EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 58–70
CILC2016. 8th International Conference on Corpus Linguistics

Graded Acceptance in Corpus-Based English-to-Spanish Machine Translation Evaluation[1]

Mario Crespo Miguel[1] and Marta Sánchez-Saus Laserna[1]
[1] Instituto Universitario de Investigación en Lingüística Aplicada, Universidad de Cádiz, Cádiz, Spain
mario.crespo@uca.es, marta.sanchezsaus@uca.es

Abstract

Traditionally, texts provided by machine translation have been evaluated with a binary criterion: right or wrong. However, in certain cases it is difficult to provide a clear-cut division between fully acceptable and fully unacceptable texts. In this paper we have selected group of different bilingual, human-translated English-to-Spanish pairs of sentences from parallel corpora (Linguee) and a group of machine translated texts with problematic linguistic phenomena in English-to-Spanish translation: polysemy, semantic equivalents, passive, anaphora, etc. We presented the translations to a group of native speakers that evaluated them in different levels of acceptability. Results show the degree of applicability of this approach.

1 Introduction

In the era of information explosion and global integration, the demand on translation in various fields is increasingly swollen. This situation calls for more and more translators competent in translating large quantity of materials in various applied fields (Erwen & Wenming, 2013). A huge demand, price competition and need for speed has led translators to a progressive incorporation of computer-assisted tools. Information and Communication Technologies play an essential role in the translator's work in terms of quality and productivity (González Boluda, 2010). Hutchins (1995) defines Machine Translation as computerized systems responsible for the production of translations with or without human assistance. Machine translation can increase the productivity of professional translators from 30% to 50% (Alcina Caudet, 2011).

[1] This paper is part of the R&D Project "Comunicación especializada y terminografía: usos terminológicos relacionados con los contenidos y perspectivas actuales de la semántica léxica" (FFI2014-54609-P), funded by the Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia of the Spanish Ministerio de Economía y Competitividad. Both authors belong to the Research Group "Semaínein" (HUM 147).

We can describe, among the most important factors to obtain a high quality machine translation, the degree of text specialization (Gutiérrez & Figueroa, 2011), the text type, the lexical density, the cultural distance, the purpose, oral or written form and degree of standardization, normalization and linguistic resources for a language (Abaitua, 2002).

Given the increasing presence of machine translation systems, different works have evaluated the performance of several freely available on-line machine translation systems into Spanish (Aiken, Vanjani, & Wong, 2009) (González Boluda, 2010), (Arenas, 2010), (Domínguez, Laurenti, & Céliz, 2013). Acceptability judgments by native speakers are generally accepted as the main type of evidence in linguistic theory. As pointed by (Schütze, 1996), this is due to a set of factors:

- Acceptability judgments allow us to examine sentences that rarely occur in spontaneous speech or corpora.
- Judgments constitute a way of obtaining negative evidence, which is rare in normal language use.
- In observing naturally occurring speech data, it is difficult to distinguish errors (slips of the tongue, unfinished utterances, etc.) from grammatical production.
- The use of acceptability judgments allows us to minimize the influence of communicative and representational functions of language. Judgment data allow us to study the structural properties of language in isolation.

These judgments about acceptability are traditionally based on a binary criterion in Linguistics, that is, acceptable and unacceptable, in order to judge the grammaticality of the different sentences. However, this constitutes an idealization since data show that the limits between acceptability and unacceptability are not clear, and different grades of acceptability can be established. Manning (2002) considers that in Linguistics such a grammatical binary criterion has been maintained by appealing to an idealized speaker/ hearer, but there is a fuzzy edge in the grammar, determined by many conflicting constraints and issues of conventionality vs. human creativity. This work presents an experimental approach to machine translation evaluation by considering different grades of acceptability. There is an absence of systematic empirical studies about the gradience-based translation evaluation.

2 Machine Translation

As pointed out by Kit & Wong (2008), the novelty of online Machine Translation (MT) services may give the impression that MT is something quite new. However, first attempts to MT started during the fifties. From the beginning of Machine Translations, approaches can be categorized into 3 main groups (Kliffer, 2008): direct (specific language-pairs, with term-by-term equivalence lists), transfer (specific language-pairs, with discrete structural and semantic analyses of source and target Ls) and interlingua (a single “universal” intermediate representation, independent of both source and target).

