Google Translate for Medical Texts: A Quantitative-Qualitative Analysis of English into Arabic Package Inserts Translation

Rania Al-Sabbagh

Assistant Professor,
Faculty of Al-Alsun
(Languages), Ain
Shams University,
Egypt.

Abstract

Although machine translation systems like Google Translate have made great strides, there are still concerns about their use for medical translation. Medical experts, researchers, and end-users doubt that Google Translate could pose serious risks, as it may distort the original meaning or omit vital information. This study argues that Google Translate should not be perceived as risky, mainly when translating package inserts from English into Arabic, as one example of medical texts. This argument stems from a quantitative-qualitative analysis of Google Translate’s translation performance, utilizing a corpus of 50 package inserts obtained from the Saudi Food and Drug Authority with their official Arabic translations. The quantitative analysis employed statistical measures to compare Google Translate’s output to the official translations, assess post-editing effort, validate whether end-users can distinguish between Google Translate’s output and official translations, and describe the accuracy and fluency error distribution. Simultaneously, the qualitative analysis involved a manual inspection of a random sample of 760 sentence pairs, employing Tezcan et al.’s (2018) taxonomy of translation errors to identify and categorize errors as accuracy-related or fluency-related. The results revealed significant differences between Google Translate’s output and the official translations, although these disparities were predominantly attributed to stylistic variations rather than errors. The results also showed that end-users were mostly unable to discern between Google Translate's output and the official translations. Moreover, only 165 out of the 760 sentences contained errors, with the majority being fluency-related rather than accuracy-related. Google Translate’s output, evaluated in this study, was generated in November 2023.

Keywords: English-Arabic translation, Google Translate, machine translation, medical translation, package inserts

Google Translate for Medical Texts: A Quantitative-Qualitative Analysis of English into Arabic Package Inserts Translation

Rania Al-Sabbagh

1. Introduction

Despite the considerable strides in machine translation, particularly with the advent of neural machine translation around 2016, there remains substantial skepticism surrounding the use of machine translation within the medical domain. Naeem Nazem, a medical advisor to the Medical and Dental Defense Union of Scotland, emphasized the potential risks, stating, “In usual clinical practice, the use of computer translations, when validated alternatives are available, is likely to heighten the risks to patient safety. This leaves doctors susceptible to criticism and, potentially, regulatory action or litigation in the event of an adverse outcome... the risk of error is significant” (as cited in Moberly, 2018, p.1).

Illustrating the risks, Khoong et al. (2019) demonstrated that using Google Translate to translate emergency department discharge instructions from English into Chinese and Spanish resulted in inaccurate and potentially life-threatening translations. For example, the instruction “Hold the kidney medicine until you have a chance to speak with your kidney doctor” was translated into Chinese as “Keep taking kidney medicine until you talk to your kidney doctor” and into Spanish as “Keep the medication for the kidney until you have the chance to talk with your kidney doctor” (Khoong et al., 2019, p. 581). In another study, Das et al. (2019) assessed the accuracy of Google Translate in translating anticipatory guidance material provided to parents in English (i.e., proactive advice on a child’s health and development) into 20 languages, including Arabic. Human evaluators rated the accuracy of Google Translate’s output on a scale from 1 to 5, where 1 was the lowest and 5 was the highest. The Arabic translations received an average rating of 3.03, categorized by the researchers as deficient, signifying that “the translation contained errors that slightly obscured or changed the meaning” (Das et al., 2019, p. 247).

Not only do medical experts and researchers such as Das et al. (2019) and Khoong et al. (2019) advise against relying on Google Translate for medical translation, but end-users, including translation professionals and the general public, also harbor skepticism regarding its accuracy. Using Google Forms, I conducted two surveys to learn about the attitudes of Arabic speakers toward Google Translate in the context of medical translation. In the first survey, 124 professional English/Arabic translators participated. Regarding their work experience, the distribution was as follows: 55 participants (49.1%) possessed 5+ years of experience, 30 participants (26.8%) had 3-5 years of experience, and 27 participants (24.1%) had 1-2 years of experience. Across these experience tiers, 111 participants (99.1%) agreed they would not trust using Google Translate for medical translation without full post-editing. This entails checking terminology against approved terminological resources, cross-referencing translations with other resources, making syntactic modifications in accordance with practices for the target language, ensuring stylistic fluency, and applying correct formatting and tagging, among other considerations. Moreover, 104 participants (92.9%) lacked trust in using Google Translate for even non-medical translation without thorough post-editing.

In the second survey, 214 native speakers of Arabic without prior translation experience and with variant levels of English proficiency participated. Out of those participants, 94 (69.1%) said that they tried using Google Translate to translate English package inserts (i.e., the documents included in the packages of each medication to provide information about that drug and its use, also known as patient information leaflets). Out of those 94 participants, 71 (75.5%) found Google Translate's output acceptable, with some unclear words and sentences that they managed to overlook and understand the overall meaning, and 21 (22.3%) found the output to be barely understandable. Furthermore, 117 (86%) participants indicated that they prefer human translation to Google Translate, and they would prefer asking a doctor, a nurse, or a pharmacist, if available, for the translation instead of using Google Translate.

The warnings of the medical experts in Moberly (2018) and even the results of Das et al. (2019) and Khoong et al. (2019) date back five to six years. Machine translation systems are updated regularly as more and more data become available. Furthermore, Das et al. (2019) and Khoong et al. (2019) used back translation to evaluate Google Translate's output, which is a problematic method. Back translation, or reverse translation, is where content is translated back to its original language and compared to the source text. Behr (2017) and Colina et al. (2017) argued that while back translation can uncover problems, it causes several false alarms, and even more importantly, many issues remain hidden.

