Is Google Translate Good Enough for Commercial Websites?
A Machine Translation evaluation of text from English websites into four different languages

Anthony Tobin

Abstract

This study is a Machine Translation evaluation of output from Google Translate, an online translation service that uses Statistical Machine Translation. The purpose of the study is to ascertain whether Google Translate is able to produce output of sufficiently high quality for use on commercial websites. Twenty sentences from the websites of four different language schools that use the Google Translate powered Google Website Translator plugin were selected. The sentences were translated into French, German, Japanese and Spanish. Native speakers of the languages volunteered to evaluate the translated output using two sets of scales; one for intelligibility and the other for accuracy. It was found that the results for accuracy and intelligibility were similar, with the German output receiving the worst evaluations for both metrics. The Japanese output for both metrics received the second worst evaluations. The Spanish output had the highest evaluation for intelligibility, and received the joint highest evaluation along with French for accuracy. Overall it was found that the

majority of the French and Spanish output was of a reasonably high quality, but should still be post-edited before appearing on a website. The German and Japanese output was of lower quality and needed more substantial correcting before being fit for publication on a website.

1 Introduction

In recent years it has become common to see websites that give the user the option of automatically translating the content of the site into another language. In most cases the translation is carried out by Google Translate, a Machine Translation system that automatically translates texts from one language to another. As of 2012 more than a million websites worldwide were using the Google Translate powered Website Translator plugin (Chin 2012). It is possible to find a wide variety of websites that use the free service, ranging from tourist information websites, language schools and even universities. In this study, translated output from Google Translate will be evaluated. The texts originate from the English language websites of four different language schools in Ireland. The target languages are French, German, Japanese and Spanish.

1.1 History of Machine Translation

The origins of Machine Translation (MT) can be dated to the years following the Second World War when researchers saw a link between translation and the cryptography employed in code breaking activities successfully carried out by computers during the war (Koehn 2010, p.15; Arnold et al. 1994, p.12; Hutchins and Somers 1992, p.5). Throughout the 1950s MT research gathered pace and received a great deal of funding in the USA, as well as in Europe, Canada and the USSR. In the USA research was largely focused on translating Russian into English (Hutchins and Somers 1992, p.6), as intelligence gathering on Soviet activity was a major pre-occupation of cold-war America (Arnold et al. 1994, p.13).

However, in the 1960s funding for MT began to dry up after a report by the Automatic Language Processing Advisory Committee (ALPAC) concluded that MT was too slow and inaccurate (Koehn 2010, p.15; Arnold et al. 1994, p13; Hutchins and Somers 1992, p.7). Interest in MT research increased again from the late 1970s, after successes such as that of the METEO system in Canada which was, and continues to be used to translate weather reports (Koehn 2010, p.16; Arnold et al. 1994, p.11; Hutchins and Somers 1992, p.7).

Perhaps the biggest boost to MT in recent years has been its proliferation on the World Wide Web. Babel Fish, which was launched in 1997, was the first MT system made available online to have a large impact (Yang and Lange 2003, pp.191-210). Many others have followed: Excite, Promt-Online, Infoseek, Amikai and Google Translate to name but a few. Now anyone with an Internet connection can translate using MT free of charge. It is no longer necessary for individuals to purchase expensive MT software.

1.2 Types of Machine Translation

There are several different approaches to designing MT systems. Some of these fall into the category of Rule-Based Machine Translation (RBMT), while others follow what is known as ‘empirical’ or corpus based approaches. Broadly speaking there are three different strategies to RBMT: the direct approach, the transfer method and the use of interlingua (Hutchins and Somers 1992, p.73). Two of the main empirical approaches include Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) (Arnold et al. 1994).

1.3 Google Translate

The type of MT used by Google Translate is Statistical Machine Translation (SMT). In SMT large bilingual aligned corpora, or text collections, are used to find translations. In SMT the concept of ‘making

optimal decisions' using statistical methods is used (Och 2005).

> The term 'statistical approaches' can be understood in a narrow sense to refer to approaches which try to do away with explicitly formulating linguistic knowledge, or in a broad sense to denote the application of statistically or probabilistically based techniques to parts of the MT task (Arnold et al. 1994, p.201).

Google explain that the system looks for patterns in hundreds of millions of documents that have already been translated by human translators and that Google Translate is able to make intelligent guesses to create an appropriate translation (Google 2010).

