DOI: https://doi.org/10.5007/2175-7968.2023.e85397

CHALLENGING MACHINE TRANSLATION ENGINES: SOME SPANISH-ENGLISH LINGUISTIC PROBLEMS PUT TO THE TEST

Argelia Peña Aguilar[1, 2]
[1]University of Ottawa
[2]Universidad Autónoma del Estado de Quintana Roo

Abstract: This work is an evaluation of machine translation engines completed in 2018 and 2021, inspired by Isabelle, Cherry & Foster (2017), and Isabelle & Kuhn (2018). The challenge consisted of testing MTs Google Translate and Bing and DeepL in the translation of certain linguistic problems normally found when translating from Spanish into English. The divergences representing a "challenge" to the engines were of morphological and lexical-syntactical types. The absolute winner of the challenge was DeepL, in second place was Bing from Microsoft, and Google was the engine that was the poorest in the management of the linguistic problems. In terms of time, when comparing the engines three years apart, it was found that DeepL was the only one that enhanced its performance by correcting a problem it had before in a test sentence. This was not the case for the other two, on the contrary, their translations were of lower quality. These machines do not seem to be consistent in the manner in which they are improved. These findings may be valuable for translators who may work with these systems as pre or post-editors so that their efforts may be better directed.
Keywords: Machine Translation; Pre-editing; Post-editing; Google Translate; Bing; DeepL

Introduction

"Machine Translation Systems are perhaps the electronic translation tools that attract the most public attention, especially among

non-translators" (Austermühl & Kortenbruck, 2001, p. 153); and it seems they will keep on drawing the attention of all kinds of users, but most particularly, that of translators/language professionals. The improvements made each time to these engines are making them superior to their previous version and these enhancements will only make them more acceptable for professional use.

Besides the upgrades, regular evaluations should also be made to test their performance in different situations. These are usually done not only by the developers of such software, but also by the people who use (or want to use) these systems. There are different approaches to assess the quality of machine translation (MT), but other considerations may be relevant:

- Reasons to use it (communication, publication, "gisting" or enabling meaning)
- Standard of quality ("professional", "human parity", "fit-for-purpose", "good enough")
- Evaluators (developers, translators, professors, students)
- Consequences of quality expectation (direct and indirect, short-long term, stakeholders, entities)
- Aspect being evaluated (a sentence at a time, a paragraph, specific linguistic features)
- Other factors (type of MT, domain, text type, language pair)
- User acceptance (preference, use, perceptions)
- Automatic metrics vs. human measures (Marshman, 2018, p. 3-17).

From an engineering perspective, Philipp Koehn in his book *Neural Machine Translation* (2020), provides an account of the myriad of methods to evaluate the progress of machine translation over time and the quality of the translation it produces. Stakeholders in language industry, computational linguistics, engineering, and translation companies, consider these to be, 'best practices.'

Koehn classifies the forms of MT assessment into three:

1. Task-based evaluation (which include real-world tasks, content understanding and translator productivity).
2. Human assessment (adequacy and fluency; ranking; continuous scale; crowd-sourcing evaluations; human translation edit rate).
3. Automatic metrics (BLEU, The Meteor metric, TER, characTER, and Bootstrap Resampling).

Incorporating Human evaluation is absolutely a paramount method for testing MT engines, as can be noticed from the above-mentioned considerations.

Testing MT engines has been a very fertile research area for some years, and there are some relevant studies in Human assessment and post-edition which relate to translators work. Brita Banitz (2020), for instance, used Human assessment along with TER score measure to evaluate rule-based and statistical machine translation and explained in an essay their performance with English-German phrases taken from Mark Twain's, *The Awful German Language*.

In a similar vein, the relevant work done by Coraline Doan (2021) in her thesis titled, *Comparing Encoder-Decoder Architectures for Neural Machine Translation: A Challenge Set Approach,* focuses on Human evaluation of MT engines from a translator's perspective. She designed the methodology inspired by the precursors of challenge sets (as in this paper): Isabelle, Cherry & Foster (2017). The set was for English to French MT translations and employed Jean Delisle and Marco Fiola's, *La traduction raisonnée* (3rd edition) to that end.

In another thought-provoking research, Guerberof-Arenas & Toral (2020, p. 254), asked readers to evaluate three different translations of a fictional story from English into Catalan. These were in three forms: machine translation, post-edited, and translated without aid (human translation). The creativity of those translations were evaluated from the readers' viewpoint and, as might be expected, the creativity was reported to be the highest when translators were involved in the process.

On the subject of post-edition studies, Parra Escartín & Goulet (2021), did an interesting enquiry on post-edition for publication purposes, but not performed by professional translators. Their aim was, "to determine whether the physician-participants would be in a position to submit research papers for publication using a general machine-translation engine followed by post-editing" (Parra Escartín & Goulet, 2021, p. 91). The results indicated that the quality of such post-edition would not be good enough for the said purpose, which made clear that the MT versions would have to be post-edited by a language professional.

