.72 out 15. In addition, for further discussion it must be notified that these four raters had done rating without the intervention of any of the other raters. In fact, they did not know each other and the researcher kept their names secret not to tip the balance during the research. Consequently, they all did their job of rating for each sentence. It can be claimed that these 50 Persian translated sentences by Google translator were rated and judged four times by four different raters

that makes the total 200 sentences in this research.

Table 4

Mean Score of Four Raters out of 15

	N	Sum	Mean
Human mean score	50	336.00	6.7200
Valid N (list wise)	50		

Therefore, there were only 33 correct sentences according to subject-verb agreement among 200 sentences scored by the four raters. This can be observed in table 4 below.

Table 5

Total Score of 4 Raters Based on Y/N

Rater	1		2		3		4	
Subject-verb agreement	Y	N	Y	N	Y	N	Y	N
Y/N	11	39	13	37	4	46	5	45
Total Sentence				200				

According to the data in table 5 above, it can be mentioned that Rater1 agreed with the subject-verb agreement of 11 out of 50 sentences he/she scored, and 39 sentences were rejected in this regard. However, Rater2 agreed with 13 sentences out of 50 sentences based on subject-verb agreement and did not agree the subject-verb agreement of 37 sentences. On the other hand, Rater3 just accepted 4 out of 50 sentences with regard to subject-verb agreement. The last but not the least is the Rater4 who did not accept 45 sentences out of 50 and just agreed 5 sentences. Therefore, the researcher intended to find the reasons why these raters agreed with some sentences' subject-verb agreement in Persian and did not agree with some others. To this end, the researcher focused on those situations that these raters were at odds with each other in respect to accepting the subject-verb agreement.

In these evaluations done by human raters, it was conspicuous that sentence 11 has scored maximum score of 15, which is the best score at the same time. Among four raters rating the sentence 11, three raters gave score 15 and one rater gave 14 to this sentence. As a result, this sentence could take the mean score of 14.75 as the highest score in this research. On the other hand, the lowest score was given to the sentences of 7, 23, 24, 46, and 50 that they all took the mean score of 4.5 from four raters in this research. This is summarized in the table 6 below, which is manual and the SPSS version is provided below.

Table 6

Max/ Min Mean Score of 4 Raters

	Sentence No	Mean score
Max mean score	11	14.75
Min mean score	7,23,24,46,50	4.5

This claim was proved through SPSS and is visible in table 7 below. In comparing the two score groups of max/min above and going further into details with respect to the subject-verb agreement, it is figured out that sentence 11 could inevitably have the best subject-verb agreement in this research that all four raters confirmed this agreement. However, none of the sentences of 7, 23, 24, 46, and 50 followed the subject-verb agreement and raters did not give score five, which shows the complete subject-verb agreement in the sentence. Moreover, cases where the raters had different viewpoints concerning subject-verb agreement are as follow. The sentences of 1, 14, 33, 34, 37, 39, and 43 reached the mean score of 3 with regard to syntactic accuracy criteria, or in this study

the subject-verb agreement. This mean score, 3, means that just two raters out of four raters had full acceptance over its subject-verb agreement and the other two raters were at odds with this viewpoint.

Table 7

Human mean score out of 15

N	Valid	50
	Missing	0
Mode		4.50[a]
Minimum		4.50
Maximum		14.75

a. Multiple modes exist. The smallest value is shown

On the other hand, there are some cases that just one rater out of four raters agreed that there was a subject-verb agreement in the Persian translated sentences. These sentences are as follow: 5, 22, 26, 27, 28, 29, 30, 32, 35, 36, and 45. Finally, sentences were found that all four raters believed Google translator could apply the subject-verb agreement in the translation and the raters gave the score 5, complete score, to these sentences. This group of sentences involves the two sentences of 11 and 25 with the mean score of 14.75 and 11.75 respectively out of 15. It must be mentioned that these two sentences could take the best mean score of human judgment in this research. These details are summarized in table 8 below in different groups of raters.

Table 8

Groups of Raters Agreed Subject-verb Agreement

Groups	Sentence No	Total
No raters	2, 3, 4, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 18, 19, 20, 21, 23, 24, 31, 38, 40, 41, 42, 44,46, 47, 48, 49, 50	30
One rater	5, 22, 26, 27, 28, 29, 30, 32, 35,36, 45	11
Two raters	1, 14, 33, 34, 37, 39,43	7
Three raters	0	0
Four raters	11, 25	2

As a whole, according to the table 8 above, it can be claimed that among 50 sentences translated through Google translator only 20 sentences met the subject-verb agreement criteria in this research. This means that Google translator is different from the human translator in applying subject-verb agreement. Consequently, this point answers the first research question. Therefore, the first null hypothesis is rejected and it can be said that Google Machine Translation does not deal better with subject-verb agreement while translating English sentences into Persian compared to human translator. The scores obtained from the F-measure method of MT evaluation for each sentence and the mean scores obtained from human evaluators for each sentences have been given to the SPSS software. Since the two sets of scores were from different measurement scales, they have been converted to standard scores, Z-score.