Although some work was done on SMT in the early years of MT research (Hutchins and Somers 1992, p.320), the principals on which current SMT systems are based were formulated as recently as 1990 (Ney 2005). Since this time improvements in algorithms, the advent of more powerful computers, as well as the creation of more powerful corpora have all led to SMT becoming a viable alternative to the more traditional types of MT (ibid.) IBM's SMT project based on a large bilingual French-English corpus from the Canadian Hansard, which records parliamentary debates in both languages, is one of the best-known experiments in SMT (Hutchins and Somers 1992, p.321). Google changed their online MT system from an RBMT system to an SMT system in 2007 (Schwartz 2007).

Google released their Google Translate powered Website Translator plugin in 2009. This allows website administrators to let users of their website instantly translate the content of the website into more than sixty languages, although most of the websites investigated for this study had a smaller range of languages available. In 2012 Google launched a new feature for the Website Translator which enables the website administrator

to edit translations and allows users to suggest a better translation. The website administrator may then accept or reject the suggested translation (Chin 2012). In other words it is now possible to post-edit the MT output in order to improve the quality of the translations.

1.4 Aims of the study

In this study the translated results of the four target languages will be evaluated by native speakers of each of the languages. They will evaluate two different criteria commonly used when evaluating MT output: intelligibility and accuracy. According to Arnold et al. (1994, p.163), intelligibility and accuracy are often closely related and accordingly the scores for both sets should be similar. Results for both metrics will be examined to see if that is the case in this study.

The main intention is to ascertain whether the quality of the MT output is high enough to be useful for a language school, or other business, to use in its online marketing activities. We will also examine whether some languages are more suited as target languages for Google Translate or if the output is of a similar quality across all languages.

2 Methodology

2.1 Website and sentence selection

It was decided to focus on the websites of businesses which market to clients who do not speak English as their native language, as it was felt there was a high likelihood of such businesses offering a translated version of their websites. English schools seemed the most obvious such business, so the selection was narrowed down to the websites of English schools in Ireland, the author's home country. The websites of twenty-five English schools in Ireland were checked to see what type of translation, if any, was available. It was found that nine of the schools had websites which seemed to offer a professionally translated version into several languages, eight of

the schools' websites used the Google Website Translator plugin or another Machine Translation system to automatically translate the content, and eight of the schools did not offer any translation of their website. Four of the websites that used Google's Website Translator were selected and five sentences from each of these websites were chosen for the translation evaluation, meaning a total of twenty sentences were selected. The sentences varied from quite short sentences with as few as twelve words, to longer sentences with up to forty-one words. The shortest and longest sentences can be seen below:

> Sentence 4. Language practice is in context and emphasis is placed on communicative ability.
> Sentence 12. In today's ever changing and demanding business, people need not only fluency of language; they also need to show they can persuade, sell and influence people in many different business situations, whether in the corporate world or running their own business.

2.2 Language selection

Although Google offers translation from English into many languages, some of the English schools' websites limited the languages to focus on their main markets. Initially it was decided to evaluate the output from English automatically translated into the following languages: French, German, Italian, Japanese, Polish and Spanish. However, as not enough evaluators could be found for Italian and Polish it was decided to focus on the remaining four languages. Most of the languages are European languages as these reflect the main markets of the language schools, but it was felt that having one non-European language with a different script might prove to be more challenging for Google Translate, and so Japanese was selected.

2.3 Method for evaluating MT output

A common evaluation technique is to have human evaluators assign scores to output sentences. Output may be evaluated for intelligibility and accuracy using scales (Trujillo 1999, pp.251-266; Arnold et al. 1994, pp.160-164; Hutchins and Somers 1992, p.164). Intelligibility is a measure of how fluent and grammatical the output of an MT system, or indeed text translated by a human translator, is (Trujillo 1999, pp.251-266). It may also be said to be a measure of how understandable a text is (Hutchins and Somers 1992, p.164). Intelligibility is also known as clarity, fluency and, sometimes, readability (FEMTI 2008). Style may or may not be taken into account when scoring for intelligibility (Arnold et al. 1994, p.161). Intelligibility is a useful measure of translation quality because even if a text is reasonably faithful to the source language input, if it is close to impossible to understand, it is not of much use. As ranks given to target language sentences are largely subjective, scores of several evaluators should be combined and a more objective assessment can be reached by means of statistics (Trujillo 1999, pp.251-266). Arnold et al. (1994, p.162) suggest a minimum of four evaluators and they also point out that they should be familiar with the chosen subject area. While scoring for intelligibility, evaluators should not be able to refer to the source language text (ibid.).

