Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation?*

Lucía Sanz-Valdivieso and Belén López-Arroyo

1, 2 University of Valladolid, Spain. P. Campus s/n, 47011, Valladolid (Spain)
lucia.sanz.valdivieso@uva.es

Abstract. Experts and professionals in specialized fields often need writing tools to communicate in English as a means to disseminate their knowledge or enter the international market. There are different tools to accomplish this and most of them are, lately, Machine Translation systems (MT) based on Neural Machine Translation (NMT), an approach using artificial neural networks to translate with outstanding fluency. Free and open systems such as Google Translate or, more recently, ChatGPT used as a translator, have popularized NMT to a multitude of users. However, there are experts and professionals who, due to their lack of command of English, often fail in their communication tasks by accepting NMT system’s output as correct. This paper examines these systems’ performance when translating terminology of the discourse in wine and olive oil tasting notes, specifically from Spanish into English. This domain may serve to represent less-studied specialized languages where general language words and terms become closely intertwined. The aim is to determine whether these systems can translate terminology accurately within the domain, and, if so, whether the GPT-3.5 model outperforms Google Translate. Results will help identify or discard possible language solutions for users who need to obtain texts in specialized English with professional and internationalization purposes, but who do not have the linguistic or economic resources to ensure the quality of the English text. Results show that, although ChatGPT yields fewer terminological errors than Google Translate in terms of error severity and number of samples affected, professionals cannot rely solely on these tools just yet.

Keywords: Languages for Specific Purposes, Terminological accuracy, Translation Quality Assessment.

* The authors belong to the ACTRES (Análisis Contrastivo y Traducción Especializada ES-EN) research group. This paper has been written within the research project Lenguajes naturales controlados, comunicación colaborativa y producción textual bilingüe en entornos 3.0 (PID2020-114064RB-100), supported financially by the Ministerio de Educación y Ciencia de España, and the project Writing Audit: Evaluación de la redacción técnica con entornos visuales (PID-078) supported financially by the University of Valladolid. Lucía Sanz-Valdivieso develops her research under a fellowship granted by the Ministerio de Educación (FPU20-00293).
1 Introduction

Since the computer was invented, humans have been, rather illusorily, aiming at Fully Automatic High Quality Machine Translation (FAHQMT) [1]. However, there has not been a model closer to that aim than Neural Machine Translation (NMT), the indisputable state-of-the-art in the field of MT. Its main advancement regarding its predecessors lies on its computational approach and its immensurable potential only limited by computer power and memory [2]. The neural approach to Natural Language Processing (NLP), based on Artificial Neural Networks (ANNs) [3], allows NMT to account for the richness of language through the principle of semantic compositionality and vectorial representation. Thus, NMT systems “build the interpretation of each sentence by combining the individual interpretations of its component words” [4, pp. 141–142].

1.1 Relevant work

Indeed, the wake of new applications of this technology into chatbots, which can be used for translation, may become a threat for some professionals, including translators [5]. There are even voices claiming that NMT systems can produce translations of such a high quality that “might and should worry some translators … [b]ecause it is close to FAHQMT” [6, p. 201]—so much so, that there have already been declarations of MT reaching human parity [7]. However, these systems are still far away from attaining FAHQMT in the majority of text types, language combinations, and when the source text is not written in a Controlled Language, although this situation is rapidly changing [6]. In fact, while NMT’s general quality is higher than other systems—in terms of fluency, accuracy, but not style—, this is not perceivable in all language pairs and it is negatively affected by sentence length [8; 9; 10]. Errors of any kind, especially critical errors, increase when translating online user-generated content, which is usually colloquial, ungrammatical, and contains emojis and other characters [11].