The first online free translation on the Internet appeared in 1997 by Babel Fish using Systran technology (Aiken, Vanjani, & Wong, 2009). In 2007 Google Translate online translator appeared, relying more on the statistical approach and the comparison of matching probabilities, than on the rule-based approach. Ever since, it has been included in almost every evaluation study.

This paper focuses on the analysis of acceptability judgments from the translations provided by two on-line translation engines: Google Translate and Systran. Google Translate is a free translation service that provides instant translations to more than 100 different languages (Google Translate, 2016). The approach to Machine Translation is statistical (González Boluda, 2010) (Arenas, 2010), so it provides translation after analysing the most likely equivalent texts from a huge amount of corpora (Seljan, Brkić, & Kucis, 2011).

Systran is a traditional program in the field of machine translation with forty years of history, and it has made remarkable efforts in research for constant updating. It was founded by Dr. Peter Toma in 1968 and has worked for many years for the United States of America Department of Defense and for the European Commission. The system in the market right now is based on hybrid technology, which combines rule-based and statistical translation (Costa-Jussà & Fonollosa, 2015). Today it is one of the most widely used systems, but often the clients are not aware of that. This system is integrated in applications such as text translators in Mac OS X operating system, the Yahoo! and Yahoo! Babel Fish online translators, and was also used by the Google search engine until 2007 (Arenas, 2010).

3 Experimentation Methodology

3.1 Corpus Preparation

Firstly, we focused on sentences traditionally known to cause problems to native Spanish speakers. These phenomena are reviewed by García, Meilán, & Martínez (2005), Gómez Manzano *et al*. (2005), and Gómez Torrego (2011). Among them we particularly concentrated on the use of gender, plural, verbal tenses, passive sentences, impersonal expressions, polysemy, use of clitics and anaphora, subordinate clauses and verbal subcategorization.

Once a list of frequent misuses phenomena was created for Spanish, we collected a small corpus of different bilingual, human-translated English-to-Spanish pairs of sentences from the online parallel corpora Linguee[2]. This tool is an aided-translation tool to search billions of bilingual, translated sentence pairs, which come from the World Wide Web, mainly from professionally translated websites of companies, organizations, universities or EU documents and patent specifications. Therefore, it is not a translation machine; it is a parallel corpus of human-translated texts.

An expert-based selection of English sentences was carried out by checking if their Spanish counterpart translation could pose difficulties to native speakers and, in addition, to Machine Translation engines. A final set of 20 different English sentences was taken from the online parallel corpus provided by Linguee. The correct Spanish equivalent was also collected. We wanted to evaluate how a translation engine behaves for such translations and how far it is from the correct human translation.

The set of English sentences selected was taken to the online automatic machine translation engines above referred. Once translated into Spanish, original human translated sentences and those provided by the machine translation were presented to a group of native speakers that evaluated them according to a 0-10 likert scale degree of acceptability.

3.2 On-line Questionnaires

We created 3 separated questionnaires, with 10 questions each. The questions were presented as Figure 1 shows: we provided the original English sentence followed by a Spanish translation. Each questionnaire only contained a translation pair, either the one from a human translator, or the translation coming from engine 1 or the translation from engine 2, but never two or more translations of the same sentence in the same questionnaire. Therefore, informants could never compare different translations of the same expression; they only had access to one translation.

The survey was designed as a Google Drive form, so the questionnaires were online accessible (both PC and mobile). Each sentence was evaluated according to a 0-10 likert scale, 0 representing a completely unacceptable translation and 10 a perfect translation one. There was no specific time to complete the survey.

[2] http://www.linguee.es/

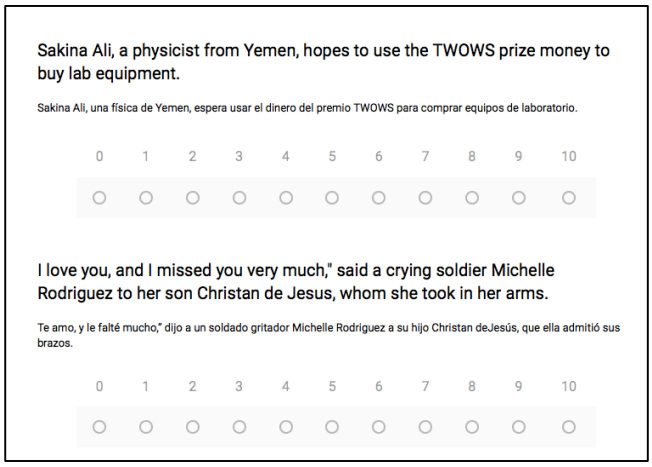

Figure 1. Example of questions in the survey

Informants evaluated sentences following a random order at a time. Previously they were required to provide information about their gender, highest level of education completed and whether they are professionally related to Translation, Linguistics or Philology or not.