In this study, I evaluated Google Translate's output for translating English package inserts into Arabic using a corpus of 50 English package inserts collected from the Saudi Food and Drug Authority (SFDA) and their official translations. Google Translate's output, evaluated in this study, was generated in November 2023. Furthermore, I used a mixed-methods approach, working directly on Arabic translations; back translation was not involved in the evaluation. The questions this study investigated were:

1. How far does Google Translate's output align with the official translations?
2. How much post-editing effort is anticipated to transform Google Translate's output into a replica of the official translations?
3. Do the variances between Google Translate's output and the official translations signify errors, or are they stylistic differences?
4. In the case of errors, which category predominates: accuracy errors or fluency errors?

2. Literature Review

Zappatore and Ruggieri (2024) conducted a systematic review of research on using machine translation in the medical domain. They analyzed 58 articles from various journals and conference proceedings. The studies focused on English-to-Spanish and English-to-Chinese translation. The researchers from the reviewed papers found that medical professionals and patients had concerns about the accuracy of machine translation systems. However, experiments about the quality of machine translation systems for the medical domain run in those papers showed that the issues in fluency, accuracy, unnatural translations, and domain adequacy could be easily addressed via two main strategies. Firstly, training machine translation systems on more domain-specific (i.e., medical) data, as a properly prepared training dataset, ensures a substantial performance enhancement; secondly, combining machine translation with in-domain human post-editing. Therefore, Zappatore and Ruggieri (2024) concluded that machine translation should be used in healthcare, especially when human translators are unavailable.

A few researchers investigated English-to-Arabic machine translation within the medical domain. However, some researchers exhibited bias against machine translation systems without offering sufficient justification. For instance, Almahasees et al. (2021) assessed Google Translate's performance in translating COVID-19 documents acquired from international organizations' websites, such as the World Health Organization, the United States Food and Agriculture Administration, and the European Center for Disease Prevention and Control. However, the researchers did not use standard quantitative evaluation metrics such as BLEU, chrF++, or TER to evaluate Google Translate's output, as commonly employed in machine translation literature (refer to Section 3.2 for more details on these metrics). Meanwhile, the researchers claimed that semantic, grammatical, lexical, and punctuation errors in Google Translate's output "inhibit the intelligibility of the translated texts" (Almahasees et al., 2021, p. 2065). However, they failed to substantiate this claim through surveys or interviews testing the intelligibility of the translated texts among end-users. Additionally, none of the examples in the researchers' article indicated any significant alterations in meaning that could impact end-users' understanding or pose risks, in contrast to the examples provided by Khoong et al. (2019) for English-to-Spanish and English-to-Arabic medical translations (refer to Section 1).

Ehab et al. (2019) tested Google Translate to translate symptoms and side effects extracted from English internal medicine journal articles. The data against which Google Translate was evaluated did not include complete sentences but rather phrases such as احتقان الرئة (lung congestion) and خلل لوظائف الكلى (kidney impaired functions). Google Translate achieved a BLEU score of 0.51, which indicates high-quality translation (see Section 3.2. for more details on BLEU). Furthermore, the researchers proved that when a medical translation memory was used to enhance Google Translate, the BLEU score increased by about 0.1 points, rendering even better translations. The researchers did not discuss the mismatches between the reference translations extracted from the *Worldwide Arabic Medical Translation Guide: Common Medical Terms* and Google Translate. They did not discuss whether these mismatches were actual errors or different styles. They did not even discuss what aspects of translation were improved when the medical translation memory was added to Google Translate.

Sharkas (2019) focused on translating English package inserts into Arabic, not to evaluate Google Translate, but to investigate the underlying reasons for the low readability and lay-friendliness of the translated package inserts. Sharkas drew inspiration from Jensen's (2013) research on the lay-friendliness of Danish package inserts translated from English. Jensen found that the Danish public was less likely to read package inserts in Danish compared to their English source texts, attributing this to translations being more challenging to read and excessively lengthy and complex, thus deviating from the original goal of providing easily accessible information. Sharkas did not validate whether the Arabic-speaking public found translated Arabic package inserts more challenging to read than their English counterparts. Instead, she presumed this to be the case and investigated the reasons directly. After analyzing 20 translated package inserts, Sharkas concluded that the challenges in the readability and lay-friendliness of translated Arabic package inserts might stem from translators relying on medical dictionaries like the *Unified Medical Dictionary* without modification. For instance, the term ‘endemic goiter’ was translated as دراق متوطن in the dictionary and likewise in the package inserts that Sharkas investigated. However, Sharkas proposed amplifying it to تضخم الغدة الدرقية المتوطن, defining amplification as a translation strategy involving the addition of words or using descriptions to clarify a term. One implication highlighted by Sharkas in her study is that trainee translators, in particular, should be informed about the impact of medical terminology on the lay-friendliness of package inserts. This awareness can guide them to reduce complexity without compromising translation accuracy. It may involve adding explanations, especially when a medical term is crucial for the proper understanding and use of the medicine, even if such explanations are not provided in the source text.

My study reported in this paper diverges from earlier research in several aspects. First, unlike Almahasees et al. (2021) and Ehab et al. (2019), I conducted a comprehensive evaluation, encompassing quantitative and qualitative analyses, as detailed in the subsequent section. Second, the evaluation focused on Google Translate’s performance on complete sentences, distinguishing itself from the evaluation of Ehab et al. (2019), which concentrated on noun phrases. Finally, in contrast to Sharkas (2019), I assessed machine translation rather than human translation, employing a significantly larger corpus of 50 package inserts instead of 20.

3. Methods

3.1. Data

As mentioned in Section 1, package inserts or patient information leaflets are documents in medication packages that offer drug details. These details encompass drugs’ composition, intended effects, potential side effects, recommended dosage, and guidance on where to seek assistance in case of side effects. Ornia (2016) describes package inserts as a hybrid textual genre due to their utilization of everyday language and medical jargon and their dual purpose, serving expository and instructive functions.

Four reasons prompted the choice of package inserts as the focal point of this study. First, they represent a medical textual genre, aligning with the study’s focus on medical translation. Second, the absence of prior research on this textual genre underscores the need to explore this area. Third, the availability of official translations through the Saudi Food and Drug Authority (SFDA) facilitated the study’s accessibility to relevant materials. Lastly, package inserts constitute a genre that directly impacts the daily lives of ordinary individuals.

