Educational Administration: Theory and Practice
2024, 30(2), 1213 -1221
ISSN: 2148-2403
https://kuey.net/

Research Article

Comparing Human Translation And Google Translate: Enhancing Communication For Oral Health

Fatema Alhammadi[1*], Aziza Eldarrat[2], Basaruddin Bin Ahmad[3], Sarliza Yasmin Bt Sanusi[4], Mohd Zulkarnain Bin Sinor[5]

[1*]School of Dental Sciences Health, Universiti Sains Malaysia (USM), Kelantan, Malaysia
[2]College of Dentistry, University of Science & Technology of Fujairah (USTF), Fujairah, UAE
[3]School of Dental Sciences Health, Universiti Sains Malaysia (USM), Kelantan, Malaysia
[4]School of Dental Sciences Health, Universiti Sains Malaysia (USM), Kelantan, Malaysia
[5]School of Dental Sciences Health, Universiti Sains Malaysia (USM), Kelantan, Malaysia

Citation: Fatema Alhammadi et.al (2024) Comparing Human Translation And Google Translate: Enhancing Communication For Oral Health, *Educational Administration: Theory And Practice, 30(2),* 1213 -1221
Doi: 10.53555/kuey.v30i2.5235

ARTICLE INFO **ABSTRACT**

Computer software-based translation of texts from one language to another is assuming increasing importance in different fields. This study aims to assess the accuracy of Google Translate (GT) in translating English SOHO-5 (E-SOHO-5) into Arabic compared to a Human Translator (HT). We evaluated the quality of translations from GT and a professional HT, by comparing them to a reference translation created by a multidisciplinary expert committee. This assessment was conducted using the BiLingual Evaluation Understudy metric. The translations produced by GT were also assessed and edited by the expert committee. The findings of this study showed that human translation consistently outperformed GT in terms of BLEU scores across unigrams, trigrams, and tetragrams while GT outperformed HT in bigrams. The average BiLingual Evaluation Understudy score for human translation was 0.447, while GT achieved an average score of 0.441. GT exhibited lower accuracy compared to human translation. To achieve linguistic and cultural equivalence in research instruments, machine translation requires post-editing.

Keywords: Cross-cultural adaptation; SOHO-5, Human Translator (HT); Google Translate (GT); Arabic language; English language

Introduction

The increasing global cooperation in clinical research has led to an increased demand for international tools to assess health-related quality of life (HRQoL**)** in different countries (Berzon, 2016; Bullinger *et al.*, 1998). To apply an instrument in a population with different language and culture, it is crucial first to subject it to cross-cultural adaptation (CCA) and psychometric assessment. This ensures that the instrument possesses the required measurement properties for its intended application. (Beaton *et al.*, 2000). The CCA is process includes both a translation of individual words and sentences between the source and target languages and an adaptation to the culture of the target language (Beaton *et al.*, 2000; Epstein *et al.*, 2015).

The process of cross-cultural adaptation entails more than just translation; it also involves evaluating the quality of the translation. According to the International Quality of Life Assessment (IQOLA) project group, the evaluation of the quality of instruments translation should consider three criteria. First, the translation clarity through the use of comprehensible terms. Second, the use of common language and avoidance of technical expressions. Lastly, insurance of conceptual equivalence (Bullinger *et al.*, 1998).

Computer software-based translation of texts from one language to another has started in the 1990s (Al-kabi *et al.*, 2013; Taylor *et al.*, 2015; Turner *et al.*, 2015) and is assuming increasing importance in different fields(Lumeras and Way, 2017). Machine translators (MT) operate by referencing a corpus, which is a collection of text in the source language paired with its corresponding translation in the target language(Taylor *et al.*, 2015). MT has been found to be associated with reduced translation costs and shorter translation time (Balk *et al.*, 2013; Taylor *et al.*, 2015; Turner *et al.*, 2017). However, while MT tools have shown proficiency in translating general text like news articles, they often encounter difficulties when it comes to translating domain-specific text that incorporates specialized vocabulary such as scientific text related to the health and medical sector, including instruments used for collecting health-related data(Turner *et al.*, 2017).