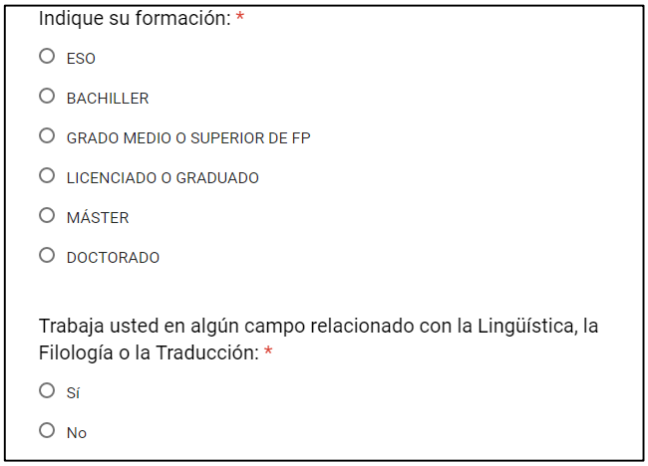

Figure 2. Demographic survey excerpt

3.3 Informants

Any possible subject was allowed to complete the evaluation questionnaire, but it was previously informed that subjects at a B2 English Level according to CEFR (Common European Framework of Reference for Languages) or higher were preferred. 83 informants answered the questionnaire during January and February 2016. 9 of them were ruled out because they were under B2 English level. In the final sample selection, 47.3% were over a B2 English level.

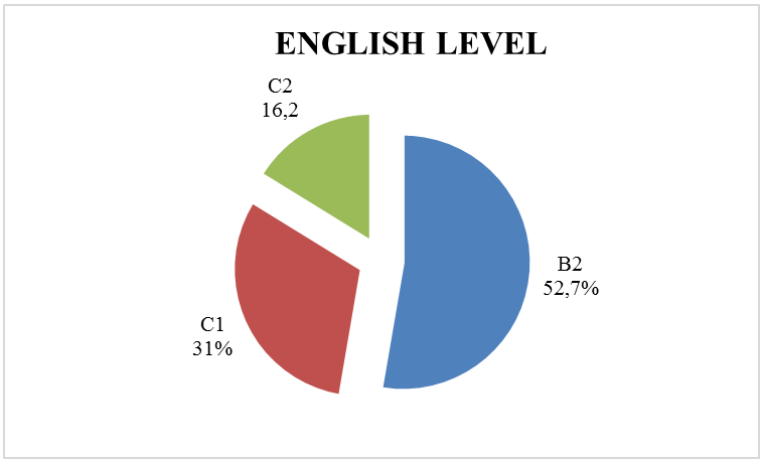

Figure 3. Informants: English level

80% of our informants had university studies. No secondary school student joined the questionnaire, so all the participants were over 16 years old. From the point of view of level of studies, more than half were master or PhD students.

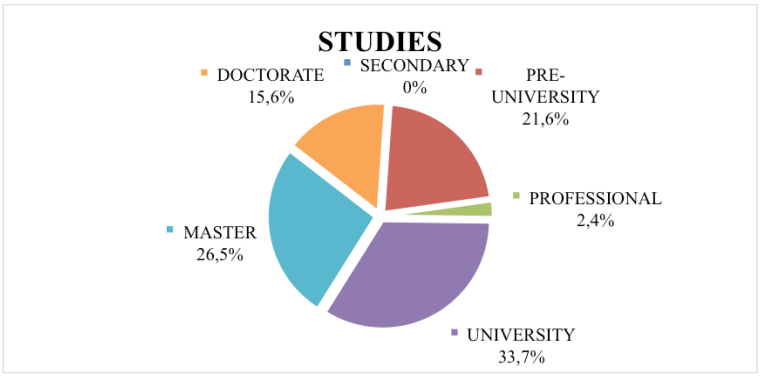

Figure 4. Informants: highest level of studies

The sample was divided into two groups: professionally related to Translation, Linguistics or Philology (45%) and not related (55%). This allows us to observe if there are significant differences between these two groups of informants.

