INNOVATION IN LANGUAGE LEARNING
International Conference

Quality in Human and Machine Translation

Nicolás Montalbán[1], Juan Manuel Dato[2]

Centro Universitario de la Defensa, Academia General del Aire, San Javier, Spain[1]
Instituto de Estudios Superiores Carlos III, Cartagena, Spain[2]

Abstract

There have always been long-winded discussions on the role played by both human and MT in quality translation processes. Which one is better? Or, should they be used in combination to achieve a quality translation? The present paper provides an answer to these matters by means of the calculation of several evaluation metrics to study the quality offered by MT compared to human translation. Moreover, there is a implementation of a new tool based upon a reference model text with some indexes including Narrativity, Readability, Referential Cohesion, Deep Cohesion, and Concreteness, which is compared to the translated texts produced by humans. To calculate the evaluation metrics and indexes, chosen samples of scientific and literary texts were included. Mentioned texts were used in two final dissertations in the university course of Translation and Interpreting at the University of Murcia.

Keywords: *Quality in Translation processes, Scientific-technical translation, Literary translation, English for Specific Purposes, Computer-based studies, Linguistics*

1. Introduction

House (2015) starts most of her works with questions such as "What is a good translation?". Quality translation should be mentioned here associated to the goals of MT and new ’interactive’ and/or ’adaptive’ interfaces have been proposed for post-editing (Green, 2015). Therefore, in this case, human and MT are inextricably linked. Some recent studies mention that MT is almost ’human-like’ or that it ’gets closer to that of average human translators’ (Wu et al., 2016) and, also that MT quality is at human parity when compared to professional human translators”. Ahrenberg (2017:1) states that the aim of MT is ‘overcoming language barriers’, although human translation is aimed at producing ‘texts that satisfy the linguistic norms of a target culture and are adapted to the assumed knowledge of its readers’.

Nevertheless, there are authors who claim that it is almost impossible to overcome the perfection of human translation (Giammarresi and Lapalme (2016). MT Translation has gone through three stages ’from early dictionary-matched machine translation to corpus-based statistical computer-aided translation, and then to neural machine translation with artificial intelligence as its core technology in recent years’ (Zhaorong, 2018). House (2018:2) defines translation as ’the result of a linguistic-textual operation in which a text in one language is re-contextualized in another language’. House (2018:5) also insists on the cognitive aspects of translation, and specifically, the process of translation in the translator´s mind; a matter studied over the last 30 years.

Ahikary (2020) states that “the equivalence is one of the most important aspects or goals of translation; translator has to focus on searching for the best equivalent terms between two different languages or dialects”.

2. Methodology

2.1 Materials used in the experiment

To carry out this work, different types of materials were used. First, a collection of texts in English dealing with: Quantum Physics, Technology, Medicine, Environment and Geology, with an extension of 600 words for each one. Then, the second one is an extract from *Red Dirt* (2016), a literary text from the narrative genre. For the MT two different tools were used: Matecat for the scientific-technical texts and Wordfast Anywhere for the literary text. Apart from that, representative texts in Spanish were selected for comparison purposes: a selection of 5 scientific-technical texts from well-known international scientific publications. As far as the literary text, an extract was chosen from the book «Escritos de un viajo indecente» by Bukowski (2006), from the same genre and full of phraseological units, including insults.

2.2 Evaluation metrics for both MT

The first evaluation metrics we are introducing here are Precision and Recall. WER (Word Error Rate) is another metric we are implementing, but the most common metric used is BLEU (Bilingual Evaluation Understudy).

2.3 CAT tools: Matecat and WordFast Anywhere

According to Matecat's site: "Matecat is a free and open source online CAT tool. It is free for translation companies, translators and enterprise users." (Matecat, 2014). du Maine and the University of Edinburgh. In Matecat translation, assignments are organized into projects in which the user specifies the source language and the target language. One project comprises one or several texts to be translated, and each project has a translations memory. Wordfast Anywhere, which is a Translation memory of the company Word have the following procedure: the text is divided into segments that are being translated and stored, creating glossaries and translations, which will appear in future translations depending on the index of coincidence of the words.

2.4 Definition of the tool used to calculate easibility of the text

To analyze the appropriateness of the texts as regards reading, a code in Python language has been developed. The first operation carried out by this code is sequencing words of the text to recover the number of paragraphs, sentences, words and syllables in total, and later, it determines five metrics based on the studies in Coh-Metrix, but simplified.