As the previous examples suggest, human involvement in the machine translation process, be it as an evaluator, or a post-editor (evidently as a translator, too) bring about better outcomes.

Our approach

Bearing in mind that, "even with ongoing automation in many aspects of translation service, revision and post-editing rely on human skill and expertise" (Konttinen, Salmi & Koponen, 2021), we believe it is crucial that human translators are able to assess MT engines by themselves and learn from other experiences on how to achieve this. To this end, the emphasis of this work revolves around MT testing and assessment by a human translator-the author.

The evaluation of machine translation engines for this paper will be examined considering the previous research of Pierre Isabelle, Colin Cherry & George Foster (2017) in their study entitled, "A Challenge Set Approach to Evaluating Machine Translation", and the investigation conducted by Pierre Isabelle & Roland Kuhn (2018), named, "A Challenge Set for French → English Machine Translation". Both studies set the course of action for this paper as described below.

The first challenge was completed on November 25, 2018, (the second, three years later in 2021), and consisted of testing MTs Google Translate, Bing and DeepL for the translation of certain

linguistic problems normally found when translating from Spanish into English. These divergences were thought to represent a "challenge" to the engines and, without doubt, the findings would aid in determining the quality of the MTs assessed.

To be able to elucidate the examples, and subsequently, the results in the next chapter, the writer has presented tables with all these linguistic problems. The three machines evaluated for this task are web-based and the belief is that they produce high-quality language translations. Google Translate and DeepL have neural architecture, but Bing seemed to use a statistical approach for the free version, although the company that produces it, (Microsoft), has announced advances in their neural version and probably has already released it. No matter what their design is, we expect interesting results.

Each table displays a particular problem and a single test sentence. There are examples of the performance of the three engines that are part of this challenge and the human evaluation is done by using a ✓ or a ✗ on the right side of the machine translations. Also, the adequate sentences are in bold for easier identification. The assessment was based on the proximity to the reference translation provided for each test sentence and the way these MTs handle the linguistic problem in turn was also considered. Unlike the other two, DeepL is the only MT that provided more than one example, and all of them were included in the chart. If one of the options was the right one, then it was considered correct. Why? Because the machine was providing "options" to the translator, who was ultimately the one who would select the appropriate one accordingly. In the beginning (and at the end) of the Results and Discussion section, there will be a summary table (years 2018 and 2021) of the performance of Google Translate, DeepL and Bing which visually helps to identify trends or general execution of these engines.

Challenge set

The following evaluation was done by the author essentially considering the performance of the engines regarding the linguistic phenomena presented. Ultimately, the author will give further suggestions taking into account the MTs feasibility after the challenge experience. Also, to clarify, the sentences used for this task were designed specifically for it. Therefore, they are intentionally short and focused on phenomena that the author has found to be challenging for human translators when translating from Spanish into English. The question here is to determine how well Google Translate, Bing and DeepL handle these same problems, in order to resolve their quality in execution.

The challenge set consisted of five language structures and are categorized into two types, morphological and lexical-syntactical. These constructions are typical or standard in the source language but unusual in the target language and for this reason they can become challenging to translate for the selected engines. A brief explanation about how different every single language construction is in Spanish and English will be in every table where there is an evaluation of the linguistic problem.

For each language structure three example sentences in the source language and reference (correct human) translations are provided. Lastly, three different machine-translated versions of the sentences are also included. In this way, the corpus will consist of 15 Spanish sentences, 15 English (human) reference translations and 45 machine translations of the source-language sentences (3 times 15 sentences).

A summary of the 15 test sentences is featured below (the machine translations are not included):

Morphological type

1. Sudden proposals in present tense to future statements

Spanish Test sentences	English Reference translations
Te llevo la maleta	→ I'll carry that case for you
¿Se lo envuelvo?	→ Shall I wrap it for you?
Yo lavo los platos hoy	→ I'll wash the dishes today

2. Present tense to present continuous for future statements

Spanish Test sentences	English Reference translations
Me voy mañana a Paris	→ I am leaving/going to Paris tomorrow
¡Te casas pronto!	→ You are getting married soon!
Salimos de viaje en una hora	→ We are leaving on a trip in an/ one hour

3. Inalienable possession

Spanish Test sentences	English Reference translations
El pelo le llega a los hombros	→Her hair falls just to her shoulders
¿Te cepillaste el cabello con cuidado?	→Did you brush your hair carefully?
La mujer ladeó un poco la cabeza	→The woman tilted her head a little

4. Definite article in Spanish to zero article in English

Spanish Test sentences	English Reference translations
¿Qué es la inmortalidad?	→ What is immortality?
Agradezco a la vida lo que tengo	→ I thank life for what I have
La política otorga poder a unos pocos	→ Politics gives power to a few

Lexical-syntactical type

1. Countable vs. Uncountable nouns

Spanish Test sentences	English Reference translations
Compré dos muebles para la sala	→ I bought two pieces of furniture for the living room
Tuvimos un clima agradable el mes pasado	→ We had nice weather last month
Dame consejos para ser mejor	→ Give me (some) advice on how to be a persona better person

Results and discussion

The 2018 assessment will start with Table 1, which presents the overall performance of the MTs that were part of this challenge. Some interesting divergences will be pinpointed and analyzed each at a time, and lastly, the general performance for year 2021 will be provided so as to compare their execution in a somewhat diachronic manner.