Figure 5. Informants professionally related to Translation, Linguistics or Philology

4 Results

The analysis of the evaluations provided by informants has given the following results: 81.3% of human translations are evaluated over 5, that is, they lay in the favourable half of the Likert scale. From them, 50% is between 9 and 10, the best of them. Favourable proportion is 63.22% for Google and 25.90% for Systran. Just 13.62% of human translations scored under 5; 28.63% scored under 5 for Google. More than a half (66.61%) of Systran translations are considered negatively and 16,99% has been considered totally unacceptable. A standard deviation computed from proportions states how human translations are focused on the highest values, whereas Google and Systran spread out among the range of values more proportionally. This also can be observed in **Error! Reference source not found.**. This plot shows how scores distribute and it can be compared to the rest of translators.

Figures 7 and 8 compare score amounts among the three translators. Human translation proportion decreases as a lower score is obtained. So is it for Google, whose values are more stable over the range of the likert scale. Inversely, Systran increases with the lowest values.

	HUMAN			GOOGLE			SYSTRAN		
	Porcentage	Cumulative	Inverse	Porcentage	Cumulative	Inverse	Porcentage	Cumulative	Inverse
10	27,64%	27,64%	100,00%	14,51%	14,51%	100,00%	3,61%	3,61%	100,00%
9	21,75%	49,39%	72,36%	13,32%	27,83%	85,49%	4,03%	7,64%	96,39%
8	14,23%	63,62%	50,61%	12,92%	40,76%	72,17%	6,16%	13,80%	92,36%
7	10,37%	73,98%	36,38%	11,33%	52,09%	59,24%	4,46%	18,26%	86,20%
6	7,32%	81,30%	26,02%	11,13%	63,22%	47,91%	7,64%	25,90%	81,74%
5	5,08%	86,38%	18,70%	8,15%	71,37%	36,78%	8,49%	34,39%	74,10%
4	5,49%	91,87%	13,62%	9,34%	80,72%	28,63%	12,10%	46,50%	65,61%
3	4,27%	96,14%	8,13%	6,56%	87,28%	19,28%	12,74%	59,24%	53,50%
2	2,85%	98,98%	3,86%	5,37%	92,64%	12,72%	12,95%	72,19%	40,76%
1	0,61%	99,59%	1,02%	4,17%	96,82%	7,36%	10,83%	83,01%	27,81%
0	0,41%	100,00%	0,41%	3,18%	100,00%	3,18%	16,99%	100,00%	16,99%
S DEVIATION	8,80%			3,89%			4,37%		

Table 1. Score percentage according to score for each type of translator

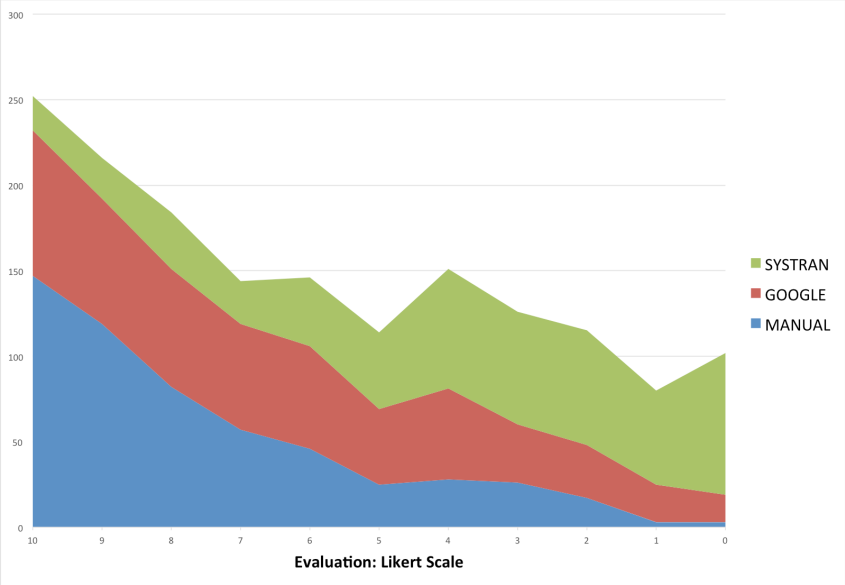

Figure 6. Area chart showing total scores according to translator

Error! Reference source not found. and **Error! Reference source not found.** compare score amounts among the three translators. Human translation proportion decreases as a lower score is obtained. So is it for Google, whose values are more stable over the range of the likert scale. Inversely, Systran increases with the lowest values.