Table 9

Descriptive Statistics

	N	Sum	Mean	Std. Deviation	Variance
Z-score(F-measure)	50	.00000	.0000000	1.00000000	1.000
Z-score(Human Mean Score)	50	.00000	.0000000	1.00000000	1.000
Valid N (list wise)	50				

To find out the relationship between these variables, the correlation between two sets of scores has been

calculated through Karl Pearson's Coefficient of Correlation formula. Table 9 demonstrates the magnitude of coefficient of correlation between these two sets of scores.

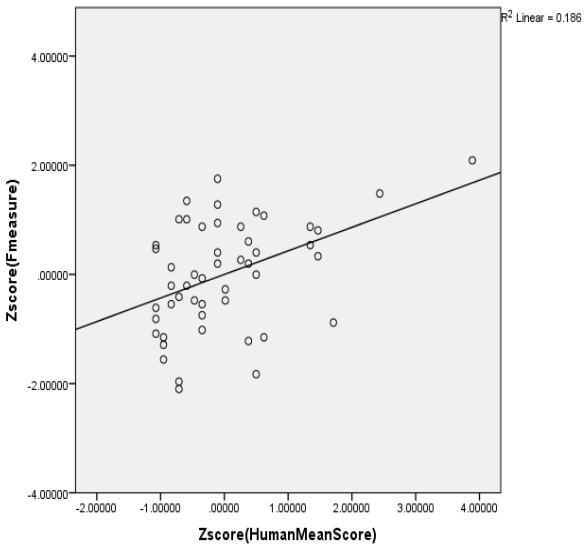
Table 10

Correlations

		Z-score (F-measure)	Z-score (Human Mean Score)
Z-score(F-measure)	Pearson Correlation	1	.432**
	Sig. (2-tailed)		.002
	N	50	50
Z-score(Human Mean Score)	Pearson Correlation	.432**	1
	Sig. (2-tailed)	.002	
	N	50	50

Note. **. Correlation is significant at the 0.01 level (2-tailed)

As table 9 illustrates, the correlation coefficient is 0.432. A scatter diagram has been drawn to show the general situation of scores in the graph. In addition line of best fit has been depicted for the scatter plot to show the direction of correlation and magnitude of how closed the scores are around the line of best fit. This can be seen in graph 1 below.



Graph 1. Scatter diagram of Z-score (F-measure) and Z-score (Human Mean Score)

The regression has been calculated. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. The result is shown in the following table 11.

Table 11

Linear regression model of two variables

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.432[a]	.186	.169	.91142206

a. Predictors: (Constant), Z score(Human Mean Score)

Since the researcher wished to use the model for predictive purposes the model summary table has been drawn. In this table, R indicates the correlation (0.432), the value of R Square (0.186) suggests that how the regression model explains the variation in the dependent variable. The value of R is a fraction between 0.0 and

1.0. An R Square value of 0.0 means that knowing X does not predict Y. The low value of R square here (.186) describes that the regression model does not explain the variation in the dependent variable well. In other words, achieving the scores by using human evaluators is not a good predictor to predict the scores obtained from F-measure evaluation method.

Table 12

ANOVA[a]

Model		Sum of Squares	*df*	Mean Square	*F*	Sig.
1	Regression	9.127	1	9.127	10.987	.002[b]
	Residual	39.873	48	.831		
	Total	49.000	49			

a. Dependent Variable: Z score(F-measure)

b. Predictors: (Constant), Z score(Human Mean Score)

ANOVA is used to test the hypothesis. In order to test it the means between two groups are equal under the assumption that the sample population is normally distributed. Once the regression model has been fit to a group of data, examination of the residuals (the deviations from the fitted line to the observed values) allows the modeler to investigate the validity of his or her assumption that a linear relationship exists. Residual is the difference between the observed value of the variable and the value suggested by the regression model. Here the difference between the Regression sum of squares and Residual sum of square is about 30.746 (39.873-9.127) that shows a high difference between the value of variables and regression model. In other words, the strength of variables cannot be determined in this ANOVA table. However, the p-value in this table is 0.002. If it is under 0.05 the variable is significant. The value we have here (0.002) is highly significant and shows that there is a significant relationship between two variables and implies that the second null hypothesis is rejected.