Table 1: Overall Performance 2018

Morphological type													Lexical-syntacticaltype		
Linguistic Problems	1.Sudden proposals			2. Present tense to present continuous			3. Inalienable possession			4.Definite article to zeroarticle			5.Countable vs. uncountable nouns		
Testsentences	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
GoogleTranslate	✓	✓̸	✓̸	✓	✓̸	✓̸	✓̸	✓	✓	✓	✓	✓	✓̸	✓̸	✓
DeepL	✓	✓	✓̸	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓̸	✓
Bing	✓	✓	✓̸	✓	✓	✓̸	✓̸	✓	✓̸	✓	✓	✓	✓̸	✓̸	✓

Source: Author.

As can be seen from Table 1, the absolute winner of the challenge was DeepL with thirteen (out of fifteen) adequate translations. The two linguistic problems in which DeepL was not good enough were the two most difficult to handle for the other two engines as well. In second place comes Bing from Microsoft with 9 acceptable translations and 6 inadequate ones. Last in order of performance is Google Translate with 8 appropriate translations and 7 wrong suggestions. We then may confirm that Google is the engine which management of the linguistic problems at hand was the poorest. Now let us look at each case more closely.

The first linguistic problem is of a morphological type (the first four belong to this same category) and it was one of the two problems in which all MTs couldn't handle an accurate translation for a particular sentence: Sudden proposals in present tense to future statements. This one is about unexpected proposals in present tense normally uttered by Spanish speakers, and the corresponding English version should resort to the use of future forms *will* or *shall*, otherwise the resulting options sound odd or atypical. For this problem, there were three test sentences.

The first one was *Te llevo la maleta*, and the translation proposed for this one was "I'll carry that case for you" (see Table 2). All of the MTs translated this one accurately, they just changed the word "case" for "suitcase" (which are the same), but the verb tense was in simple future, as expected.

Table 2: Problem 1 S1: Sudden Proposals in Present Tense to Future Statements

Problem 1 Sentence 1	Sudden proposals in present tense to future statements (morphological type)	Evaluation
In actions that are suddenly proposed and the speaker also seeks for approval, present tense is used in Spanish; the tendency is to avoid present tense in English and use future forms Will or Shall.		
Source sentence	Te llevo la maleta	
Reference translation	I'll carry that case for you	
Google Translate	I'll take your suitcase	✓

DeepL	I'lltakeyoursuitcase,I'lltakeyourbag,I'llbringyour suitcase	✓
Bing	I'll take your suitcase	✓

Source: Author.

The second example was ¿se lo envuelvo?, which corresponding reference translation was, "Shall I wrap it for you?" (Table 3, below). Here Google Translate was the only one which mistranslated that phrase, using simple present tense for the question, e.g. "Do I wrap it?" which does not sound like a proposal or sudden offer in English. DeepL had it right with the three options proposed, the same as Bing.

Table 3: Problem 1 S2: Sudden Proposals in Present Tense to Future Statements

Problem 1 Sentence 2	Suddenproposalsinpresenttensetofuturestatements (morphological type)	Evaluation
Inactionsthataresuddenlyproposedandthespeakeralsoseeksapproval,Spanishuses presenttense:thetendencyistoavoidpresenttenseinEnglishandusefutureformsWillor Shall.		
Sourcesentence	¿Se lo envuelvo?	
Referencetranslation	Shall I wrap it for you?	
GoogleTranslate	Do I wrap it?	✗
DeepL	ShallIwrapitforyou?,ShallIwrapit?,ShallIwrap it up?	✓
Bing	Should I wrap it up?	✓

Source: Author.

The third example was *Yo lavo los platos hoy*, for which the reference translation was "I'll wash the dishes today" (Table 4). Interestingly, the three engines failed to provide an accurate translation and provided options in simple present tense. I believe this happened due to the "tricky" word "today", employed in this last sentence. However, it was used as it was the only way in which near future could be expressed while still using present tense in Spanish. Apparently, the engines did not discern this clue.