A random selection of 50 package inserts was collected from the SFDA website, which hosts several English package inserts accompanied by Arabic translations. The data collection process went as follows: first, the package inserts were collected in HTML format from the SFDA website; second, the HTML files were converted into text files using Sotoor AI[1], an artificial intelligence optical character reader; third, the text files were manually checked for typos; finally, the English sentences and their Arabic translations were manually aligned, creating bilingual tables similar to Table 1. The statistics of the final corpus are listed in Table 2.

Table 1: A Sample Bilingual Table

Source Text	Official Translations
This medicine contains methylprednisolone, which belongs to a group of medicines called steroids.	يحتوي هذا الدواء على ميثيل برينيزولون، الذي ينتمي إلى مجموعة من الأدوية تُسمى الستيرويدات.
Their full name is corticosteroids.	الاسم الكامل لهذه المجموعة هو الستيرويدات القشرية.
Corticosteroids are produced naturally in your body and are important for many bodily functions.	تُنتج الستيرويدات القشرية بصورة طبيعية في جسمك وهي مهمة للعديد من وظائف الجسم.
Boosting your body with extra corticosteroids such as Medrol can help if your body cannot produce enough corticosteroids due to problems with your adrenal glands (e.g., adrenal insufficiency).	يُمكن لتعزيز جسمك بستيرويد قشري إضافي مثل ميدرول أن يساعد في حالة كان جسمك لا يستطيع إنتاج ما يكفي من الستيرويدات القشرية نتيجة لمعاناتك من مشكلات في غدتيك الكظريتين (مثل القصور الكظري).
Corticosteroids can also help following surgery (e.g., organ transplants), injuries, or other stressful conditions.	يُمكن أن تساعد الستيرويدات القشرية أيضًا عقب الجراحات (مثل عمليات زراعة الأعضاء) أو الإصابات أو الحالات الأخرى المسببة للإجهاد.

Table 2: Corpus Statistics

	English Words		Arabic Words	
Sentence Pairs	**Tokens**	**Types**	**Tokens**	**Types**
6,966	84,165	9,720	80,630	14,780

Note: Word tokens refer to the total number of words in the texts, while word types refer to the total number of unique (i.e., non-duplicate) words.

3.2. Quantitative Evaluation

Three quantitative metrics were used in this study. The first is BLEU (bilingual evaluation understudy; Post, 2008). The second is chrF++ (character n-gram F-score; Popović, 2015). The last is TER (translation edit rate; Snover et al., 2006).

BLEU is the most used quantitative evaluation metric in machine translation literature; it is an *n*-gram sequence-based metric that counts the number of similar words between machine and reference translations while penalizing brevity (i.e., if the machine translation is shorter than the reference translation, the overall score is reduced). BLEU scores range from 0 to 1, with 1 indicating a perfect similarity between machine and reference translations.

One drawback of BLEU is that it does not consider synonyms and word-order alternations. This is especially problematic for flexible word-order languages like Arabic. Furthermore, it gives equal weights to content and function words, so a translation error in the verb predicate will be penalized equally as a translation error in an article or a preposition. That is why other metrics should be used along with BLEU, such as chrF++ (Popović, 2015) and TER (Snover et al., 2006).

chrF++ (Popović, 2015) operates at the character and word levels. It measures word-level similarity and gives partial credit for morphologically similar words based on the number of shared characters. For instance, if يجب (must) is translated as تجب (must) by a machine translation system, it will still be given partial credit instead of being completely penalized. Such a metric can be beneficial for a morphologically rich language like Arabic. The chrF++ scores range from 0 to 100, with higher values indicating closer matching between machine and reference translations.

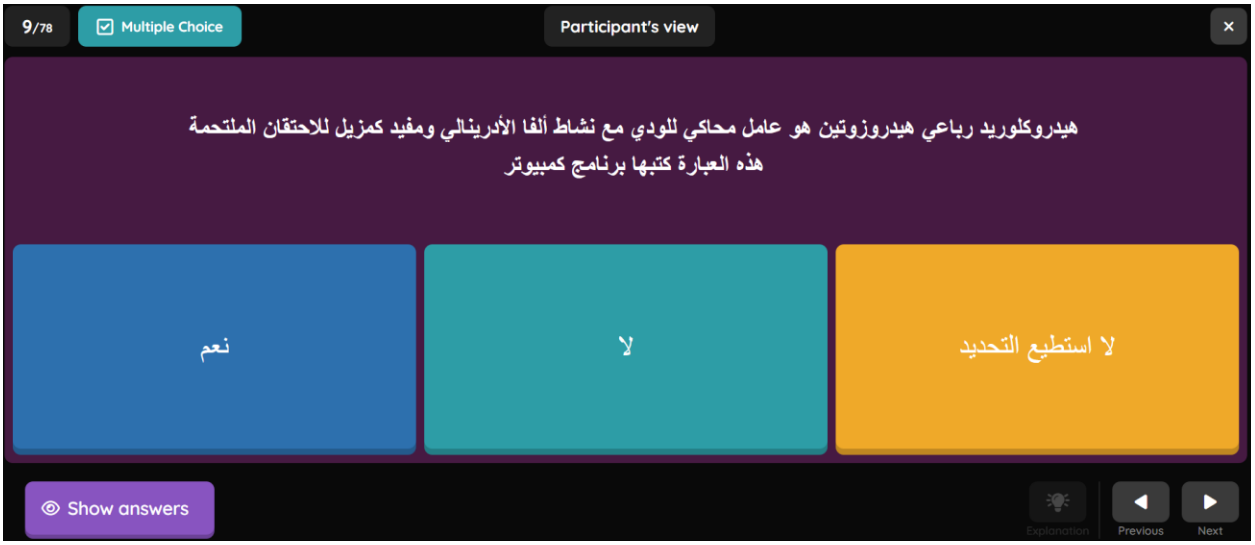
The TER (Snover et al., 2006) metric estimates the work required to turn the machine translation output into the reference translation. Specifically, it quantifies the number of edit operations (insert, delete, substitute, shift) required to change the machine translation output into the reference translation. The metric can be interpreted as the required post-editing effort since one could manually carry out these edit operations with a keyboard and a mouse (O'Brien, 2011). TER score can be a value between 0 and 1: the lower the score, the better (i.e., the fewer edits are required).