This new technique is called CohLitheSP since it is based upon Coh-Metrix, and does not need large dictionaries nor corpuses formed by thousands of words to offer consistent results. Furthermore, on the other hand, specific formulae have been introduced for tests written in Spanish, when just a few changes have to be made to adapt it to any language without any extra cost.

To apply the aforementioned metrics, the following are needed:

A reference text conforming to a valid corpus,

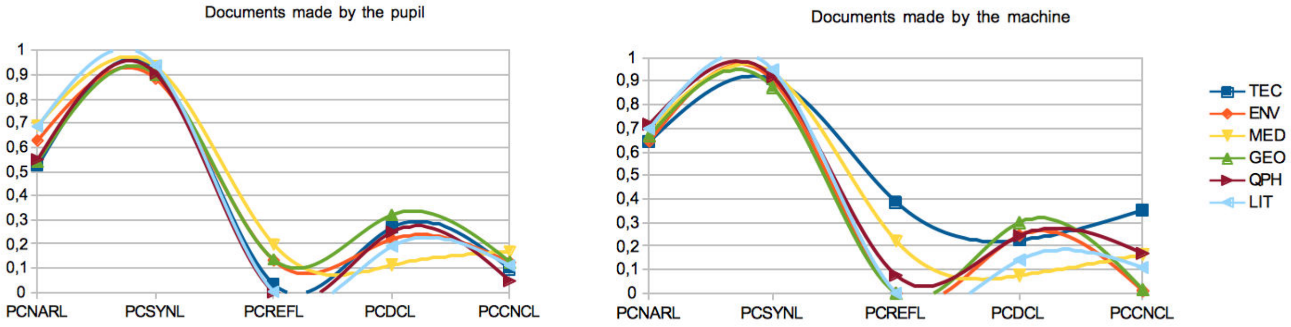
A glossary of technical or specific terms which is helping to know which words are specific within a corpus. These terms will not include measurement units nor "words of stop" (prepositions, determiners, etc), and

A set of connectors allowing to know when, in a sentence, something is being inferred from something previously said.

The selected metrics and their changes are:

- PCNARL. Narrativity. It is calculated determining which words of the text to be evaluated are already being recognized in the reference text.
- PCSYNL. Readability. It determines the simplicity of the text in its language. In the case of Spanish, the readability of Fernández (1959) has been chosen (based on Flesch), which is using a number of sentences, syllables and words. If someone wants to do it for the English language, it only needs to be changed with the Flesch-Kincaid1, whose formula is also based on a similar calculation.
- PCREFL. Referential Cohesion. In this version, the same referential cohesion as in Coh-Metrix is calculated; but instead of considering all nouns, it is only applied in technical or specific terms recognized in the glossary.
- PCDCL. Deep Cohesion. It determines the incidence of the connector over the recognized sentences.
- PCCNCL. Concreteness. In this version, instead of calculating the concreteness over the whole corpus of the language, the incidence of the terms of the glossary is determined from the recognized words in the reference text within the text to be evaluated.

After applying this simplified version of Coh-Metrix over the produced texts in Spanish, it is possible to see how, after being evaluated separately with a mark from 0 to 10, they seem to describe a similar curve:



Tables 1 and 2. Students' and MT documents

As can be seen in the above figures, different types of written texts for different technical corpuses seem to be minor differences in marks, but with a pattern that seems to say that measurements are not random. Therefore, it seems that, in addition, the texts used as references, representing a corpus without errors, have a mark below below 10 so students can never get that mark. Therefore, not only must each Coh-Lithe metric be weighted in such a way that favours the distinction among students' faculties, but, in addition, the results must be amplified so the reference texts have the same mark. For this reason, now there is an explanation on how to calculate the weighting of each metric and the constant used to amplify the mark.

2.5 Calculation of the amplification constant for each specific corpus

Below, the results of evaluating the reference texts can be seen.