Table 4: Problem 1 S3: Sudden Proposals in Present Tense to Future Statements

Problem 1 Sentence 3	Sudden proposals in present tense to future statements (morphological type)	Evaluation
In actions that are suddenly proposed and the speaker also seeks for approval, present tense is used in Spanish; the tendency is to avoid present tense in English and use future forms Will or Shall.		
Source sentence	Yo lavo los platos hoy	
Reference translation	I'll wash the dishes today	
GoogleTranslate	I wash the dishes today	✗
DeepL	I do the dishes today, I wash the dishes today, I'm doing the dishes	✗
Bing	I wash the dishes today	✗

Source: Author.

Problem number 2 focusses on another verb tense change: this time it is present tense to present continuous for future statements. The first statement *Me voy mañana a Paris* was proposed to be translated as, "I am leaving/going to Paris tomorrow" (Table 5). All the engines did well for this first example.

Table 5: Problem 2 S1: Present Tense to Present Continuous for Future Statements

Problem 2 Sentence 1	Present tense to present continuous for future statements (morphological type)	Evaluation
Spanish expresses future with the simple present tense but in English present continuous is the correct tense to express near future for similar statements.		
Source sentence	Me voy mañana a Paris	
Reference translation	I am leaving/going to Paris tomorrow	
GoogleTranslate	I am going to Paris tomorrow	✓
DeepL	I am going to Paris tomorrow (just 1 option)	✓
Bing	I am going to Paris tomorrow	✓

Source: Author.

As for the second, ¡Te casas pronto! / "You are getting married soon!" (Table 6), Google system failed to provide the example in present continuous and used simple present instead. DeepL and Bing did better.

Table 6: Problem 2 S2: Present Tense to Present Continuous for Future Statements

Problem 2 Sentence 2	Present tense to present continuous for future statements (morphological type)	Evaluation
Spanish expresses the future with the simple present tense but in English present continuous is the correct tense to express near future for similar statements.		
Source sentence	¡Te casas pronto!	
Reference translation	You are getting married soon!	
Google Translate	You get married soon!	✗
DeepL	You're getting married soon!, get married soon!	✓
Bing	You're getting married soon!	✓

Source: Author.

Source sentence 3 (in Table 7), *Salimos de viaje en una hora* which reference translation was, "We are leaving on a trip in an/ one hour" had two unsatisfactory translations, by Bing and Google, which resorted to simple past tense rather than present continuous. Their sentences had no sense whatsoever. DeepL gave a suitable translation that used simple present as the source text and ended up being a good choice. However, that was not the expected answer since present continuous was originally believed to be appropriate. In spite of this failure, DeepL proved to be "smart" enough to employ a different verb tense and still get it right.

Table 7: Problem 2 S3: Present Tense to Present Continuous for Future Statements

Problem 2 Sentence 3	Present tense to present continuous for future statements (morphological type)	Evaluation
Spanish can express the future with simple present tense but in English present continuous is the correct tense to express near future for similar statements.		

Source sentence	Salimos de viaje en una hora	
Reference translation	We are leaving on a trip in an/one hour	
GoogleTranslate	We went on a trip in one hour	✗
DeepL	We leave in an hour (just 1 option)	✓
Bing	We went on a trip in an hour	✗

Source: Author.

Problem 3 is about Inalienable possession, which is the use of definite articles in Spanish to refer to parts of the body, whereas English uses the possessive adjectives instead. The first example, on Table 8, *El pelo le llega a los hombros* / "Her hair falls just to her shoulders" turned out to be difficult to translate for the three machines. Both Google and Bing failed to use possessive adjectives and produced unacceptable translations. DeepL had a partial correct translation as it used possessive in the second part of the sentence, but not in the first one. They considered, "Her hair" to be a correct answer for the first part of the statement, but as there was no previous reference, the machines interpreted *El pelo* as a general concept and did not use any possessive element here.

For the second part of the statement, DeepL provided two options: "his shoulders" and "her shoulders" as there was no specification of the gender of the subject. A half point was given/ earned for this MT.

Table 8: Problem 3 S1: Inalienable Possession

Problem 3 Sentence 1	Inalienable possession (morphological type)	Evaluation
Where in Spanish we use definite articles to refer to parts of the body, English uses the possessive for all the references to parts of someone's body.		
Source sentence	El pelo le llega a los hombros	
Reference translation	Her hair falls just to her shoulders	
GoogleTranslate	The hair reaches the shoulders	✗
DeepL	The hair reaches his shoulders, the hair reaches her shoulders	✓
Bing	The hair comes to the shoulders	✗

Source: Author.

Example 2, (Table 9) the engines translated this easily and all provided accurate translations. The source sentence, ¿Te cepillaste el cabello con cuidado?, was translated as "Did you brush your hair carefully?" as originally proposed in the reference translation.