The Scale of Oral Health Outcomes for 5-year-old children (SOHO-5) is an instrument designed to measure the Oral Health-Related Quality of Life (OHRQoL) in 5-year-old children in the English language. It includes reports from both children and parents(Tsakos *et al.*, 2012). It has been translated and cross-culturally validated into many languages, such as Portuguese(Abanto *et al.*, 2013), Indonesian(Rachmawati *et al.*, 2017), Chinese(Gao *et al.*, 2020), Spanish(Abanto *et al.*, 2013) (Abreu-Placeres *et al.*, 2017), Persian(Asgari and Kazemi, 2017), Bengali (Mishu *et al.*, 2021), Turkish(Bani, 2021), Thai (Ariyavutikul *et al.*, 2023), and Myanmar (Burmese-speaking) language (Min *et al.*, 2023).

We conducted a CCA process to adapt the E-SOHO 5 questionnaire into Arabic, following the recommended guidelines by Beaton et al. (2000), that went through five stages, namely: stage I: initial translation/ forward translation, stage II: synthesis of the translations, stage II: synthesis of the translations, stage IV: expert committee, stage V: test of the prefinal version, stage VI: submission of documentation to the developers or coordinating committee for appraisal of the adaptation process. However, in this article, we will focus on the initial stage and list some examples comparing what HT produced and what GT produced regarding translation accuracy. Our process involved forward translation by two bilingual native Arabic speakers proficient in English and the machine translator GT. The two translators had distinct profiles and backgrounds. One of them was a dentist familiar with the concepts being examined, while the other was an official translator without a dental or medical background was not familiar with the construct under study. The translated versions were then synthesized into unified versions by three expert members combine the two Arabic translated version into a single Arabic version independently. Back translation was performed by two native English speakers to ensure equivalence to the original English version to translate the single Arabic version back into English language. Then, all translations were reconciled into a final unified A-SOHO-5 questionnaire by a multidisciplinary expert committee: (consisting of the forward and backward translators, language professionals, research methodologist and health-care professionals) met and resolved any inconsistencies in previous stages and finalize a draft version of the assessment questionnaire. Apart from a limited number of reports, there is a shortage of comprehensive evaluations regarding the use of MT in healthcare (Ridha *et al.*, 2020; Taylor *et al.*, 2015) particularly in translating from English to Arabic.

Therefore, the aim of this study was to assess the accuracy of GT in translating E-SOHO-5 into Arabic language compared to HT.

Material and Method

The instrument

The E-SOHO 5 questionnaire comprises two versions: one for children and one for parents. This study primarily focused on the child questionnaire, as the questions in both versions are largely similar, with only a few items differing between them. The child versions of E-SOHO-5 includes an initial section that focuses on experiences of toothache (comprising three items). Following that, seven items assess if children have encountered difficulties related to their teeth during activities such as eating, drinking, speaking, playing, avoided from smiling due to tooth pain and avoided smiling due to the appearance of their teeth and sleeping. Two additional items were included in the questionnaire, namely global rating questions that assessed satisfaction with oral health and the occurrence of dental cavities. In this study, we evaluated the translation of the first ten items; pain experience questions and the seven items of E-SOHO-5 questionnaire. Due to the simplicity of the responses, which mostly consisted of one or two-word phrases, they were excluded from this study.

Detailed information about the E-SOHO-5 items and scoring for both child and parent versions can be found elsewhere (Tsakos et al., 2012).

Measures of translation accuracy

BiLingual Evaluation Understudy (BLEU)

We evaluated the quality of the translations generated by both the professional HT and GT, referred to as candidate translations, by comparing them to a gold standard. The gold standard consists of the final items produced by the multidisciplinary expert committee, known as the reference translation. This is the committee that reconciles the previous steps and approved the translated questionnaire to be used as draft in the first stage of the study. In final step of our CCA process, a multidisciplinary expert committee conducted a thorough comparison of all translation and back-translation versions, meticulously considering semantic, idiomatic, and cultural aspects relevant to the target population. Discrepancies in word choices were carefully discussed and resolved through consensus. Ultimately, the expert committee merged all versions of the questionnaire, leading to the creation of the unified A-SOHO-5 questionnaire. We utilized the insights and inputs from the expert committee deliberations to assess the accuracy of Google Translate.