Table 13

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	*t*	Sig.
1	(Constant)	6.321E-18	.129		.000	1.000
	Z score(Human Mean Score)	.432	.130	.432	3.315	.002

a. Dependent Variable: Z score(F-measure)

The final table, table 13, presents the regression coefficients. The B weight in the Constant row refers to as the intercept. The B weight in the Predictor row (human mean score) refers to as the slope. The coefficients are used to form the following linear regression equation.

$$Y'=6.321 + 0.432$$

The model reveals that human evaluation accounts for 0.186 percent of the variance in software evaluation with a Pearson r=0.432, F(1,48)=10.987, p=.002. The result of the linear equation is:

$$Y'= 6.321 + 0.432$$

Through examining the results achieved from SPSS software, we can see that the amount of coefficient of correlation is 0.432, which shows a positive correlation between scores obtained from Human judges and F-measure MT evaluation method. Since the optimistically high positive coefficient of correlation is 1, the magnitude of coefficient of correlation obtained from this research is still markedly different from high positive correlation.

In Graph 1, the pattern of scattered dots shows a thinly scattered population around the line of best fit. The line of best fit is the line that comes as close as possible to as many scores as possible. In this research, the slope of this line is 0.432, which shows moderately positive coefficient of correlation between two variables. Also, p=0.002 is less than 0.05 and it shows that null hypothesis is rejected. However, the low value of R Square here (0.186) describes that the regression model does not explain the variation in the dependent variable well. In other words, having human evaluation scores is not a good predictor to predict the scores obtained from F-measure MT evaluation method. In addition, in the ANOVA table the difference between the Regression sum of squares and Residual sum of square is about 30.746 (39.873_9.127) that shows a high difference between the value of variables and regression model. In addition, the strength of variables cannot be determined in the ANOVA table. So, scores obtained from F-measure method cannot be predicted by scores obtained from Human evaluators. Therefore, researcher in this research has come to the conclusion that "there is a meaningful correlation between scores obtained from human judges and the scores obtained from F-measure MT evaluation method".

3.1 Summary of the Findings

As mentioned in previous section, among 50 sentences translated through Google translator only 20 sentences met the subject-verb agreement criteria in this research. This means that Google translator is different from the Human translator in applying subject-verb agreement because human translators are aware of the agreement principles but Google translator as translating statistically does not know these principles. Google translator can improve this shortcoming by being given more translation of English to Persian Therefore, the first null hypothesis is rejected and it can be said that Google Machine Translation does not deal better with subject-verb agreement while translating English sentences into Persian compared to human translator. It means that Google Machine Translator has long way to properly deal with subject-verb agreement in translating Persian sentences. On the other hand, through examining the results achieved from SPSS software, we can see that the amount of coefficient of correlation is 0.432, which shows a positive correlation between scores obtained from Human judges and F-measure MT evaluation method. Therefore there is a significant relationship between these two variables which rejects the second null hypothesis of the research.

4. References

Ahangar, A., Jahangiri, N., & Mohammadpour, F. (2012). A lexical-functional model for machine translation of English zero-place predicators into Persian. *International Journal of English Linguistics*, *2*(3). http://dx.doi.org/10.5539/ijel.v2n3p2

Catford, J. C. (1965). *A linguistic theory of translation.* Oxford: Oxford University Press.

Corbett, G. (2000). *Number.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139164344

Hutchins, J. (1985). *Machine translation past, present, future.* Chichester: Ellis Horwood.

Hutchins, W., & Somers, H. (1992). *An introduction to machine translation.* London: Academic Press.

Koehn, P. (2010). *Statistical machine translation.* Cambridge: Cambridge University Press.

Lotfi, A. R. (2006). Agreement in Persian. *Linguistik Online, 29*(4/06), 124.

Matthews, P. (1981). *Syntax.* Cambridge: Cambridge University Press.

Mirzaeian, V. (2010). *Challenges of machine translation in Persian, using Three MT systems* (Vol. 7). Tehran, Iran: Translation Studies.

Robinson, D. (2003). *Becoming a translator* (2nd ed.). Oxon: Routledge.

Shahahbi, M. (2009). An evaluation of output quality of machine translation program. *Student Research Workshop*, 71-75.

Taghavipour, M. (2004). On machine translation for Persian. *Translation Stu*dies, *2*(7-8), 33-50.