Table 9: Problem 3 S2: Inalienable Possession

Problem 3 Sentence 2	Inalienable possession (morphological type)	Evaluation
WhereinSpanishweusedefinitearticlestorefertopartsofthebody,Englishusesthe possessive for all the references to parts of someone's body.		
Sourcesentence	¿Te cepillaste el cabello con cuidado?	
Referencetranslation	Did you brush your hair carefully?	
GoogleTranslate	Did you brush your hair carefully?	✓
DeepL	Did you brush your hair carefully?	✓
Bing	Did you brush your hair carefully?	✓

Source: Author.

Lastly, source sentence number 3 was, *La mujer ladeó un poco la cabeza* and the proposal for translation was, "The woman tilted her head a little", for which the three systems were expected to have a satisfactory performance, nevertheless, Bing was unsuccessful in this one (see Table 10). No clear explanation can be given, except for the possibility that this MT (Bing) cannot cope with the identification of possessive for the third subject as it failed in examples 1 and 3, for which he/she as subject was used.

Table 10: Problem 3 S3: Inalienable Possession

Problem 3 Sentence 3	Inalienable possession (morphological type)	Evaluation
WhereinSpanishweusedefinitearticlestorefertopartsofthebody,Englishuses possessive for all the references to parts of someone's body.		
Sourcesentence	La mujer ladeó un poco la cabeza	
Referencetranslation	The woman tilted her head a little	
GoogleTranslate	The woman tilted her head a little	✓

DeepL	The woman tilted her head a little	✓
Bing	The woman cocked a little head	✗

Source: Author.

The last of the morphological problems, number 4, was concerned with the use of the definite article in Spanish for a general concept, whereas in English the zero article is normally employed. The first example, ¿Qué es la inmortalidad?, was translated correctly by the machines (on Table 11). Their answer omitted the use of 'la', the article 'the', as it was done in the reference translation: "What is immortality?"

Table 11: Problem 4 S1: Definite Article in Spanish to Zero Article in English

Problem 4 Sentence 1	Definite article in Spanish to zero article in English for general concepts (morphological type)	Evaluation
In English, the use of articles is avoided in general concepts like Life, Immortality, Resurrection, among others, whereas in Spanish definite articles should be utilized.		
Source sentence	¿Qué es la inmortalidad?	
Reference translation	What is immortality?	
Google Translate	What is immortality?	✓
DeepL	What is immortality?	✓
Bing	What is immortality?	✓

Source: Author.

Google, DeepL and Bing had the same performance throughout the other two examples (Tables 12 and 13), so this particular problem did not challenge the MTs at all.

Table 12: Problem 4 S2: Definite Article in Spanish to Zero Article in English

Problem 4 Sentence 2	Definite article in Spanish to zero article in English for general concepts (morphological type)	Evaluation

InEnglish,theuseofarticlesisavoidedingeneralconceptslikeLife,Immortality, Resurrection,amongothers,whereasinSpanishdefinitearticleshouldbeutilized.		
Sourcesentence	Agradezco a la vida lo que tengo	
Referencetranslation	I thank life for what I have	
GoogleTranslate	I thank life what I have	✓
DeepL	IthanklifeforwhatIhave,IthanklifeforwhatI've got	✓
Bing	I thank life what I have	✓

Source: Author.

Table 13: Problem 4 S3: Definite Article in Spanish to Zero Article in English

Problem 4 Sentence 3	DefinitearticleinSpanishtozeroarticleinEnglishfor general concepts (morphological type)	Evaluation
InEnglish,theuseofarticlesisavoidedingeneralconceptslikeLife,Immortality, Resurrection,amongothers,whereasinSpanishdefinitearticleshouldbeutilized.		
Sourcesentence	La política otorga poder a unos pocos	
Referencetranslation	Politics gives power to a few	
GoogleTranslate	Politics grants power to a few	✓
DeepL	Politicsempowersafew,politicsgivespowertoafew	✓
Bing	Politics gives power to a few	✓

Source: Author.

Problem number 5, which was a lexical-syntactical type, was concerned with divergences in the use of countable nouns in Spanish and their corresponding uncountable nouns in English. In example 1, on Table 14, systems Google and Bing failed to provide an adequate translation for *Compré dos muebles para la sala* into English. DeepL provided an acceptable translation by quantifying the non-countable noun "furniture": "I bought two pieces of furniture for the living room".

Table 14: Problem 5 S1: Countable vs. Uncountable Nouns

Problem 5 Sentence 1	Countable vs. uncountable nouns (lexical-syntactical type)	Evaluation
InSpanishsomenounscanbecountedorpluralized,butinEnglishsomeofthese samenounsshouldalwaysbeinsingularformorarenotnaturallycountable,so expressions of quantity have to be added.		
Sourcesentence	Compré dos muebles para la sala	
Referencetranslation	Iboughttwopiecesoffurniturefortheliving room	
GoogleTranslate	I bought two furniture for the living room	✗
DeepL	Iboughttwopiecesoffurnitureforthelivingroom	✓
Bing	I bought two furniture for the living room	✗

Source: Author.