Accuracy, or fidelity as it is sometimes known, is a measure of the extent to which a translated text preserves the content of a source text (Trujillo 1999, pp.251-266). To get a broader picture of translation quality both intelligibility and accuracy are ranked as a pair. Arnold et al. (1994, p.162) point out that scoring for accuracy should be carried out after scoring for intelligibility has already been completed. Therefore the evaluators in this study were requested to leave a gap of three days after completing the intelligibility evaluation before starting the accuracy evaluation. Unlike intelligibility, when scoring for accuracy the evaluators need to be able to refer to the source text, so evaluators should have the

necessary linguistic skills (ibid.). As with intelligibility there may be a lack of objectivity with some evaluators scoring more strictly than others, but by using as many evaluators as possible a reasonably clear picture of the accuracy of a given MT system should emerge. As well as this problem of inter-rater inconsistency, there is also the problem of intra-rater inconsistency where one evaluator will mark the same sentence differently on different occasions. This highlights the inherent difficulties in using humans to evaluate MT output.

2.4 Intelligibility scale selection

Many scoring scales have been developed to rate the intelligibility of output from MT systems. In such scales sentences that resemble perfect sentences in the target language are given top marks, while sentences that have become so mangled that it would be close to impossible for an evaluator to even hazard a guess at the meaning of, are given bottom marks. Scales for intelligibility have ranged from two-point scales to as high as nine-point scales (Arnold et al. 1994, p.161). A two point-scale would only have the options of either “intelligible” or “unintelligible”, but as pointed out by Arnold et al. this gives no indication as to the seriousness of the errors that affect intelligibility. They also mention the nine-point scale which featured in the ALPAC report, but found that as it was also produced to evaluate human translation it was not very suitable to evaluating the output of MT as it included judgements on very subtle differences in style, etc.

The scale selected for this study was the one developed by Trujillo (1999, pp.251-266). This scale, seen in Table 1, is a five-point scale that includes very clear, unambiguous descriptions for each point on the scale. In this scale sentences with a score of 1 have the highest intelligibility and sentences with a score of 5 have the lowest intelligibility.

1	The meaning of the sentence is clear, and there are no questions. Grammar, word usage, and/or style are all appropriate, and no rewriting is needed.
2	The meaning of the sentence is clear, but there are some problems in grammar, word usage and/or style, making the overall quality less than 1.
3	The basic thrust of the sentence is clear, but you are not sure of some detailed parts because of grammar and word usage problems. You would need to look at the original source language sentence to clarify the meaning.
4	The sentence contains many grammatical and word usage problems, and you can only guess at the meaning after careful study, if at all.
5	The sentence cannot be understood at all.

Table 1: Trujillo's intelligibility scale

2.5 Accuracy scale selection

When scoring for accuracy a similar type of scale to that used for intelligibility is usually used. The main difference in the procedure for evaluating accuracy is that the evaluators need to be able to refer to the source text to gauge how closely its meaning is transferred to the translated output.

The accuracy scale chosen (Table 2) was, like the intelligibility scale, developed by Trujillo (1999, pp.251-266). Unlike the intelligibility scale, the scale for accuracy is a seven-point scale. While it may have been convenient to use a five-point scale to be able to directly compare a correlation, or otherwise, of accuracy with intelligibility in a straightforward manner, it was decided to use Trujillo's scale as it was very detailed in its specifications. In this scale sentences with a score of 1 have the highest accuracy and sentences with a score of 7 have the lowest accuracy.

1	The content of the source language (SL) sentence is faithfully conveyed to the target language (TL) sentence. The translated sentence is clear to a native speaker of the TL and no rewriting is needed.
2	The content of the SL sentence is faithfully conveyed to the TL sentence, and can be clearly understood by a native speaker, but some rewriting is needed.
3	The content of the SL sentence is faithfully conveyed in the TL sentence, but some changes are needed in word order.
4	While the content of the SL sentence is generally conveyed faithfully in the TL sentence, there are some problems with things like relationships between phrases and expressions, and with tense, plurals, and the position of adverbs. There is some duplication of nouns in the sentence.
5	The content of the SL sentence is not adequately conveyed in the TL sentence. Some expressions are missing, and there are problems with the relationships between clauses, between phrases and clauses, or between sentence elements.
6	The content of the SL sentence is not conveyed in the TL sentence.
7	The content of the SL sentence is not conveyed at al. The output is not a proper sentence; subjects and predicates are missing.

Table 2: Trujillo's accuracy scale

2.6 Profile of evaluators

As Arnold et al. (1994, p.162) stated that four was the minimum number of evaluators acceptable in an evaluation study, it was decided to use at least this number. All of the evaluators spoke the target language they evaluated as their native language and it was also required that they had an excellent understanding of English as it was necessary for them to be able to understand the instructions of the evaluations and also to be able to understand the original English sentences in the accuracy evaluation, where sentences from the source language and target language were compared. All of the evaluators were university educated, and among them were language teachers, university lecturers, post-graduate students, translators and engineers. The evaluators were assigned letters; A-D

evaluated the French output, E-H evaluated the German output, I-L evaluated the Japanese output, and M-P evaluated the Spanish output. Evaluator C for French submitted only the intelligibility evaluation, so it was necessary to find another French evaluator for accuracy. This new evaluator was designated as C2. In total there were seventeen evaluators involved in the study – four for each language except for French which had five due to the circumstances explained above.