Still, numerous studies point towards NMT as the highest quality kind of MT in different pair combinations and using different assessment methods in high-to-moderate resource settings [1; 12]. While general NMT quality is indeed higher than Statistical MT [8; 9; 10; 13; 14; 15; 16; 17; 18], most errors tend to be lexical [19], even though NMT produces fewer word order and morphological errors [16]. NMT also outperforms Phrase-Based MT in technical translation quality in the business language, except in the categories of terminology and formatting tags [20; 21; 22; 23]. Post-editing effort is also lower in NMT systems’ output [24; 25]—the most frequent changes are related to word substitutions and word form [24], confirming too NMT’s relative terminological and lexical weakness [26]. This is especially relevant since many texts to be translated using MT belong to specialized domains, where terminology takes a central role [27]. Comparing two free open-source NMT systems, Google Translate and DeepL, when translating Spanish phraseological units both show a similar performance which is weakened when encountering low-frequency expressions [28]. Other studies confirm such results in Portuguese-French, where phraseology, calque and nonsense were the most frequent errors [29].
1.2 Motivation and focus of the study

The purpose of this paper is to test the terminological accuracy of the latest NMT systems in a very specific text type within a specialized discourse: olive oil and wine tasting notes written in Spanish. Tasting notes are usually short texts describing a product’s organoleptic attributes, composed of relatively long sentences and full of lexical and terminological richness [30; 31; 32; 33]. The wider study in which this experiment develops addresses a very common type of user, i.e., professional or technical experts who, in spite of not being able to produce specialized texts in English by themselves, do need to obtain such texts. These users need texts in English as the international lingua franca for a diversity of purposes, ranging from marketing, to labelling, to touristic promotion and education. These factors ultimately determine the international economic performance of sectors as important as the wine and olive oil industries in Spain, in this case. Nevertheless, the ongoing project aims at extrapolating results to other specialized fields where speakers belong to small-medium organizations and need to obtain English texts but do not have the ability to compose such texts on their own nor the means to adopt quality language services.

The relevance of the kind of expert described above is in their lack of ability to identify an inadequate translation. This would not pose a problem if these users’ aim when using MT was gisting-related [3]—but these users know the content of the source text and are translating into a language they do not fully command. An added issue to this profile is the lack of economic means that most multi-national companies can invest in high-quality Language for Specific Purposes (LSP) translations to promote their internationalization. The reality is that these experts cannot possibly post-edit a faulty translation in the way a professional translator would. Rather, they will usually just copy and paste the MT system’s output, or directly integrate a Google Translate plug-in in their website to be able to offer its English version in some way, even if flawed (Fig. 1 below).

![Fig. 1. Examples of resources currently used by these users to translate content into English.](image)

This is just an instance showing how experts in this kind of small specialized domain take NMT as FAHQMT, even though translation professionals and English-speaking members of the discourse community would identify possible errors in the text [34]. While errors most frequently result in unnatural expressions unrecognizable for the target discourse community, they may also reach the extent of impeding successful
communication. Unfortunately, this may have serious consequences for companies individua-

In this sense, few works have examined NMT in specialized contexts, with the ex-

Tasting notes (TNs) may indeed be viewed as suitable candidates for FAHMQMT given

The selection of MT system(s) to be tested is determined by the purpose of each

2 Methodology

2.1 Dataset

The dataset used consists of samples from olive oil and wine TNs Spanish corpora

In the case of CGPT-3.5, the prompt “Traduce de español a
inglés” [Translate from Spanish into English] was used for it to act as a translator without finetuning the results through a more accurate prompt—a path currently under study in the wider project where this experiment develops.

2.2 Methods

To test the terminological accuracy of current NMT systems at translating TNs, this paper aims at analyzing translations performed by free, open and popular systems, as part of the project which considers complementary means of human Translation Quality Assessment (TQA) as well as automated metrics. For this purpose, we used our familiarity with the LSP of olive oil and wine tasting in English and our background in translation and linguistics to perform the human evaluation of the target texts from both NMT systems. Hence, there were two annotators in this experiment, where inter-annotator agreement was calculated through Cohen’s Kappa with a result of $K=0.7242$, indicating substantial agreement.