For source sentence 2 (Table 15), *Tuvimos un clima agradable el mes pasado*, none of the systems proposed accurate translations as they all used a quantifier for the word "weather", which is an uncountable noun in English. The reference translation was "We had nice weather last month". The last sentence was easier to handle by the three engines.

Table 15: Problem 5 S2: Countable vs. Uncountable Nouns

Problem 5 Sentence 2	Countable vs. uncountable nouns (lexical-syntactical type)	Evaluation
InSpanishsomenounscanbecountedorpluralized,butinEnglishsomeofthese samenounsshouldalwaysbeinsingularformorarenotnaturallycountable,so expressions of quantity have to be added.		
Sourcesentence	Tuvimos un clima agradable el mes pasado	
Referencetranslation	We had nice weather last month	
GoogleTranslate	We had a nice climate last month	✗
DeepL	We had a nice weather last month	✗
Bing	We had a nice weather last month	✗

Source: Author.

On Table 16, the source sentence is *Dame consejos para ser mejor persona*, for which the reference translation was, "Give me

(some) advice on how to be a better person", considering that the word "some" is optional. Interestingly, Bing supplied a different word to "advice" and resorted to "tips" as in a way to quantify the noun similarly as in Spanish. Overall, the systems do not seem to cope consistently well with the differences in use of countable and uncountable nouns for the Spanish-English language combination.

Table 16: Problem 5 S3: Countable vs Uncountable Nouns

Problem 5 Sentence 3	Countable vs. uncountable nouns (lexical-syntactical type)	Evaluation
In Spanish some nouns can be counted or pluralized, but in English some of these same nouns should always be in singular form or are not naturally countable, so expressions of quantity have to be added.		
Source sentence	Dame consejos para ser mejor persona	
Reference translation	Give me (some) advice on how to be a better person	
Google Translate	Give me advice to be a better person	✓
DeepL	Give me advice on how to be a better person	✓
Bing	Give me tips to be a better person	✓

Source: Author.

In general, we need to give some recognition to these machines for their performance, especially to DeepL, which did surprisingly well for most of the test sentences utilized for this challenge. As can be seen from Table 1 (on page 8), seven of the fifteen sentences were not difficult to handle by these engines; they all had the translation right for these seven. That is roughly 46% of the total amount. The most difficult linguistic problems were, 'Sudden proposals,' and 'Countable vs. uncountable nouns,' morphological and lexical types of problems respectively.

Other difficult aspects to handle for these MTs were the change in verb tense, present tense to present continuous (problem 2), and inalienable possession (problem 3), especially for Google and Bing.

2021 Assessment

In order to update and compare performance of these engines over time, the author completed the same exercise three years later (on December 20, 2021). It is usually stated that machine translation improves every year, but after challenging them with the same set of phrases and using the same considerations for assessment, interesting results came out. Below, in table 17, there is a comparison of their performance including that of 2018 on the left side of each box and the 2021 results on the right side. This display of results was the most appropriate manner to identify the evolution of execution.

In the case of Google Translate, its performance worsened substantially and proved to be the least improved engine in regard to the linguistic problems analyzed. It even got wrong what it had got correct in a set of three test sentences used three years ago, such as the use of definite article to zero article.

Bing's performance was pretty much the same in terms of numbers and only had one additional, inadequate sentence (see Table 18). However, some type of sentences that were acceptable in 2018 were now unacceptable, such as in, 'Sudden proposals.' By contrast, it became better at Countable vs. uncountable nouns in test sentences. This tells us that these machines do not seem to be very consistent in the manner in which they are improved.

DeepL is the only one that enhanced its performance by only having one inadequate sentence (in Sudden proposals). Three years ago it had two unsatisfactory results. We can only deduce that DeepL maintained its high quality level compared with the other two.

Table 17: Compared Overall Performance 2018-2021

Morphological type													Lexical-syntactical type		
Linguistic Problems	1.Sudden proposals			2.Presenttense to present continuous			3.Inalienable possession			4.Definite articletozero article			5.Countablevs. uncountable nouns		
Test sentences	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Google Translate	✓Ø	ØØ	ØØ	✓✓	ØØ	ØØ	ØØ	✓Ø	✓✓	✓✓	✓Ø	✓✓	ØØ	ØØ	✓Ø
DeepL	✓✓	✓✓	ØØ	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	Ø✓	✓✓
Bing	✓Ø	✓Ø	ØØ	✓✓	✓Ø	ØØ	ØØ	✓✓	ØØ	✓✓	✓✓	✓✓	Ø✓	Ø✓	✓✓

Source: Author.

Numbers in Table 18 complement the previous analysis by summarizing the achievements and shortcomings identified over different years.

Table 18: Comparison of percentages

	✓ Correct	Ø Incorrect	
Google Translate	53%	47%	2018
	26%	74%	2021
DeepL	87%	13%	2018
	93%	7%	2021
Bing	60%	40%	2018
	53%	47%	2021

Source: Author.