2.7 Rules for experiment

The evaluators were each given a pre-experiment briefing and a set of instructions for intelligibility and accuracy. In the pre-experiment briefing they were instructed to carefully read the instructions before beginning, and to make themselves familiar with the scoring scales they would be using. The evaluators were informed in the pre-experiment briefing that they would be evaluating MT output that was automatically translated into their language and taken from language school websites offering this MT function. The evaluators were also instructed to leave a gap of three days between doing the first and second evaluations. This was to avoid the evaluators working on the second evaluation while the first one was still fresh in their memories. The evaluators were also requested not to discuss the task with any other evaluators involved in the study. They were also recommended to print out each evaluation, as this would be easier than reading from the screen. The final recommendation was to carry out the task in a quiet room, so that they could give their full attention to the task.

3 Analysis

In this section the data from the intelligibility and accuracy evaluations will be analysed. The data may be helpful in determining which of the languages, if any, should be considered by businesses when using Google's Website Translator on their websites.

3.1 Intelligibility

In Figure 1 the average intelligibility score for each evaluator can be seen. This shows the average score given by each of the sixteen evaluators for intelligibility in their evaluation of the Google Translate output into their target language. As can be seen, the vertical axis ranges from 1 to 5 to reflect the scoring system for intelligibility explained in the Methodology. The lower the score a language receives, means the better the average intelligibility evaluation of the output from Google Translate. The evaluators for French showed a great deal of consistency with all the evaluators giving an average score of around 2.5 on the 5-point scale. The evaluators for German also showed a reasonable amount of consistency with all average scores being close to 3. For Japanese, evaluator J gave a much harsher evaluation than the other three evaluators, and for the Spanish evaluations, evaluator P has given a noticeably harsher evaluation than the other three evaluators.

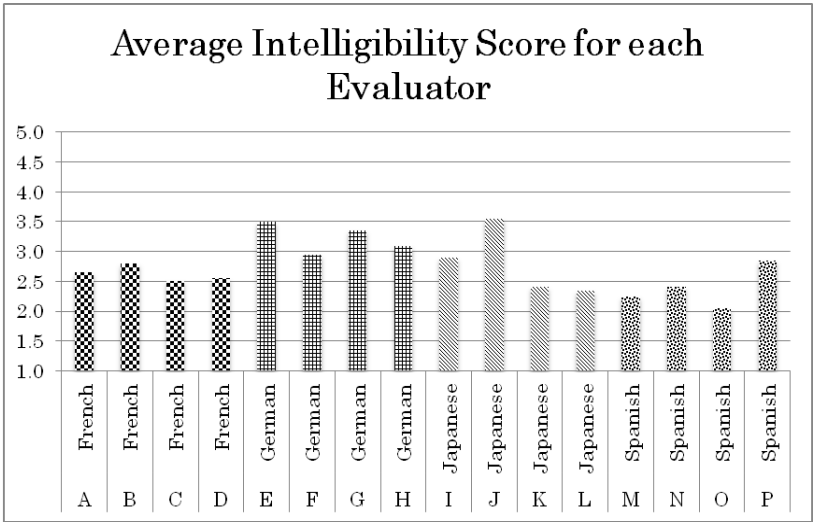


Figure 1: Average intelligibility score for each evaluator

In Figure 2 the overall average intelligibility score for each of the languages can be seen. The average score of each of the four evaluators for each language (Figure 1) was taken and an average score for each language was calculated. With a score of 2.4, Spanish received the best score. French had the second best score with 2.6, and Japanese was third with a score of 2.8. The German output was given the worst evaluation with a score of 3.2.