I used the Python libraries Sacrebleu[2] and PyTer[3] to compute BLEU and TER, respectively. For chrF++, I used Popović's (2015) code from GitHub[4]. Quantitative evaluation metrics are fast, free, objective, and language-independent. However, they do not provide insight into the types of errors Google Translate generates. For that reason, I combined quantitative and qualitative evaluations.

The last part of the quantitative evaluation focused on addressing the third research question: whether disparities observed between the output of Google Translate and official translations could be attributed to errors or were merely indicative of stylistic variations. To investigate this, an online survey was administered, wherein participants were presented with sentences and asked to determine whether the sentences were generated by a machine system (see Figure 1). The survey encompassed 75 sentences: 35 were generated by Google Translate, while the remaining 40 were extracted from the SFDA's official translations. The 35 sentences from the output of Google Translate exhibited low BLEU scores, falling within the range of 0.1 to 0.3. Furthermore, each sentence in the survey comprised more than three words.

The survey started with demographic inquiries, capturing participants' ages, Arabic language proficiency (whether it is their mother tongue or a second language), and their professions, categorizing them as students, translators, English language teachers, or others. The survey purpose was intentionally undisclosed to participants, aiming to prevent the influence of any negative stereotypes about Google Translate on their responses. Recognizing the substantial time and mental effort required, ranging from 20 to 30 minutes, the survey adopted a game-like format (see Figure 1). Utilizing Quizzizz, the design aimed to enhance engagement, incorporating music and random memes for added humor.

Figure 1: A Screenshot from the Online Survey



هيدروكلوريد رباعي هيدروزوتين هو عامل محاكي للودي مع نشاط ألفا الأدرينالي ومفيد كمزيل للاحتقان الملتحمة
هذه العبارة كتبها برنامج كمبيوتر

نعم

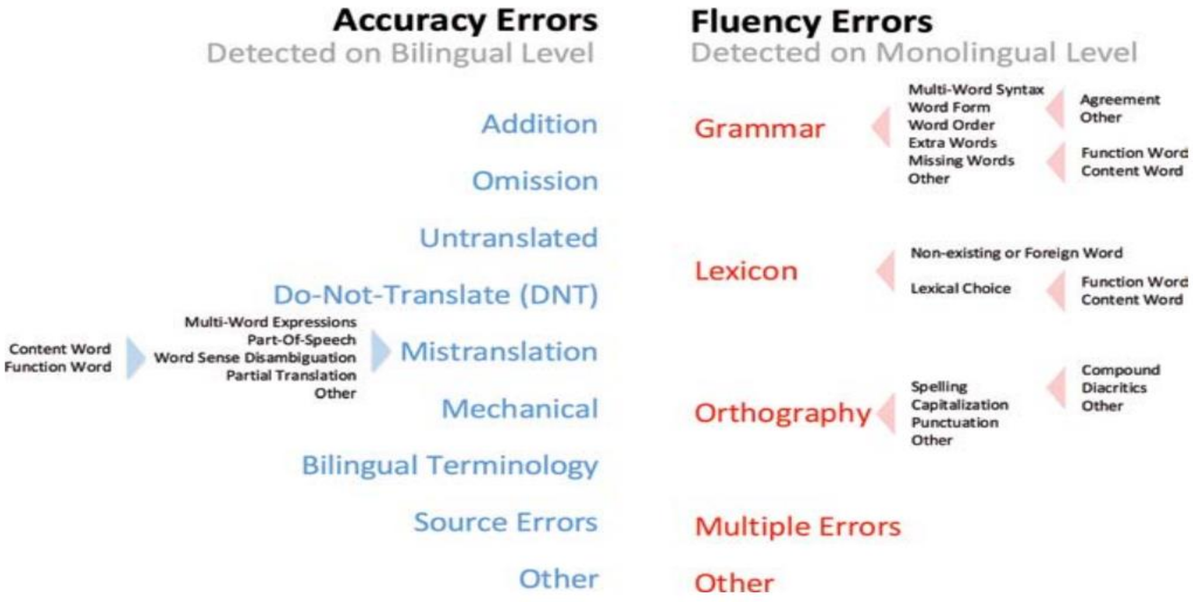
لا

لا استطيع التحديد

3.3. Qualitative Evaluation

For qualitative evaluation, I followed Tezcan et al.'s (2018) translation error typology known as SCATE (Smart Computer-Aided Translation Environment) to identify and categorize translation errors within a randomly selected sample of 760 sentence pairs. All the sentences had BLEU scores ranging from 0.1 to 0.3. SCATE, illustrated in Figure 2, assesses translation based on two primary dimensions: accuracy and fluency. Accuracy pertains to "how much of the source content and meaning is retained in the target text" (Tezcan et al., 2018, p. 222). Fluency addresses "the extent to which the translation flows well, regardless of sentence meaning" (Tezcan et al., 2018, p. 222).

Figure 2: SCATE Translation Error Typology



Tezcan et al. (2018) define several error types under the accuracy dimension, including addition, omission, mistranslation, and bilingual terminology. Addition errors refer to adding information in the target text that was not originally presented in the source text. Omission errors occur when information from the source text is deleted from the target text. Mistranslation errors mean that source content is translated incorrectly; for example, idiomatic expressions are translated literally; the wrong meaning of an English polysemous word is selected; numeric values are incorrectly converted from imperial to metric systems or vice versa; or quantities, dates, and times are inconsistent between source and target texts. Bilingual terminology errors result from translating terms incorrectly or inconsistently (i.e., when the same term is translated in multiple ways within the same document or across different documents within the corpus).