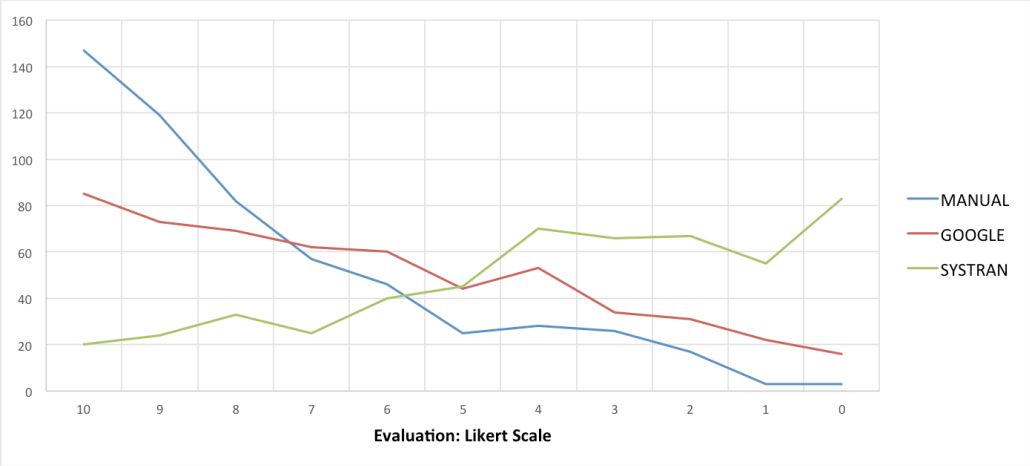

Figure 7. Area chart showing percentage of scores according to translator

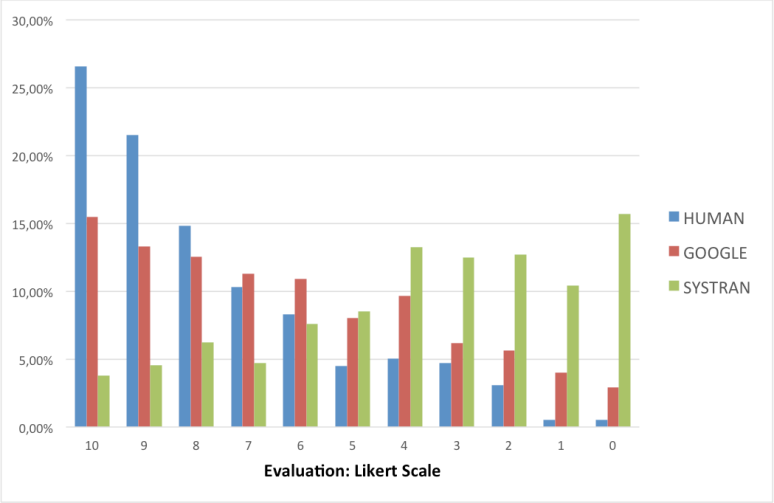

Figure 8. Bar plot showing percentage of scores according to translator

Table 2 shows the descriptive statistics for the three translators. Human translations are those with the highest mean and less standard deviation. This means that informants tend to agree on scores. Confidence interval for the mean at 95% shows that general perception for human translations are over 7. Google translation also performed over 5 and confidence interval for the mean at 95% confirms results. Systran mean is under 5 and confidence interval confirms this perception. Figure 9 compares means among three translators.

HUMAN	GOOGLE	SYSTRAN
Mean: 7.690883	Mean: 6.251543	Mean: 3.656118
Median: 7.684211	Median: 6.392857	Median: 3.842105
St. Dev., s: 0.9101108	St. Dev., s: 1.630817	St. Dev., s: 1.388645
95% CI for the Mean:	95% CI for the Mean:	95% CI for the Mean:
7,2522 < mean <8,1295	5,4655 < mean <7,0376	2,9868 < mean <4,3254