The translated versions were then synthesized into unified versions by experts in public health, research, and dentistry. Backward translation to English was performed by bilingual native English speakers, and all translations were reconciled into a final unified Arabic SOHO-5 (A-SOHO 5) questionnaire by a multidisciplinary expert committee. We used the BiLingual Evaluation Understudy (BLEU), which is one of the

methods used to assess systems of MT(Al-kabi *et al.*, 2013; Papineni *et al.*, 2002). BLEU is founded on a fundamental concept that evaluates the quality of MT systems by measuring the proximity of the system's generated output to a reference translation performed by a professional human translator (Al-kabi *et al.*, 2013). For a BLEU implementor, the main programming task is to compare the words and phrases (n-grams) in the candidate translation with those in the reference translation (Papineni *et al.*, 2002) count the number of matches between them, and calculate the *precision score* of the translation by dividing the number of matches by the total number of word and phrase in the candidate translation (Al-kabi *et al.*, 2013; Papineni *et al.*, 2002). An n-gram refers to a subsequence of n-items within a given sequence of words, which can be characters, words, or sentences depending on the context. N-grams can vary in the number of words they contain, and each type is assigned a specific name. When the sizes of the n-grams are one, two, three, or four words, they are referred to as unigram, bigram, trigram, and tetragram, respectively(Al-kabi *et al.*, 2013).
To illustrate the method of extracting n-grams, below is an example using the first item from the E-SOHO 5 questionnaire translated by the human translator. We extracted unigrams, bigrams, trigrams, and tetragrams from the translation and then computed the precision score:
Original item: Did your teeth hurt when they were coming.

Reference translation:

هل"did" آلمتك"hurt you" أسنانك "your teeth” عندما " when" كانت "was" تظهر "erupting"؟

Candidate translation:

هل "did" كانت "was" تؤلمك "hurting you" أسنانك "your teeth" عند "when" خروجها "erupting"؟

Unigrams:
Reference:

هل "did", آلمتك "hurt you", أسنانك "your teeth", عندما "when", كانت "was" تظهر "erupting"؟

Candidate:

هل"did", كانت"was", تؤلمك"hurting you", أسنانك"your teeth", عند"when", خروجها"erupting"؟

Bigrams:

Reference:

هل آلمتك”did hurt you”, آلمتك اسنانك”your teeth hurt you“ , أسنانك عندما”your teeth when”, عندما كانت”when was”,
كانت تظهر”was erupting”؟

Candidate:

هل كانت”was it”, كانت تؤلمك”was hurting you”, تؤلمك أسنانك”you teeth hurting you”, أسنانك عند”your teeth when”,
عند خروجها”when erupting”؟

Trigrams:
Reference Trigram
هل آلمتك اسنانك”did your teeth hurt you”, آلمتك اسنانك عندما”your teeth hurt you when”, اسنانك عندما كانت ”your
teeth when they were”، عندما كانت تظهر”when they were erupting”؟

Candidate Trigrams:
هل كانت تؤلمك”did it hurt you”, كانت تؤلمك أسنانك”your teeth were hurting you”, تؤلمك أسنانك عند ”your teeth
hurting you when”, أسنانك عند خروجها”your teeth when erupting”؟

Tetragrams:
Reference:
هل آلمتك اسنانك عندما”did your teeth hurt you when”, آلمتك اسنانك عندما كانت”your teeth h you when”, اسنانك عندما
كانت تظهر” your teeth when they were erupting "

Candidate:
هل كانت تؤلمك أسنانك”did your teeth hurt you”, كانت تؤلمك أسنانك عند”your teeth hurt you when”, تؤلمك أسنانك عند
خروجها”your teeth hurt you when they were erupting”؟

We detected four unigrams(هل did) (أسنانك your teeth) (عند when) (كانت was) that occur in both the candidate and reference translations. The same procedure was repeated for the bigrams, trigrams and tetragrams for both candidate translations to get the precision scores.