Tezcan et al.'s (2018) fluency errors include grammatical, lexical, and orthographic errors. Grammatical errors relate to incorrect subject-verb agreement, pronoun-reference agreement, word forms, word order, and tense usage. They also include incorrect and missing function words. Lexical errors relate to incorrect lexical choices that violate target language collocations. Orthographic errors relate to spelling and punctuation.

4. Results and Discussion

Table 3 shows the BLEU, chrF++, and TER scores achieved by Google Translate. A BLEU score of 0.255, as reported by Google Cloud (2024), indicates clear meaning. This surpasses many scores attained by Google Translate in translating questions and answers from diagnostic patient interviews from English into seven other languages. Costa-Jussà et al. (2012) reported the following BLEU scores for translating 500 questions and answers from diagnostic patient interviews from English into various languages: 0.24 into French, 0.2 into Portuguese, 0.26 into Spanish, 0.17 into German, 0.13 into Russian, and 0.72 into Basque. It is worth noting that attaining a BLEU score of 1 is nearly impossible. Doshi (2021) argues that a BLEU score within the range of 0.6 to 0.7 represents the optimal performance for a machine translation model; scores outside this range might suggest overfitting.

Table 3: Quantitative Evaluation Scores

BLEU	**chrF++**	**TER**
0.255	51.13	0.597

However, the scores in Table 3 are worse than those achieved by Google Translate for translating other textual genres, such as newspaper articles from English into Arabic. For example, Kadaoui et al. (2023) reported a BLEU score of 0.66, a chrF++ score of 78.97, and a TER score of 0.286. Likewise, Moslem et al. (2023) reported a BLEU score of 0.44, a chrF++ score of 62, and a TER score of 0.58. The difference between the good scores attained by Kadaoui et al. (2023) and Moslem et al. (2023) and the scores Google Translate achieved in this study can be attributed to the abundance of English-Arabic parallel corpora with billions of words featuring general-purpose textual genres such as newspaper articles in contrast to the lack of large specialized English-Arabic corpora that cover the medical domain. Available corpora are either comparable or monolingual. For example, the comparable corpus of Moreno-Sandoval and Campillos-Llanos (2013) contains Spanish, Japanese, and Arabic biomedical articles collected from several websites: Altibbi, Alawsat, Youm7, and Alkhabar. There is also the corpus of Boudjellal et al. (2020), which contains 49,856 sentences on the Altibbi website, yet it is a monolingual corpus.

Another monolingual corpus of Arabic medical texts is that of Abdelhay et al. (2023), which comprises 430,000 Arabic questions and answers distributed across 20 medical specializations extracted from the Altibbi website. Finally, there is also a monolingual corpus of 2,026 medical Arabic tweets collected by Alayba et al. (2017).

Despite the clarity of meaning that a BLEU score of 0.255 shows, the values of BLEU, chrF++, and TER in Table 3 reveal significant disparities between Google Translate's output and the official translations on the SFDA website. The question is whether these disparities signify errors or stylistic variations. The survey, as detailed in Section 3.3, was taken by 33 individuals aged 18 and above. Approximately 91% (30 out of 33) of the respondents were native Arabic speakers, with the remaining participants being second-language speakers. Regarding professional backgrounds, the participants included 9 students, 9 translators, 6 teachers, and 9 individuals with various other occupations. The accuracy rate was 46%, representing the average frequency participants correctly identified sentences originating from either Google Translate or SFDA. Table 4 displays the distribution of survey responses.

Table 4: The Distribution of Survey Responses

	Google Translate	**Official Translation**	**Can't Discern**
Participants who got it wrong	498	591	173
Participants who got it right	579	634	173

Utilizing the statistics presented in Table 4, I conducted a chi-square test with a significance *p*-value < 0.5 to assess the null hypothesis, suggesting no association between participants' choices and the translation source. The alternative hypothesis posited an association between participants' choices and the translation source. In simpler terms, the test aimed to validate whether participants could distinguish between sentences generated by Google Translate or SFDA. The chi-square test results revealed a value of 1.7986, yielding a *p*-value of 0.406853. Excluding the "can't discern" option, the chi-square value was 0.9246, with a *p*-value of 0.336274. In either case, the observed differences were statistically insignificant, leading to the acceptance of the null hypothesis. This implies that participants could not confidently and consistently determine whether the displayed sentences originated from Google Translate or SFDA. Further evidence supporting this conclusion is apparent in the examples provided in Tables 5 and 6. In Table 5, all the sentences are generated by Google Translate, yet most participants believed SFDA generated them. In Table 6, it is the other way around.

Table 5: Sample Google Translate's Sentences that Were Misperceived as SFDA Translations

Google Translate: ومن شأن هذه التدابير أن تساعد على حماية البيئة	
Participants who *correctly* identified Google Translate as the source:	10
Participants who *mistakenly* identified SFDA as the source:	20
Participants who opted out (i.e., chose "can't discern"):	3

SFDA Translation: فمن شأن هذه الإجراءات حماية البيئة	
Google Translate: لا تستخدمه إلا إذا كان المحلول واضحًا والختم سليمًا	
Participants who *correctly* identified Google Translate as the source:	9
Participants who *mistakenly* identified SFDA as the source:	20
Participants who opted out (i.e., chose “can’t discern”):	4
SFDA Translation: لا تقم بالاستعمال إلا إذا كان المحلول صافيًا وكان الغطاء سليمًا	
Google Translate: يجب إيقاف قطرات العين أوبتيزولين عند أول ظهور لطفح جلدي أو أي علامة أخرى لتفاعل فرط الحساسية	
Participants who *correctly* identified Google Translate as the source:	10
Participants who *mistakenly* identified SFDA as the source:	19
Participants who opted out (i.e., chose “can’t discern”):	4
SFDA Translation: يجب إيقاف قطرات العين أوبتيزولين عند أول ظهور لطفح جلدي أو أي علامة أخرى لتفاعل فرط الحساسية	