Human assessment of MT output is considered to be subjective (Koehn, 2020; Rossi & Carré, 2022; Barreiro & Ranchhod, 2005), but as strategy, we kept the test sentences short and clear to avoid the essence of the linguistic problems be lost or distorted in the translation and in the evaluation. As there was only one evaluator,

the short-sentences-strategy was useful to maintain the focus on the specific problems.

Equally interesting, in the study done by Doan (2021) that challenged MT engines with short and long sentences, she reported that length sentences did not seem to affect the performance of the engines (Portage, Google Translate and DeepL translator) as the author originally thought. That said, short sentences might not be a limitation of this study either. However, another important limitation remains, which is the short amount of test sentences used in the set. This should be taken into consideration for future studies with similar aims.

Conclusions

All things considered, why does it matter to learn about this? Well, these findings are valuable information that should be taken into account for translators who may work with these systems as post-editors as their efforts would be better directed. For example, in light of some lexical problems (e.g. countable/uncountable nouns) for the engines, translators would focus their attention on whether or not the MT has translated properly this kind of phenomenon and worry less about those aspects that are known to be better handled by MTs. Another application could be to pre-edit documents where it is possible to make it easier for the system to handle.

Some other general observations are drawn from the experience. First, Google is not as reliable as most people may think, at least as far as the handling of linguistic issues of this type is concerned. Second, the fact that some engines propose just one option for translation makes them not a great choice for a translator who is looking for other possibilities. DeepL provides options and that is an asset. Third, there is a tendency to provide similar or same translations for a given text or statement. Although creativity is part of human nature, these machines seem to lack creativity. Despite the fact that this exercise was created to measure other aspects, two

systems (Google and Bing) tended to provide identical or similar translations, except for DeepL, which provided more than one option on most occasions. This absence of creativity was confirmed in the research conducted by Guerberof-Arenas & Toral (2020) in which readers assessed different types of translation. Fourth, we still have to consider, however, that the language employed for the testing was controlled, and the MTs seem to work outstandingly well for these short statements (except for Google). "The quality of MT output is closely connected to how MT-friendly the input is," (Austermühl, 2014, p. 163) and the input for this case was especially designed to cope with linguistic problems. We need to keep an eye on their upgrades (and on research work) about the management of longer texts, which ideally would be of greater help to the translator.

From my vantage point, it is somewhat easier to see now why professional translators are making more use of these systems and are already becoming literate in machine translation (machine translation literacy is actually a new concept in the field, created by Bowker & Buitrago Ciro (2019). In this sense, knowledge about MT is now also recommended in translator education. Some strategic sub-competences specific to post-editing are the following: "knowledge about MT systems and their capabilities," and "knowledge of typical MT errors," (Konttinen, Salmi & Koponen 2021, p. 194). The inclusion of such skills' development in translation schools would lead to a better integration of human translation with MT in the near future (Konttinen, Salmi & Koponen, 2021, p. 188) and supports exercises like this one.

What is surprising at this point is that DeepL MT is not as popular for the general public as is Google Translate. Translators are more aware of this, and it is more evident now why it can be a valuable support for translation assignments. Hopefully, this challenge may not only make people aware of the possible drawbacks, but also show some of the advantages that the use of engines such as DeepL could bring to the labor of any professional translator. As Lagoudaki (2008, p. 262) states, "the use of MT is

now considered a common practice among translators who prefer to have a rough draft of a translation before they produce a final translation, by editing the first draft," and this practice seems to be a good idea now after witnessing such a valuable addition.

Suggestions of use, however, would be in the less specialized use of language, for non-official translation and for an analysis of the use of language. Nevertheless, machine translations would always need to be revised by a translator or a fluent target language speaker. Language pair combinations matter, as well, and should be taken into account when assessing MT performance. We should consider that English, which is a lingua franca, when combined with any other language spoken by many (like Spanish, in this case) has an increased opportunity for better matches in machine translation.

Apparently, most people seem to use machine translation for quick communication and "gisting", as they are appropriate for the transferring of ideas to interact with others or for functioning in the world. It would not matter much to use it to translate a recipe or section of an e-mail from a foreign friend. However, these engines are also being used for publication on the web, other social media or for more serious work in the field. In any case, post-edition is necessary to make the resulting translation more natural.

In conclusion, MT systems do not give the impression of threatening translators' jobs in the near future, as their proposals are not flawless yet, but they surely are getting better. Human parity does not seem achievable so far, but still, there could be a harmonious collaboration between humans and machines. We had better delve into this technology now and start being a part of this functional relationship.

Acknowledgements

I would like to thank Professor Elizabeth Marshman for introducing the topic of challenge sets to test machine translation systems in our doctoral class in the University of Ottawa. And for assisting me in selecting my challenge set in order to keep it in realistic terms so

that these engines would be able to tackle the linguistic problems in mind. My appreciation also goes to Peter Clabrough for help in editing this paper.