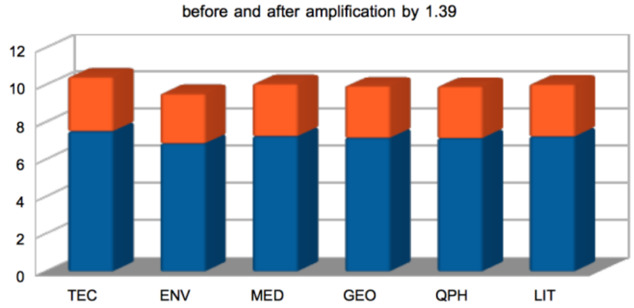


Table 3. Marks of reference texts

Due to the fact that reference texts have a mark below 10 (as it can be observed in Figure 3 in blue bars, after applying an amplification of 1.39, the results would be near 10. To be able to calculate a specific amplification to the text belonging to its corpus, the following formula could be applied:

$$K_{TEC} = \frac{10}{\frac{PCNARL^{REF}_{TEC}}{1000} \cdot 0.49 + \frac{PCSYNL^{REF}_{TEC}}{206.82} \cdot 0.2 + \frac{PCREFL^{REF}_{TEC}}{1000} \cdot 0.09 + \frac{PCDCL^{REF}_{TEC}}{1000} \cdot 0.17 + \frac{PCCNCL^{REF}_{TEC}}{1000} \cdot 0.05}$$

2.6 Calculation of marks of easibility of texts

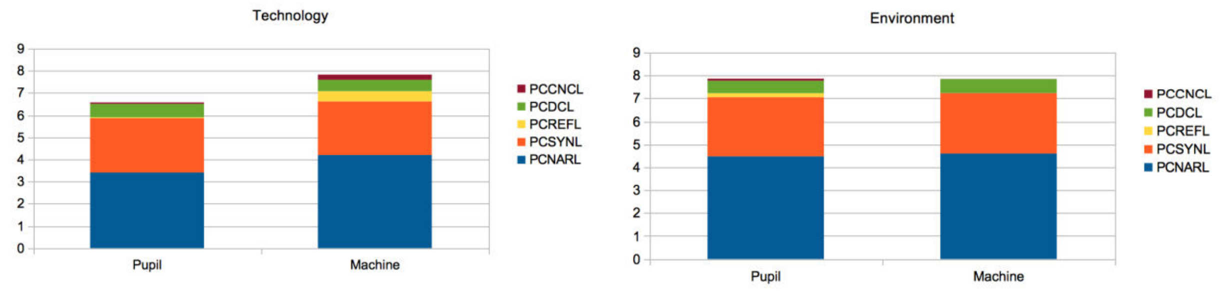
Regarding the calculation of the marks of the texts, the amplification constant must be applied by the addition of each metric divided by its maximum and multiplied by its weight. For example, the following formula can be observed over the technology texts:

$$Score^{PUPIL}_{TEC} = K_{TEC} \cdot \left(\frac{PCNARL^{PUPIL}_{TEC}}{1000} \cdot 0.49 + \frac{PCSYNL^{PUPIL}_{TEC}}{206.82} \cdot 0.20 + \frac{PCREFL^{PUPIL}_{TEC}}{1000} \cdot 0.09 + \frac{PCDCL^{PUPIL}_{TEC}}{1000} \cdot 0.17 + \frac{PCCNCL^{PUPIL}_{TEC}}{1000} \cdot 0.05 \right)$$