This TQA was performed through the Multidimensional Quality Metrics (MQM) framework, developed to provide a comprehensive and standardized quality assessment model [36]. It comprises a set of 182 issue types hierarchically organized into dimensions; not all of the types are to be covered in the assessment of a translation—rather, they are to be used to ensure that said translation “meets specifications” [47, p. 119]. In other words, the MQM proposes a functionalist framework where the translation’s purpose in context plays a central role in how its quality should be assessed [48]. While the broader project on which this work-in-progress paper is based covers more MQM aspects, only a limited set of issues are reported here.

The terminology dimension of the MQM “relate[s] to the use of domain- or organization-specific terminology” and is made up of three possible kinds of issues: inconsistency with termbase, with domain, or inconsistent use of terminology along the text [49]. Our analysis focuses on terminological inconsistency with domain, since there is not a specific termbase that standardizes tasting terminology. Hence, a term was flagged in a translation when it “is used contrary to general domain expectations” [49]. Nevertheless, a forthcoming terminological and phraseological glossary we developed for related projects was consulted for guidance when necessary. In addition and for the sake of accounting for terminological accuracy comprehensively, the issue “mistranslation” within the “accuracy” dimension (i.e., “does not accurately represent the source content” [49]) was also considered when terminology was affected. Untranslated expressions were noted as well.

To quantify the extent of the issues detected through the selected parameters, the MQM provides four severity levels: critical errors, where a translation is unfit for its purpose, involving legal, safety or usability consequences; major errors, which “make the intended meaning of the text unclear …[and] the user cannot recover the meaning” [47, p. 120]; minor errors, without an impact on usability; and null level (changes that are not errors) [47]. Since tasting notes hardly ever entail critical danger, level 1 is excluded from the analysis, as well as level 4, since we are concerned with errors per se.
3 Results and discussion

After following the MQM to examine the 50 samples and their translations, the analysis revealed that samples translated by CGPT-3.5 yield 21.57% fewer errors, mistranslations and untranslated elements of the tasting LSP than GNMT:

<table>
<thead>
<tr>
<th>MT system</th>
<th>Terminological inconsistency with domain</th>
<th>Mistranslated expressions</th>
<th>Untranslated expressions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error severity levels</td>
<td>Major</td>
<td>Minor</td>
<td>Major</td>
<td>Minor</td>
</tr>
<tr>
<td>GNMT</td>
<td>23</td>
<td>50</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>CGPT-3.5</td>
<td>14</td>
<td>45</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

Results seem to show CGPT-3.5 used as a translator outperforms GNMT in terms of general terminological accuracy when working with TNs from Spanish into English. Most errors belong to the terminological issue type, i.e., translations that could work outside the specific domain in question. In many cases, both systems overlooked domain-specific terminology and used other general language equivalents instead:

1. Source text: … un vino de capa alta, de gran *brillante* …
   GNMT: … a wine with a high robe, of great *brilliance* …
   Source text: … el aceite presenta un aspecto *brillante*.
   CGPT-3.5: … the oil has a *shiny appearance*.

Other common terminological errors found in the translations include “capa,” translated as “layer” and not as “robe” in most of the cases by both systems; “entrada” as “entrance” and not “entry”; “paso” and “recorrido” often translated as “step”, “passage” and “journey” instead of “mid palate”; “recuerdos” as “memories” or “reminders” and not as “hints,” “notes,” or even “reminiscences”; or the color descriptor “teja” translated as “tile” instead of “brick”; which CGPT-3.5 used interchangeably.

Tasting verbs were also not correctly translated, such as “ofrecer” or “regular” being translated as “give” instead of “offer”, or progressive expressions such as “apreciándose” instead of “appreciating”. Other mistranslations include “aceituna de pre-envero” translated as “pre-veraison olive” (GNMT) or “olive in pre-winter” (CGPT-3.5) and not “green olive”. Untranslated expressions were recurrently “alloza” [green almond] and “bodega” [cellar, winery, vineyard].

However, CGPT-3.5 was slightly more accurate than GNMT:

2. Source text: El *picor* es ligero pero se nota.
   GNMT: The *itch* is slight but it is noticeable.
   CGPT-3.5: The pungency is light but noticeable.