Table 2. Descriptive statistics

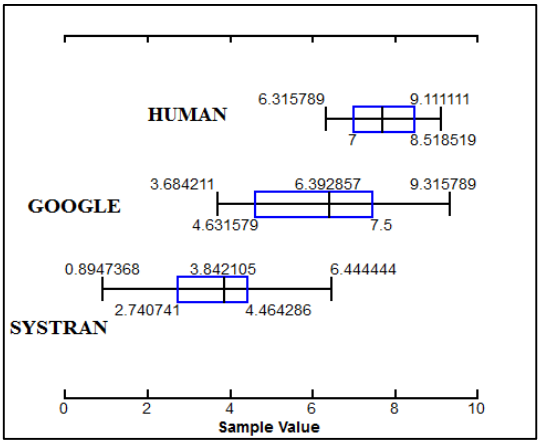

Figure 9. Boxplot of 95% confidence interval for the Mean

An ANOVA revealed a significant main effect of degree of acceptance among the human, Google and Systran translations (F-value (3,1)= 41,4 and p-value = 0,000). Bomferroni test also shows that all means are different from each other. This means that general performance of the three translators shows different grades of acceptability according to data provided by informants. So, according to this work, the evaluation that informants have made on the translations provided by human translators, Google and Systran are significantly different.

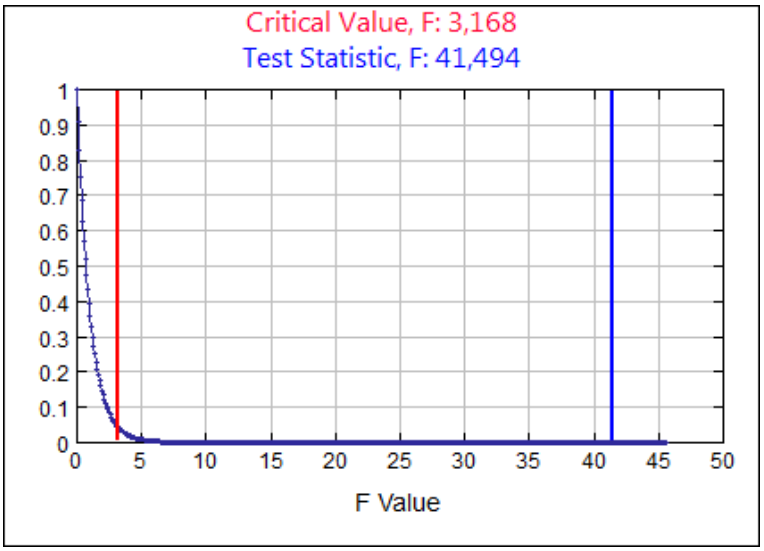

Figure 10. Plot for ANOVA results

5 Discussion

Error! Reference source not found. shows general results achieved by the three translators. As we referred above, general approaches to machine translation quality assessment are based on a binary criterion, that is, grammatical and ungrammatical, in order to judge the quality of the different translated sentences. This is an idealization and it does not show how much human and machine translation engines differ from each other. The assessment methodology proposed in this work assumes that human translations might not be considered as perfect (it achieves 7.690883 mean) and it can be compared to other results. From this perspective, Google is much closer than Systran to human realizations.

Regarding the results, it can be seen that the standard deviation coming from machine translation is higher than that from human translation. It indicates that there is much variability in the automatic translation. We assume that there are phenomena that the engine will work best with and some other with a poor translation in Machine Translation.

We allowed to complete online questionnaire to any subject, however, finally we decided to exclude those informants who were under B2 in English proficiency from general results. **Error! Reference source not found.** shows mean and standard deviation provided by these informants with a lower English level:

HUMAN	GOOGLE	SYSTRAN
Mean 7,43	Mean 7,32	Mean 4,98

Standard deviation 1,43	Standard deviation 1,86	Standard deviation 1,82

Table 3. Descriptive statistics of scores provided by under-B2 informants

As it can be observed, scores provided for human and Google translations look similar. The mean for Systran is 5 for this group of informants, much better than results for those subjects from B2 and higher. A T-test did not reveal a significant main effect of degree of acceptance among the human and Google (F-value (1.9) = 1,1 and p-value = 0.2345), so there is no statistical difference between them. **Error! Reference source not found.** shows results of t-test. This is explained by the fact that informants under B2 do not have a good level of proficiency to perceive adequately style and grammatical differences in sentences in English.