Table 6: Sample SFDA Translations that Were Misperceived as Google Translate Sentences

SFDA Translation: ينبغي استخدام الكميات التي جُهزت للتسريب على الفور، ولكن إذا لم يكن ذلك ممكنًا، يمكن في ظروف معينة تخزينها لمدة تصل إلى ٣٠ يومًا في الثلاجة شريطة تحضيرها بطريقة تمنع التلوث الجرثومي	
Participants who *correctly* identified SFDA as the source:	9
Participants who *mistakenly* identified Google Translate as the source:	22
Participants who opted out (i.e., chose “can’t discern”):	2
Google Translate: ينبغي استخدام الحقن المحضرة على الفور، ومع ذلك، إذا لم يكن ذلك ممكنًا، فيمكن، في ظروف معينة، تخزينها لمدة تصل إلى 30 يومًا في الثلاجة بشرط أن يتم تحضيرها بطريقة تستبعد التلوث الميكروبي	
SFDA Translation: بشكل خاص إذا كنت تتناول أي من الأدوية التالية، سيقوم الطبيب بمراقبتك للتأكد من الأدوية التي تتناولها تعمل بشكل مناسب، عند البدء بتناول أوروتكس	
Participants who *correctly* identified SFDA as the source:	7
Participants who *mistakenly* identified Google Translate as the source:	23
Participants who opted out (i.e., chose “can’t discern”):	3

Google Translate: على وجه الخصوص، إذا كنت تتناول أيًا من الأدوية التالية، فقد يرغب طبيبك في مراقبتك للتأكد من أن أدويتك تعمل بشكل صحيح، بمجرد البدء في تناول أوروتيكس

SFDA Translation: إذا أخبرت من قبل الطبيب أنك تعاني من عدم القدرة على تحمل بعض أنواع السكريات، قم بالاتصال مع الطبيب قبل البدء بتناول هذا الدواء

Participants who *correctly* identified SFDA as the source: 10

Participants who *mistakenly* identified Google Translate as the source: 23

Participants who opted out (i.e., chose "can't discern"): 0

Google Translate: إذا أخبرك طبيبك بأنك غير قادر على تحمل بعض السكريات، اتصل بطبيبك قبل تناول هذا المنتج الطبي

The differences between Google Translate's output and SFDA official translations are not always errors. The fact that the participants could not clearly and consistently identify whether Google Translate or SFDA generated a sentence means that both versions of the sentence are meaningful and sound natural to end-users. This can also be seen in the examples in Tables 5 and 6.

In the 760 sentences I analyzed following Tezcan et al.'s (2018) translation error typology, 595 (78.4%) were error-free despite being different from the official translations. Table 7 shows some of those sentences. As for those sentences with errors, 29.7% (49 out of 165) had accuracy errors. None of the accuracy errors had to do with addition or omission; instead, the errors were distributed as follows: 33 mistranslation errors, 12 terminology errors, and 5 untranslated words.

Table 7: Official Translations and Google Translate's Output to Show that Differences Do not Always Indicate Errors

Source Text	Official Translations	Google Translate
If you are taking a medicine containing nelfinavir (used for HIV infection).	إذا كنت تتناول دواء يحتوي على نيلفيناڤير (يستعمل لعلاج التهاب ڤيروس نقص المناعة المكتسبة).	إذا كنت تتناول دواء يحتوي على نلفينافير (المستخدم لعلاج الإصابة بفيروس نقص المناعة البشرية).
Subacute cutaneous lupus erythematosus (SCLE): Proton pump inhibitors are associated with very infrequent cases of SCLE. If lesions occur, especially in sun-exposed areas	الذئبة الحمامية الجلدية شبه الحادة ترتبط مثبطات مضخة البروتون بحالات نادرة جداً من الذئبة الحمامية الجلدية شبه الحادة، في حال حدوث آفات، خاصة في مناطق الجلد المتعرضة للشمس، وإذا كانت	الذئبة الحمامية الجلدية تحت الحادة (SCLE): ترتبط مثبطات مضخة البروتون بحالات نادرة جدًا من الذئبة الحمامية الجلدية. في حالة حدوث آفات، خاصة في المناطق المعرضة للشمس من الجلد، وإذا كانت مصحوبة

of the skin, and if accompanied by arthralgia, the patient should seek medical help promptly and the health care professional should consider stopping this medication.	مصحوبة بألم مفصلي، يجب على المريض طلب المساعدة الطبية فوراً، ويجب على مقدم الرعاية الصحية النظر في وقف استعمال هذا الدواء.	بألم مفصلي، يجب على المريض طلب المساعدة الطبية على الفور ويجب على أخصائي الرعاية الصحية أن يفكر في إيقاف هذا الدواء.
Medicines that are used to thin your blood, such as warfarin or other vitamin K blockers.	أدوية تستعمل للوقاية من تجلط الدم، مثل الوارفارين أو حاصرات ڤيتامين ك الأخرى.	الأدوية التي تستخدم لتسييل الدم، مثل الوارفارين أو حاصرات فيتامين ك الأخرى.
Remember to also mention any other ill-effects like pain in your joints.	لا تنسى ذكر الآثار المرضية الأخرى مثل: آلام المفاصل.	تذكر أيضًا أن تذكر أي آثار سيئة أخرى مثل الألم في المفاصل.
However, your doctor may give you a further dose of 50 IU to 100 IU (0.5 to 1 mg) for every kilogram of your body weight, if necessary.	ومع ذلك، فإن طبيبك قد يعطي لك جرعة إضافية من 50 وحدة دولية الى 100 وحدة دولية (0.5 إلى 1 ملغ) لكل كيلوغرام من وزن الجسم الخاص بك، إذا لزم الأمر.	ومع ذلك، قد يعطيك طبيبك جرعة إضافية تتراوح من 50 إلى 100 وحدة دولية (0.5 إلى 1 مجم) لكل كيلوغرام من وزن جسمك، إذا لزم الأمر.