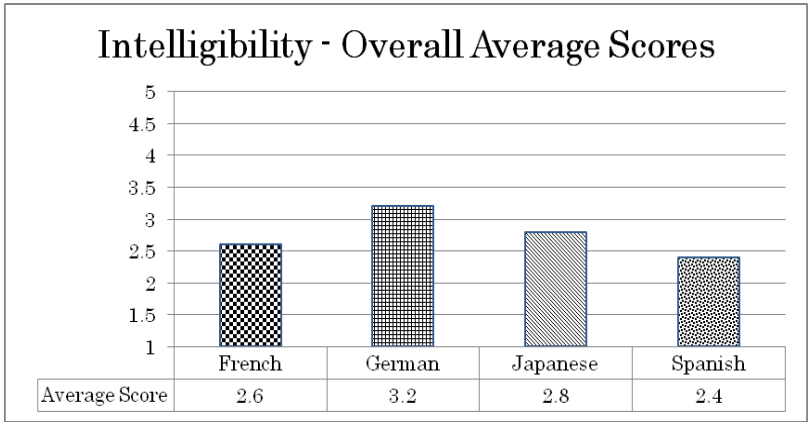
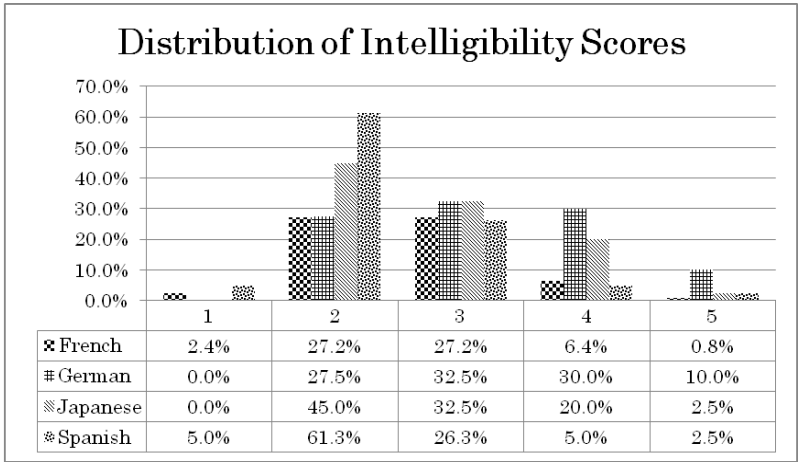


Figure 2: Overall average intelligibility scores

According to these scores the Spanish output was the most intelligible and the German output the least intelligible. Excluding Spanish, if one rounds off the scores to the nearest whole number, all of the languages scored an average of 3. This equates to the following:

> 3. The basic thrust of the sentence is clear, but you are not sure of some detailed parts because of grammar and word usage problems. You would need to look at the original source language sentence to clarify the meaning.

In Figure 3 the total distribution of the intelligibility scores can be seen. On the horizontal axis the scores for each language are shown and on the vertical axis the percentage of scores given for each point on the scale is shown. As there were twenty sentences and four evaluators for each language, each language received a total of eighty scores.



	1	2	3	4	5
French	2.4%	27.2%	27.2%	6.4%	0.8%
German	0.0%	27.5%	32.5%	30.0%	10.0%
Japanese	0.0%	45.0%	32.5%	20.0%	2.5%
Spanish	5.0%	61.3%	26.3%	5.0%	2.5%

Figure 3: Distribution of intelligibility scores

It is clear that very few top scores were given for the translations into any of the languages. Not a single score of 1 was given to the German or Japanese output and Spanish had the highest number of top scores, but at just 5% it is not a significant amount. The Spanish output received a very high number of scores at 2 on the scale – 61.3%. Almost half the scores given to Japanese were at 2 on the scale, but just over a quarter of the scores for French and German output were at this point on the scale. All of the languages received a significant number of scores at 3 on the scale, with Spanish receiving the lowest number with 26.3%, and Japanese and German receiving the highest number of 3 scores with 32.5% each. Point 4 on the scale represents translation with quite poor intelligibility and both German and Japanese had a significant proportion of such scores with 30% and 20% respectively. French and Spanish had a smaller amount of scores

of 4 with just 6.4% and 5% respectively. Point 5 represents translation that cannot be understood at all and most of the languages had very few scores at this point on the scale, but a significant 10% of the German scores were at this point on the scale.

Overall it is clear that the Spanish output from Google Translate was the most intelligible with an average score of 2.4 and very few scores at 4 or 5 on the scale. According to the average scores, the difference in quality of the French and Japanese output was not very large. However, when one examines the distribution of the scores it seems that the Japanese output was very mixed with many scores of 2, but also 22.5% of scores were at 4 or 5 on the scale. In contrast, the French output received only 7.2% of its scores at 4 or 5 on the scale. The German output was evaluated as being the least intelligible with an average score of 3.2 and 40% of scores at 4 or 5 on the scale. Perhaps the most surprising outcome is that the German output received the lowest score. One might have expected the Japanese output to receive the lowest score due to the large differences between the English and Japanese languages compared with the European languages. It may be possible that the German output was not actually less intelligible than the Japanese output, but the German evaluators were harsher in their evaluations than the Japanese evaluators. With such a large difference between the English and Japanese languages, perhaps the Japanese evaluators were more forgiving of errors and therefore more lenient in their evaluations. After all, it is not unusual to see low quality human translations of Japanese into English and vice versa, whereas German speakers may not be accustomed to seeing translation of such low quality.