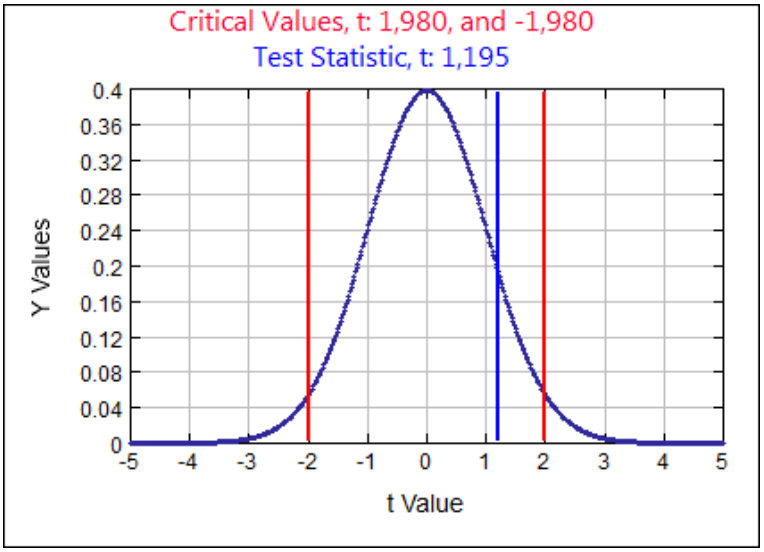

Figure 11. Plot for T-test results: evaluation of Google and Systran translations by under-B2 informants

We also compare general results obtained between expert and non-expert informants.

Expert	**Non-expert**
Mean: 6.007264 Median: 6.357143 St. Dev., s: 2.167579 95% CI for the Mean: 5,4424 < mean <6,5721	Mean: 5.756963 Median: 6.125 St. Dev., s: 2.197708 95% CI for the Mean: 5,1842 < mean <6,3297

Table 4. Descriptive statistic for results provided by expert and non-expert informants

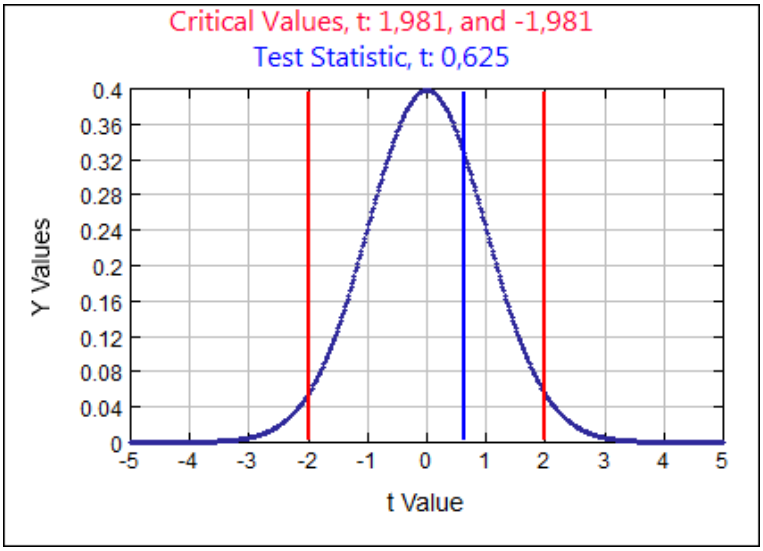

Figure 12. Plot for T-test results between expert and non-expert scores

An Student's t-test did not yield a significant main effect of both expert-based evaluation and not expert-based evaluation (F-value (1.9)= 0.625 and p-value = 0.533) for translated sentences. This difference is considered to be not statistically significant, so both groups give same values to sentences and there are no statistical differences. That means that this method for assessing a translator's general performance provides the same results no matter if informants are professionally related to Linguistics or they are not. Therefore, results show that most important variable for evaluation is language proficiency. This detail can be explained by the fact that gradience-based methodology is not based on the notion of grammaticality, so it can work with no professional expert informants as long as they have a good language proficiency.

6 Conclusions

The main purpose of this paper has been to show the applicability of graded acceptance in translation evaluation, instead of the traditional binary method (right/wrong) based on grammaticality. In order to do this, we have compared the results of the evaluations of several sentences translated from Spanish into English by human translators, Google translator and Systran. The evaluators were Spanish informants with a B2 or higher level of English.

Focusing on the results given by the comparison of translations, statistical analysis revealed that the evaluation that informants have made on the translations provided by human translators, Google and Systran are significantly different. The lowest scores are for Systran translations: only 25,90% of its translations are scored over 5; more than half (66.61%) are considered negatively and 16,99% has been considered totally unacceptable. Google translations are considerable higher: 63.22% over 5 and 28.63% under 5. The highest scores are recorded by human translations: 81.3% of human translations are evaluated over 5 in a likert scale, 50% are scored as 9 or 10; just 13.62% of human translations scored under 5. An ANOVA revealed that general performance of the three translators shows different grades of acceptability according to data provided by informants.