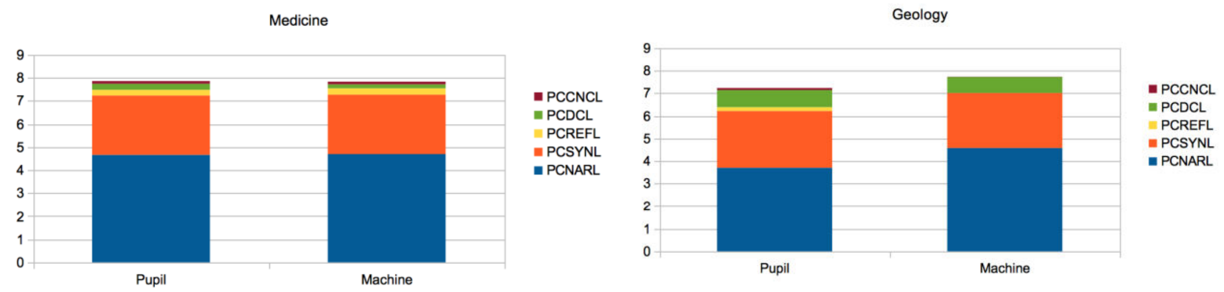
3. Results

3.1 Calculation of marks of easibility of texts

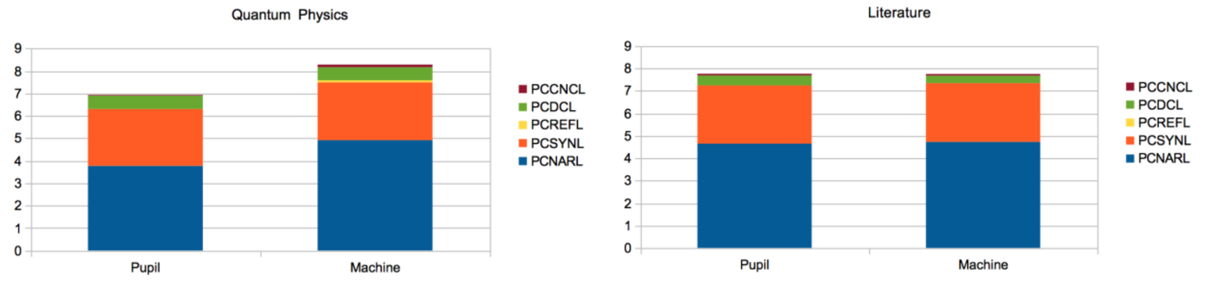
After applying the corresponding formulas already described above, the following results are achieved:



Tables 4 and 5. Evaluation amplified by its reference (Technology and Environment)



Tables 6 and 7. Evaluation amplified by its reference (Medicine and Geology)



Tables 7 and 8. Evaluation amplified by its reference (Quantum Physics and Literature)

4. Conclusions

In this work, a new and different tool has been shown which adds a supplementary challenge for students: the possibility of improving the readability of their own translations from English into Spanish. Given the facts, the technique explained before is working properly mainly due to two results: on the one hand, it is proved that different texts coming from different typologies, including MT texts, get good or bad marks in the same metrics. On the other hand, the tests also show that, after refining the final mark, the result is approximate to a student's evaluation.

Moreover, it is important to stress the easy programming, which does not require large corpuses, despite the fact it comes from systems needing an enormous extra charge in the development of programming. This last feature is complemented by the fact that it is easily transformed to be working in any language.

5. Software

The programme written in Python used to calculate the statistics with commentaries in English can be found in the following address: https://archive.org/details/coh-lithe-sp-012

References

[1] Adhikary, R. P. "Degrees of equivalence in translation – a case study of Nepali novel 'seto bagh' Adhikary, R. P. "Degrees of equivalence in translation – a case study of Nepali novel 'seto bagh' into English as 'the wake of the white tiger'", *European Journal of Multilingualism and Translation Studies*, [S.l.], v. 1, n. 1, May 2020. Available at: <https://oapub.org/lit/index.php/EJMTS/article/view/173>. Date accessed: 29 October 2020

[2] Ahrenberg, L. “Comparing machine translation and human translation: A case study”, In Irina Temnikova, Constantin Orasan, Gloria Corpas and Stephan Vogel (eds), RANLP 2017 The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT) Proceedings of the Workshop, September 7th, 2017; 2017, pp. 21-28.

[3] Alvarez Flórez, J.M. & Pérez. “Escritos de un viejo indecente”, Anagrama, 2006.

[4] Fernández Huerta J. “Medidas sencillas de lecturabilidad”, Consigna 1959; (214): 29-32.

[5] Giammarresi, S., & Lapalme, G. “Computer science and translation: Natural languages and machine translation”, DOI:10.1075/BTL.126.10GIA

[6] Green, S. “Beyond post-editing: Advances in interactive translation environments”, *ATA Chronicle* www.atanet.org/chronicle-on-line/, Volume 64, Issue 3,
p. 490 – 494. DOI: https://doi.org/10.1075/babel.00047.kan, 2015

[7] House, J. “Translation Quality Assessment: Past and present”, Routledge, 2015

[8] Matecat.
https://site.matecat.com/benefits/?gclid=Cj0KCQjw6uT4BRD5ARIsADwJQ1-s67PUj9ENqpo1g9Yl5kP2SQyfikdDv3DU_dHUMG-rxM2J_XsFG6QaAk0yEALw_wcB, 2014

[9] Raepy, E.M. “Red Dirt”, Head of Zeus, 2016

[10] Wu, Y.; Schuster, M.; Chen, Z.; V. Le, O.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. “Google’s neural machine translation system: Bridging the gap between human and machine translation”, *CoRR* abs/1609.08144. Http://arxiv.org/abs/1609.08144., 2016

[11] Zhaorong Zong “Research on the Relations Between Machine Translation and Human Translation”, Journal of Physics: Conference Series, Volume 1087, Issue 6, 2018