3.2 Accuracy

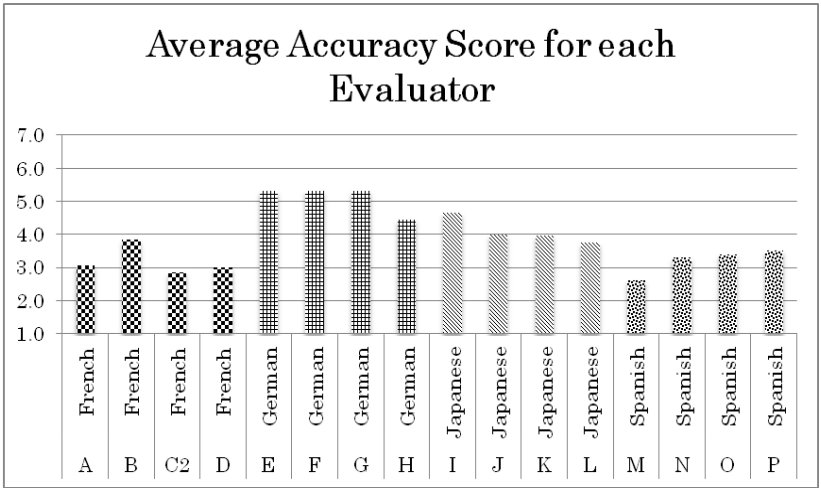


Figure 4: Average accuracy score from each evaluator

The average accuracy score for each evaluator can be seen in Figure 4. As discussed in the Methodology, this time the scale is from 1 to 7 with lower scores representing better accuracy and higher scores representing worse accuracy. The evaluators for each language generally seem to be quite consistent, with the possible exception of evaluator B who was harsher in his evaluations than the other French evaluators, and evaluators H and M who seem to have been more lenient than the other evaluators for their respective languages.

In Figure 5 we can see that there are quite large differences between the languages with German having the worst average accuracy score with 5.1, Japanese in third place with 4.1, and Spanish and French both having an average score of 3.2. A score of 3 on the seven-point scale equates to the following:

3. The content of the source language (SL) sentence is faithfully

conveyed in the target language (TL) sentence, but some changes are needed in word order.

Therefore these average accuracy scores for French and Spanish suggest that the output is very accurate, but the word order is not entirely correct.

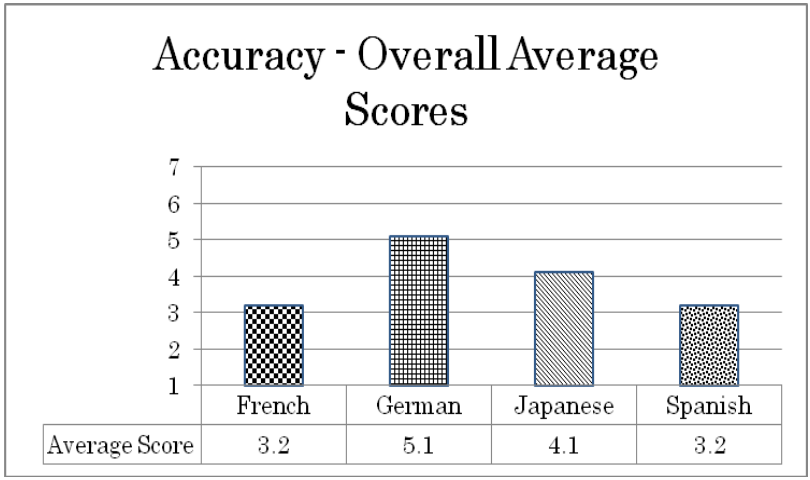


Figure 5: Overall average accuracy scores

The average score for the Japanese output of just over 4 on the scale equates to the following:

> 4. While the content of the SL sentence is generally conveyed faithfully in the TL sentence, there are some problems with things like relationships between phrases and expressions, and with tense, plurals, and the position of adverbs. There is some duplication of nouns in the sentence.