Data analysis

To calculate the *precision score*, we counted the number of unigrams, bigrams, trigrams, and tetragrams that appeared simultaneously in both the candidate and reference translations. We then divided these counts by the total number of respective unigrams, bigrams, trigrams, and tetragrams in the candidate translation. For example, in the above example, we divided the number of common unigrams (four) by the total number of unigrams in the candidate translation (six) which yielded a precision score for the unigram of 0.66.

Next, we combined the previous precision score values in a single overall score (called BLEU-score). This is done in two steps by taking the geometric mean of the precision scores and then multiply the result by an exponential brevity penalty factor using the following formulae:
Formula (1) (Al-kabi *et al.*, 2013; Papineni *et al.*, 2002) explains the first step: computing the Brevity Penalty (BP) where r is the length of reference that has more common n-grams and c is the candidate translation length:

$$BP = \begin{cases} 1 & if\ c > r \\ e\left[1 - \frac{r}{c}\right] & if\ c \leq r \end{cases} \quad (1)$$

Formula (2) (Al-kabi *et al.*, 2013; Papineni *et al.*, 2002) explains the second step of computing the BLEU score using the BP from the first step, where N = 4 and uniform weights wn = (1/N).

$$BLEU = BP\ \times \exp\left[\sum_{n=1}^{N} W_n\ \ log\ P_n\right] \quad (2)$$

The BLEU score ranges from 0 to 1, where the translation that has a score of 1 is considered as identical to a reference translation(Al-kabi *et al.*, 2013; Papineni *et al.*, 2002).

Results

BLEU Score

Table 1 presents the ten original items from the E-SOHO 5 questionnaire that were included in the study, along with their translations by both HT and GT. The table also displays the finalized items agreed upon for inclusion in the A-SOHO 5 questionnaire by the expert committee.
The overall BLEU scores for unigrams, trigrams, and tetragrams were consistently higher for the HT compared to GT, while the BLEU score for the bigrams was higher for GT as indicated in Table 2. Overall, the HT demonstrated an average BLEU score of 0.447, whereas GT achieved an average score of 0.441 (Table 2). Notably, for both the HT and GT, the BLEU scores were higher for unigrams and lower for tetragrams (Table 2).
Table 3 illustrates the ranking of items from the SOHO 5 questionnaire based on the quality of translation, as indicated by their overall BLEU scores. Both the HT and GT obtained the highest BLEU scores for items number seven, ten, and five, and for both of them, items number two and three were among the three items with the lowest scores.

Table 1: Comparison of Original Items, Human Translated Items, Google Translated Items, and Final Expert-Approved Items for Scale of Oral Health Outcomes for 5-year-old Children

Original	Candidate translation	Reference
Item 1: Did your teeth hurt when they were *coming through?*	HT: هل كانت تؤلمك أسنانك عند خروجها ؟ Did your teeth hurt when they were coming out?" GT : هل أصيبت أسنانك عندما كانت تتأذى؟ " Did your teeth get injured when they were hurting?"	هل آلمتك /أزعجتك اسنانك عندا كانت تظهر؟ "Did your teeth hurt/bother you when they were erupting?"
Item 2: Do your teeth *hurt* now (other than when they were coming through)?	HT:هل تؤلمك أسنانك الآن بخلاف خروجها ؟ Do your teeth hurt now aside from their eruption?" GT: هل تتألم أسنانك الآن بخلاف عندما كانت تتألم؟ "Are your teeth hurting now besides when they used to hurt?"	هل تؤلمك/ تزعجك اسنانك الان غير وقت ظهورها؟ "Are your teeth hurting/bothering you now, aside from the time of their eruption?"
Item 3: Have your teeth ever *hurt* you (other than when they were coming through)?	HT: هل تؤلمك أسنانك بخلاف خروجها ؟ "Do your teeth hurt, aside from when they come out?" GT : هل سبق لك أن أضرت بك أسنانك بخلاف ما كانت عليه في السابق؟ "Have your teeth ever harmed you, aside from how they used to be before?"	هل آلمتك /أزعجتك اسناك سابقاً غير وقت ظهورها؟ "Did your teeth hurt/bother you previously, aside from the time of their eruption?"
Item 4: Has it ever been hard for you to eat because of your teeth?	HT: هل كان من الصعب عليك أن تأكل بسبب أسنانك؟ "Was it difficult for you to eat because of your teeth?" GT: هل كان من الصعب عليك تناول الطعام بسبب أسنانك؟ "Was it difficult for you to take a meal because of your teeth?"	هل كان من الصعب عليك ان تاكل في اي وقت بسبب اسنانك؟ "Was it difficult for you to eat at any time because of your teeth?"