On the other hand, when focusing on the methodology of evaluation, the one we presented here has shown that it is not necessary that evaluators are professional experts in Translation, Linguistics or Philology: nontechnical users –provided that their level of English is B2 or higher– showed the

same scores as technical informants. Informants with a lower level could not see differences between human and Google translations. English proficiency was a determinant factor for the evaluation of Machine Translation.

Finally, a Machine Translation evaluation based on a graded acceptance opens new ways to measure the "real distance" between automatic and manual translations. In addition to that, it could also allow knowing what linguistic errors users consider less acceptable and how they affect the general output of the Machine Translation engine. This issue will be developed in future papers.

References

Abaitua, J. (2002). Tratamiento de corpora bilingües. *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita* , 61-90.

Aiken, M., Vanjani, M. B., & Wong, Z. (2009). Measuring the accuracy of Spanish-to-English translations. *Issues in Information Systems , 7* (2), 125-128.

Alcina Caudet, M. A. (2011). Los traductores automáticos en la red. *El español en el mundo. Cuadernos Cervantes* , ??-??

Alcina, A. (2008). Translation technologies scope, tools and resources. *Target , 20* (1), 79-102.

Arenas, A. G. (2010). Exploring Machine Translation on the Web. *Tradumàtica: traducció i tecnologies de la informació i la comunicació , 8*, 1-6.

Carlisle, D. (2010, April). *graphicx: Enhanced support for graphics.* Retrieved from http://www.ctan.org/tex-archive/ help/Catalogue/entries/graphicx.html

Costa-Jussà, M. R., & Fonollosa, J. A. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language , 32* (1), 3-10.

Erwen, Z., & Wenming, Z. (2013). Application of Computer-Aided Translation Technology in Translation Teaching. *International Journal of Emerging Technologies in Learning (iJET) , 8* (5), 15-20.

Domínguez, M., Laurenti, L., & Céliz, C. (2013). Google Translate: una experiencia con alumnos de inglés técnico en el nivel superior. *Virtualidad, Educación y Ciencia , 4* (6), ¿¿-44.

Gutiérrez, I. C., & Figueroa, D. F. (2011). Traducción automática. Técnicas y aplicaciones.

García, S., Meilán, A., & Martínez, H. (2005). *Construir bien en español: la forma de las palabras.* Oviedo: Universidad de Oviedo.

González Boluda, M. (2010). Estudio comparativo de traductores automáticos en línea: Systran, Reverso y Google. *Núcleo , 22* (27), 187-216.

González Boluda, M. (2010). Estudio comparativo de traductores automáticos en línea: Systran, Reverso y Google. *Núcleo , 27*.

Gómez Manzano, P., & et al. (2005). *Ejercicios de gramática y de expresión. Con nociones teóricas.* Madrid: Centro de Estudios Ramón Areces.

Gómez Torrego, L. (2011). *Hablar y escribir correctamente.* Madrid: Arco/Libros.

Hutchins, W. J. (1995). Machine Translation: A Brief History. In E. F. Koerner, & R. E. Asher, *Concise History of the Language Sciences: from the Sumerians to the Cognitivists* (pp. 431-455). Elsevier.

Keller, F. (2000). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality.* University of Edinburgh.

Kit, C., & Wong, T. M. (2008). Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal , 100* (2), 299-321.

Kliffer, M. D. (2008). Post-editing Machine Translation as an FSL Exercise. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras , 9*, 53-68.

Lacorte, M. (2007). *Lingüística aplicada del español.* Madrid: Arco/Libros.

Lee, J., & Liao, P. (2011). A Comparative Study of Human Translation and Machine Translation with Post-editing. *Compilation & Translation Review , 4* (2), 105-149.

Manning, C. (2002). *Probabilistic Syntax.* Stanford University: Departments of Linguistics and Computer Sciences.

Real Academia Española. (2005). *Diccionario panhispánico de dudas*. Madrid: Santillana.

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Seljan, S., Brkić, M., & Kucis, V. (2011). Evaluation of free online machine translations for Croatian-English and English-Croatian language pairs. *Proceedings of the 3rd International Conference on the Future of Information Sciences: INFuture2011-Information Sciences and e-Society*, (pp. 331-345).