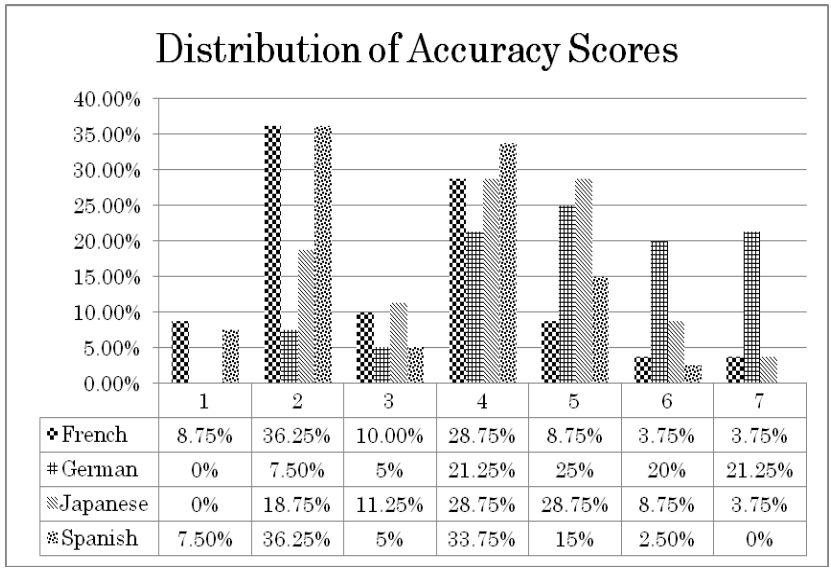
While the quality of the output is not as high as the Spanish and French output, it seems that it may be possible to correct the mistakes with some editing.

The German score of just over 5 on the scale equates to the following:

> 5. The content of the SL sentence is not adequately conveyed in the TL sentence. Some expressions are missing, and there are problems with the relationships between clauses, between phrases and clauses, or between sentence elements.

This suggests that the level of accuracy is really quite poor and that correcting the content would require more than some simple editing, and that it may be necessary to have the content translated from scratch by a human translator.

As we have seen above, French and Spanish had the same overall average score for accuracy with 3.2. In Figure 6 the distribution of the accuracy scores can be seen and here also there are many similarities in the scores for the French and Spanish output. Both languages received quite a small number of scores at 1 on the scale, but both had more than a third of their scores at 2 on the scale, with 36.25% each. Strangely, both languages received quite a small percentage of scores at 3 on the scale despite achieving an average score of 3.2, but they did both receive many scores at 4 on the scale, with 28.75% for French and 33.75% for Spanish. French had less than 9% of its scores at 5 on the scale, while Spanish had a more significant 15% of its scores at this point. Both languages had very few scores of 6 or 7 on the scale.



	1	2	3	4	5	6	7
French	8.75%	36.25%	10.00%	28.75%	8.75%	3.75%	3.75%
German	0%	7.50%	5%	21.25%	25%	20%	21.25%
Japanese	0%	18.75%	11.25%	28.75%	28.75%	8.75%	3.75%
Spanish	7.50%	36.25%	5%	33.75%	15%	2.50%	0%

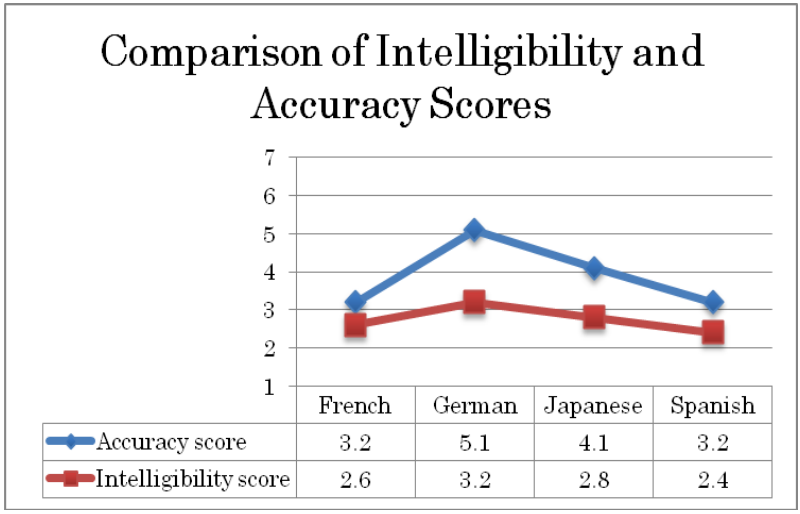
Figure 6: Distribution of accuracy scores

The Japanese output did not receive any scores of 1, but quite a few scores of 2 (18.75%) and a smaller percentage of scores at 3 on the scale with 11.25%. Slightly under 60% of the scores for the Japanese output were are at 4 or 5 on the accuracy scale, with 28.75% each. A total of 12.5% of the scores for the Japanese output were at 6 or 7 on the scale. Overall, despite some good scores the Japanese output is quite mixed, and this suggests that editing the output may not be so simple.

As we have seen, the German output had the worst average accuracy score with 5.1, and the distribution of the scores shows that only 12.5% of the scores were either 2 or 3, and there was not a single score at 1 on the scale. The vast majority of the scores for the German output were between 4 and 7 on the scale with all points having around 20%, except for point 5, which had a little more with 25% of the scores.