Some mistranslation errors were significant as they could impact the comprehension of package inserts. For instance, in one medication, a listed side effect was ‘aggression,’ which Google Translated as ‘عدوان’ instead of ‘عدوانية.’ Similarly, ‘hives’ was translated by Google Translate as ‘خلايا النحل,’ which is not suitable in this medical context and may lead patients to overlook this side effect. In SFDA’s official translation, ‘hives’ was rendered as ‘شرى,’ and it can also be translated as ‘ارتكاريا,’ ‘طفح جلدي,’ or ‘ طفح جلدي على شكل خلايا النحل.’ Other mistranslation errors were less critical. For example, ‘replace’ was translated as ‘ضع’ instead of ‘بدل’ or ‘استبدل,’ which may cause some confusion among patients but is not life-threatening.

Some translations may be difficult to comprehend when it comes to terminology errors. For instance, ‘regurgitation’ was translated by Google Translate as ‘القلس.’ While this translation exists in certain medical dictionaries like *The Unified Medical Dictionary* by the World Health Organization, it is unlikely to be understood by the general public. It is better translated as ‘ارتجاع’ or ‘ارتجاع في المرئ.’ Additionally, Google Translate sometimes exhibited inconsistency when translating the same term. For example, in the same package insert, ‘STEMI (ST-segment Elevation Myocardial Infarction)’ was at times translated as ‘احتشاء عضلة القلب,’ while at other times, it was left untranslated.

Accuracy errors remain relatively small compared with fluency errors, as 126 of the 165 sentences (76.4%) had one or more fluency errors. The most prevalent fluency errors, accounting for 68 out of 126 instances, were stylistic. These errors were associated with verbose expressions, redundant phrases, and repeated words. For example, Google Translate used the lengthy expressions of ‘محلول عن طريق الفم’ and ‘محلول فموي’ instead of ‘شراب’ to translate ‘syrup.’ Stylistic errors also encompassed the inclusion of words and phrases that did not enhance the overall meaning, such as the translation of ‘Treat blood clots that are in your blood’ into ‘ علاج جلطات الدم التي تكون في دمك,’ where the relative clause ‘التي تكون في دمك’ adds no new information and would be better omitted. Additionally, stylistic errors manifested in the unnecessary repetition of words (refer to Table 8).

Table 8: Examples of Unnecessarily Repeated Words in Google Translate’s Output

Source Text	Official Translation	Google Translate
ENOXA is usually given by injection underneath the skin (subcutaneous).	عادة ما يتم إعطاء إنوكسا عن طريق الحقن تحت الجلد.	يتم إعطاء إنوكسا عادةً عن طريق الحقن تحت الجلد (تحت الجلد).
feeling sick (nausea)	الشعور بالمرض والغثيان	الشعور بالغثيان (الغثيان)
The National Pharmacovigilance and Drug Safety Center (NPC)	المركز الوطني للتيقظ والسلامة الدوائية	المركز الوطني للتيقظ الدوائي والسلامة الدوائية (NPC)

Note: unnecessarily repeated words are underlined

After stylistic errors, lexical errors emerged as the second most frequent fluency issues, featuring 22 unnatural or uncommon collocations. For instance, ‘crush the capsules’ was translated as ‘تسحق الكبسولات’ rather than the more natural collocation ‘تطحن الكبسولات.’ Similarly, the phrase ‘as advised by your doctor’ was rendered as ‘حسب نصيحة الطبيب,’ though more naturally fitting collocations would be ‘ حسب تعليمات الطبيب’ or ‘حسب إرشادات الطبيب.’ Additionally, ‘a clear solution’ was translated as ‘ محلول واضح,’ not as ‘محلول صافي.’

Grammatical fluency errors ranked third and were distributed as follows: 22 word-form errors, 8 function word errors, and 4 agreement errors. Word form errors were clear in lists such as Table 9. The list illustrated instructions on self-injection. Ideally, each new item on the list should start with the same word form. However, Google Translate correctly used the imperative verb forms for the first and third instructions on the list but used a noun form for the second instruction.

Table 9: An Example of Inconsistent Word Forms Rendered by Google Translate in Lists

Source Text	Official Translation	Google Translate
Carefully pull off the needle cap from the syringe.	اسحب غطاء الإبرة بعناية من المحقنة.	اسحب غطاء الإبرة بعناية من المحقنة.
Throw away the cap.	تخلص من الغطاء.	رمي بعيدا الغطاء.

Do not press on the plunger before injecting yourself to get rid of air bubbles.	لا تضغط على المكبس قبل أن تحقن نفسك للتخلص من فقاعات الهواء .	لا تضغط على المكبس قبل حقن نفسك للتخلص من فقاعات الهواء.

5. Limitations of the Study

No information is available on the methods used to generate the SFDA's translations, whether the package inserts were manually translated from scratch or initially generated by a machine translation system and then post-edited. Similarly, SFDA's website does not provide details about the guidelines for translation or post-editing. The lack of clarity regarding the process for generating the SFDA's translations may raise questions about using them as a benchmark for evaluating Google Translate. Nevertheless, the SFDA's translations remain one of the few publicly accessible resources that offer translations for medical documents. As the SFDA is an official organization in the Kingdom of Saudi Arabia, where Arabic is the primary language with approximately 18 million native speakers (Saudi Census, 2022), these translations can still be considered official and serve as a reference. If other resources providing translations for medical documents in English and Arabic become available, the study could be replicated to ensure that the conclusions drawn are accurate and generalizable.

6. Conclusion and Implications

This study quantitatively and qualitatively analyzed Google Translate's output against the official SFDA translations in 50 package inserts translated from English into Arabic. The analysis showed that on a statistical level, the translations of Google Translate were understandable, and the meanings were clear. However, it also showed that Google Translate's output differed from official translations. The qualitative analysis showed that not all differences indicated errors; in the survey, end-users managed to differentiate between official and machine translations with an accuracy rate of only 46%. Furthermore, looking into 760 random sentences, only 21.7% of the sentences were identified as containing errors, and out of those sentences, most errors were fluency-related rather than accuracy-related. Such fluency errors make the translations sound weird or unnatural or do not read very smoothly, but they do not pose risks to end-users; they do not alter meanings or omit crucial information.