Other instances where only CGPT-3.5 was able to find the correct tasting term include simple agrarian terms such as “tomatera”, translated as “tomato plant” by CGPT-3.5 but as “tomato” by GNMT, which also output “nariz voluminosa” as “bulky nose”; or “zumo cordobés [from Córdoba, Spain]” as “Cordovan juice”. Similarly, CGPT-3.5 was able to correctly translate the verbs “finalizar, terminar” as “finish” and not “end”.

102
Yet, particularly note-worthy are GNMT’s incorrect translations of “vista” [appearance] as “sight” and “view”, and, “nota de cata” [tasting note], as “Cata’s note”.

Not only did CGPT-3.5 produce translations with fewer errors as a whole—it also output more texts free of any terminological errors, with a 38.89% difference from GNMT texts:

<table>
<thead>
<tr>
<th>MT system</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error severity levels</td>
<td>None</td>
</tr>
<tr>
<td>GNMT</td>
<td>7</td>
</tr>
<tr>
<td>CGPT-3.5</td>
<td>18</td>
</tr>
</tbody>
</table>

Still, even though 36% of the samples translated by CGPT-3.5 were error-free in contrast to the 14% of the texts by GNMT, this does not mean that they are ready to be regarded as acceptable. It is worth taking a closer look at one of the translations where no terminological issues were detected by GNMT (example 3) and CGPT-3.5 (example 4):

3. Very greenish yellow color. The nose is intense, complex, fresh grass, artichoke hints and a touch of dried fruit. On the palate it is dense, fruity, of good intensity, spicy, clean, with a vegetal touch, well-balanced sweet sensations and good length.

4. Wine with an intense cherry red color and soft violet notes, which indicates its aging with bright terracotta and amber edges. A nose of great subtlety with aromas of ripe fruit well combined with the aging in wood, leading us to special aromas (vanilla, cinnamon) to the stimulating aroma of coffee or toasted notes. Very meaty on the palate, with a long finish and balanced acidity.

These are accurate translations terminologically speaking, but both the fluency and some grammatical structures are questionable to different extents. For instances, in example 3, the main verb “is” is used to embed a series of phrases into an enumeration that results rather unnatural in English; while, in example 4, the first two sentences become so long—which is completely natural in Spanish—that even grammatical mistakes can be spotted (“lead from X to Y”, and not “lead to X to Y”). In this sense, there is ample room for improvement in most terminologically accurate translations, as well as in those with errors, in terms of fluency. In other words, even though NMT’s output sounds surprisingly more natural than previous systems’, these outputs prove how there are still robotic transfers of grammatical patterns into the target language that prevent them from being acceptable within the target discourse community. Besides, issues such as consistency are worth a more detailed analysis, since most errors were frequently, but not always present in both of the systems’ output.

4 Conclusion

This work-in-progress paper has focused on how the latest and most popular free and open MT systems treat terminology within the specialized field of tasting, which differs from other more pervasive, technological and more objective domains. This
rendered a sample of olive oil and wine tasting notes a rich and interesting ground for research in this regard—many human activities and sectors use specialized languages which are often too small, subjective or intermingled with general language for MT systems to be able to correctly translate them, even in two of the largest languages in the world. Our MQM-based, terminology-focused analysis has proved the enormous potential of these systems while revealing terminology is not the greatest strength of neural systems. Most importantly, results show a better performance by CGPT-3.5 used as a translator than GNMT terminology-wise.

Still, none of these systems outputs texts that are acceptable for the user who does not have the training, linguistic or economic means necessary. The wider project in which this work in progress develops overcomes some of the limitations of the present paper, and is currently looking at other aspects of the MQM, as well as applying automatic TQA metrics in order to obtain a complete picture of the behavior of these systems. In any case, there is a long way ahead in order to develop tools which can help this kind of user achieve their goal, which may range from finetuned MT systems to complementary tools such as terminological aids that may help them obtain their TNs in English and so promote and internationalize their products, businesses, and cultural assets.

References