References

Austermühl, Frank. *Electronic Tools for Translators*. London: Routledge, 2014.

Austermühl, Frank & Kortenbruck, Anke. "A Translator's Sword of Damocles? An Introduction to Machine Translation". In: Austermühl, Frank. *Electronic Tools for Translators*. London: Routledge, 2001/2014. p. 153-176.

Banitz, Brita. "Machine Translation: A Critical Look at the Performance of Rule-Based and Statistical Machine Translation". *Cadernos de Tradução*, 40(1), p. 54-71, 2020. DOI: https://doi.org/10.5007/2175-7968.2020v40n1p54

Barreiro, Anabela & Ranchhod, Elisabete. "Machine Translation Challenges for Portuguese". *Lingvisticæ Investigationes*, 28(1), p. 3-18, 2005. DOI: https://doi.org/10.1075/li.28.1.03bar

Bing Translator. Available at: https://www.bing.com/translator. Accessed in: Nov. 25, 2018 and Dec. 20, 2021.

Bowker, Lynne & Buitrago Ciro, Jairo. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald Publishing Limited, 2019. DOI: https://doi:10.1108/978-1-78756-721-420191002

DeepL Traductor. Available at: https://www.deepl.com/es/translator. Accessed in: Nov. 25, 2018 and Dec. 20, 2021.

Doan, Coraline. *Comparing Encoder-Decoder Architectures for Neural Machine*

Translation: A Challenge Set Approach. 2021. 274 f. Thesis (Master in Translation Studies) – University of Ottawa, Faculty of Arts, School of Translation and Interpretation, Ottawa, 2021.

Google Traductor. Available at: https://translate.google.com/. Accessed in: Nov. 25, 2018 and Dec. 20, 2021.

Guerberof-Arenas, Ana & Toral, Antonio. "The Impact of Post-Editing and Machine Translation on Creativity and Reading Experience". *Translation Spaces*, 9(2), p. 255-282, 2020. DOI: https://doi.org/10.1075/ts.20035.gue

Isabelle, Pierre; Cherry, Colin & Foster, George. "A Challenge Set Approach to Evaluating Machine Translation". In: Conference on Empirical Methods in Natural Language Processing, 55., 2017, Copenhagen. *Proceedings* [...]. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 2486-2496. DOI: https://doi.org/10.18653/v1/D17-1263

Isabelle, Pierre, & Kuhn, Roland. "A Challenge Set for French → English Machine Translation". *arXiv preprint arXiv:1806.02725*, 2018. DOI: https://doi.org/10.48550/arXiv.1806.02725

Koehn, Philipp. *Neural Machine Translation*. New York: Cambridge University Press, 2020. DOI: https://doi.org/10.1017/9781108608480

Konttinen, Kalle; Salmi, Leena & Koponen, Maarit. "Revision and Post-Editing Competences in Translator Education". In: Koponen, Maarit; Mossop, Brian; Robert, Isabelle S. & Scocchera, Giovanna (Eds.). *Translation Revision and Post-Editing: Industry Practices and Cognitive Processes*. London: Routledge, 2021. p. 187-202.

Lagoudaki, Elina. "The Value of Machine Translation for the Professional Translator". In: Conference of the Association for Machine Translation in the Americas, 8., 2008, Waikiki. *Proceedings* [...]. Waikiki, USA: Association for Machine Translation in the Americas, 2008. p. 262-269. Available at: https://aclanthology.org/2008.amta-srw.4.pdf. Accessed in: Feb. 2, 2023.

Marshman, Elizabeth. *Evaluating MT*. Ottawa: University of Ottawa, 2018. p.

1-20.

Parra Escartín, Carla & Goulet, Marie-Josée. “When the Post-Editor Is Not a Translator”. In: Koponen, Maarit; Mossop, Brian; Robert, Isabelle S. & Scocchera, Giovanna (Eds.). *Translation Revision and Post-Editing: Industry Practices and Cognitive Processes*. London: Routledge, 2021. p. 89-106.

Rossi, Caroline & Alice Carré. “How to Choose a Suitable Neural Machine Translation Solution: Evaluation of MT Quality”. In: Kenny, Dorothy (Ed.). *Machine Translation for Everyone: Empowering Users in the Age of Artificial Inteligence*. Berlin: Language Science Press, 2022. p. 51-79. DOI: https://doi.org/10.5281/zenodo.6653406

Recebido em: 12/08/2022
Aprovado em: 16/01/2023
Publicado em fevereiro de 2023

Argelia Peña Aguilar. Ottawa, Ontario, Canada. Chetumal, Quintana Roo, México. E-mail: argelia@uqroo.edu.mx https://orcid.org/0000-0002-3591-1985.