The results of this study contribute to dispelling stereotypes surrounding machine translation in the medical field, at least as far as package inserts are concerned. Accepting that Google Translate does not pose serious risks and that its output might be different yet still correct can help translators move forward by encompassing it into their workflow. They might adapt it to the textual genres they work on using options like Google AutoML and openly offer Google Translate and post-editing as a cost-effective option for their clientele.

Notes

[1] https://sotoor.ai/en/home

[2] https://huggingface.co/spaces/evaluate-metric/sacrebleu

[3] https://pypi.org/project/pyter/

[4] https://github.com/m-popovic/chrF

References

Abdelhay, M., Mohammed, A., & Hefny, H. A. (2023). Deep learning for Arabic healthcare: MedicalBot. *Social Network Analysis and Mining, 13,* article no. 71. https://doi.org/10.1007/s13278-023-01077-w

Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2017). Arabic language sentiment analysis on health services. *Proceedings of the* 1st *International Workshop on Arabic Script Analysis and Recognition (ASAR)* (pp. 114–118). https://doi.org/10.1109/ASAR.2017.8067771

Almahasees, Z., Meqdadi, S., & Albudairi, Y. (2021). Evaluation of Google Translate in rendering English COVID-19 texts into Arabic. *Journal of Language and Linguistics Studies, 17*(4), 2065 – 2080. https://doi.org/10.52462/jlls.149

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, *20*(6), 573–584. https://doi.org/10.1080/13645579.2016.1252188

Boudjellal, N., Zhang, H., Khan, A., Ahmad, A., Naseem, R., & Dai, L. (2020). A silver standard biomedical corpus for Arabic language. *Complexity, 2020*, Article ID 8896659. https://doi.org/10.1155/2020/8896659

Colina, S., Marrone, N., Ingram, M., & Sanchez, D. (2017). Translation quality assessment in health research: A functionalist alternative to back-translation. *Evaluation & Health Professions*, *40*(3), 267–293. https://doi.org/10.1177%2F0163278716648191

Costa-Jussà, M. R., Farrú, M., & Pons, J. S. (2012). Machine translation in medicine: A quality analysis of statistical machine translation in the medical domain. *Advanced Research in Scientific Areas*, 1995–1998.

Das, P., Kuznetsova, A., Zhu, M., & Milanaik, R. (2019). Dangers of machine translation: The need for professionally translated anticipatory guidance resources for limited English proficiency care-givers. *Clinical Pediatrics*,*58*(2), 247–249. https://doi.org/10.1177/0009922818809494

Doshi, K. (2021, May 9). *Foundations of NLP explained – Bleu score and WER metrics*. Medium. https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b

Ehab, R., Gadallah, M., & Amer, E. (2019). English-Arabic hybrid machine translation system using EBMT and translation memory. *International Journal of Advanced Computer Science and Applications, 10*(1), 195–203.

Google Cloud. (2024). Evaluating models. https://cloud.google.com/translate/automl/docs/evaluate

Jensen, M. N. (2013). *Translators of patient information leaflets: Translation experts or expert translators? A mixed methods study of lay-friendliness* [Ph.D. dissertation]. Aarhus, Denmark, Aarhus School of Business and Social Sciences.

Kadaoui, K., Magdy, S. M., Waheed, A., Khondaker, M. T. I., El-Shangiti, A. O., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Tarjamat: Evaluation of Bard and ChatGPT on Machine Translation on Ten Arabic Varieties [Preprint]. ArXiV Preprints. https://doi.org/10.48550/arXiv.2308.03051

Khoong, E. C., Steinbrook, E., Brown, C., & Fernandez, A. (2019). Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Internal Medicine,179*(4), 580–582. https://doi.org/10.1001/jamainternmed.2018.7653

Moberly, T. (2018). Doctors are cautioned against using Google Translate in consultations. *BMJ, 363*, Article k4546. https://doi.org/10.1136/bmj.k4546

Moreno-Sandoval, A. & Campillos-Llanos, L. (2013). Design and annotation of MultiMedica – A multilingual text corpus of the biomedical domain. In C. Vargas-Sierra (Ed.), *Corpus Resources for Descriptive and Applied Studies: Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013)*, 95 (pp.33 – 39). ScienceDirect. https://doi.org/10.1016/j.sbspro.2013.10.619

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive machine translation with large language models. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, & H. Moniz (Eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)* (pp. 227–237). European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.22

O'Brien, S. (2011). Towards predicting post-editing effort. *Machine Translation*, *25*(3), 197–215. https://doi.org/10.1007/s10590-011-9096-7

Ornia, G. F. (2016). *Medical brochures as a textual genre*. Cambridge Scholars Publishing. https://www.cambridgescholars.com/product/978-1-4438-8727-4

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, & P. Pecina (Eds.), *Proceedings of the 10th Workshop on Statistical Machine Translation* (pp.392–395). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W15-3049

Post, M. (2008). A call for clarity in reporting BLEU scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the 3rd Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W18-6319

Saudi Census. (2022). Saudi Census 2022. https://portal.saudicensus.sa/portal

Sharkas, H. (2019). Translation methods used in Arabic translations of medical patient information leaflets. In M. J., M. Taibi, & I. H.M. Crezee, *Multicultural health translation, interpreting and communication* (pp. 123–137). Routledge.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A study of translation error rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223–231). Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25

Tezcan, A., Hoste, V., & Macken, L. (2018). SCATE taxonomy and corpus of machine translation errors. In G. C. Pastor & I. Durn-Mouz (2018), *Trends in E-tools and resources for translators and interpreters* (pp. 219–244). Brill.

Zappatore, M. & Ruggieri. (2024). Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language, 84*, 101582. https://doi.org/10.1016/j.csl.2023.101582