3.3 Comparison of intelligibility and accuracy scores

In Figure 7, we can see that there is a high correlation between the overall average intelligibility and accuracy scores. Naturally, the intelligibility scores appear below the accuracy scores due to the use of a five-point scale and a seven-point scale respectively. However, we can see the German content received the worst evaluations overall for both metrics and that Japanese was second worst. There is a slight difference between the results of the two metrics for Spanish and French. Spanish was ranked slightly better than French for intelligibility, but both languages received the same average score for accuracy.



	French	German	Japanese	Spanish
Accuracy score	3.2	5.1	4.1	3.2
Intelligibility score	2.6	3.2	2.8	2.4

Figure 7: Comparison of overall average intelligibility and accuracy scores

Conclusion

We have seen that Google Translate was able to produce the most intelligible and accurate output from English into French and Spanish. The Japanese output only scored slightly worse in its average intelligibility score than French, but the Japanese score for accuracy showed that the

quality was not very high. German fared the worst overall, though whether this indicates that the output was actually of lower quality than the Japanese output is debateable due to the possibility that the German evaluators may have been stricter than the Japanese evaluators.

The main question is whether the MT output produced by Google Translate from English into any of these languages is acceptable for use on a commercial website. For the German output, and probably also for the Japanese output, the quality of the translation is not high enough for dissemination and therefore not very suitable for a website. The translated output may lead to confusion due to its lack of intelligibility and accuracy. It may also lead to prospective clients, or in this case students of the language school, holding a poor opinion of the business due to the lack of a professionally translated website. In the majority of the sentences chosen, it would seem that the Spanish and French output is understandable, but not perfect. Output of such quality may be fine within a company or for an individual to use, as it is certainly useful for information purposes. However, once again it is questionable whether such output is suitable for marketing purposes on a website as it may deter some prospective clients.

Ideally, language schools and other businesses would use professional translators for their websites. However, this is usually quite costly and not all businesses have a budget for translation. Using the Google Website Translator plugin may seem like an attractive alternative for businesses in this situation, but it should be made clear to website users that Machine Translation is being used and that the quality may be low and is for information purposes only. To be fair, Google Translate does not guarantee perfect translations as can be seen on the About Google Translate page (Google 2010), but prospective students may not be aware of this. Therefore a disclaimer stating such information should appear when using the Website Translator.

As mentioned in the Introduction of this study, the Google Website Translator plugin does offer a function which enables administrators and users of the website to edit translations by suggesting a better translation. This function seems to have been ignored by the language schools in this study, as very few sentences from any of the languages scored top marks in either intelligibility or accuracy. Businesses could have native speakers edit the output to produce translations of much higher quality. In the cases of Spanish and French it may not even be necessary for the editors to have a detailed knowledge of the source language, as most of the output was understandable and just needed corrections in grammar and word order. This would be a much lower cost alternative to having the content professionally translated, and in the case of a language school it should be possible to find native speakers of the target languages amongst their students.

References

Arnold, D., Balkan, L., Humphreys, R., Meijer, S.and Sadler, L. 1994. *Machine Translation: An Introductory Guide.* Oxford: NCC Blackwell.

Chin, J. 2012. Now you can polish up Google's translation of your website. *The Official Google Translate Blog* [Online], 30 May. Available from:http://googletranslate.blogspot.jp/2012/05/now-you-can-polish-up-googles.html [Accessed 1 September 2014]

FEMTI. (Homepage). [Online]. Available from: http://www.isi.edu/natural-language/mteval/ [Accessed 3 September 2014]

Google 2010. *Inside Google Translate* [Online]. Available from: http://translate.google.com/about/intl/en_ALL/ [Accessed 1 September 2014]

Hutchins, W.J. and Somers. H.L. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge: Cambridge University Press

Ney, H. 2005. One Decade of Statistical Machine Translation. *IN*: *Proceedings from MT Summit X*. September 2005. Phuket, Thailand. pp.i-12-17.

Och, F.J. 2005. Statistical Machine Translation: Foundations and Recent Advances. *IN*: *Proceedings from MT Summit X*. September 2005. Phuket, Thailand. Tutorial Note.

Schwartz, B. 2007. *Google Drops Systran for Home Brewed Translation* [Online]. Available from: http://searchengineland.com/google-translate-drops-systran-for-home-brewed-translation-12502 [Accessed 1 September 2014]

Trujillo, A. 1999. *Translation Engines: Techniques for Machine Translation*. London: Springer-Verlag.

Yang, J. and Lange, E. 2003. Going live on the internet. *IN*: Somers, H. (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam-Philadelphia: John Benjamins.