Item 5: Has it ever been hard for you to drink because of your teeth?	HT: هل كان من الصعب عليك الشرب بسبب أسنانك؟ "Was it difficult for you to drink because of your teeth?" GT : هل كان من الصعب عليك الشرب بسبب أسنانك؟ "Was it difficult for you to drink because of your teeth?"	هل كان من الصعب عليك ان تشرب في اي وقت بسبب اسنانك؟ "Was it difficult for you to drink at any time because of your teeth?"
Item 6: Has it ever been hard for you to *speak* because of your teeth?	HT: هل كان من الصعب عليك الكلام بسبب أسنانك؟ "Was it difficult for you to talk because of your teeth?" GT : هل كان من الصعب عليك التحدث بسبب أسنانك؟ "Was it difficult for you to speak because of your teeth?"	هل كان من الصعب عليك الكلام في اي وقت بسبب اسنانك؟ "Was it difficult for you to talk at any time because of your teeth?"
Item 7 :Has it ever been hard for you to play because of your teeth?	HT : هل كان من الصعب عليك اللعب بسبب أسنانك؟ " is "Was it difficult for you to play because of your teeth?" GT: هل كان من الصعب عليك اللعب بسبب أسنانك؟ " is "Was it difficult for you to play because of your teeth?"	هل كان من الصعب عليك اللعب في اي وقت بسبب اسنانك؟ "Was it difficult for you to play at any time because of your teeth?"
Item 8: Have ever not smiled because your teeth were hurting	HT: هل لم تكن تستطع ان تبتسم من قبل لأن أسنانك كانت تؤلمك؟ Were you unable to smile before because your teeth were hurting? GT : هل لم تبتسم من قبل لأن أسنانك كانت تؤلمك؟ "Did you not smile before because your teeth were hurting?"	هل حدث سابقا ان منعك الم اسنانك من ان تبتسم؟ "Has it happened before that the pain in your teeth prevented you from smiling?"
Item 9: Have ever not smiled because of how your teeth look?	GT: هل لم تكن تستطيع ان تبتسم من قبل بسبب شكل أسنانك؟ Were you unable to smile before because of the shape of your teeth? HT: هل لم تبتسم من قبل بسبب شكل أسنانك؟ Did you not smile before because of the shape of your teeth?	هل حدث سابقا ان منعك شكل اسنانك من ان تبتسم؟ "Has it happened before that the shape of your teeth prevented you from smiling?"
Item 10: Has it ever been hard for you to sleep because of your teeth?	HT: هل كان من الصعب عليك النوم بسبب أسنانك ؟ Was it difficult for you to sleep because of your teeth?" GT: هل كان من الصعب عليك النوم بسبب أسنانك؟ Was it difficult for you to sleep because of your teeth?"	هل كان من الصعب عليك النوم في اي وقت بسبب اسنانك؟ "Was it difficult for you to sleep at any time because of your teeth?"

GT: Google Translate, HT: Human translator

Table 2: Comparison between human translation and Goole Translate translation of the Scale of Oral Health Outcomes for 5-year-old children

	Unigram	Bigram	Trigram	Tetragram	Overall
Human Translator	0.728	0.602	0.544	0.496	0.447
Google Translate	0.703	0.605	0.539	0.459	0.441

Table 3: The overall Bleu score for the Scale of Oral Health Outcomes for 5-year-old children 10 translated items by human translator and Google translate

Item number	Human translator		Item number	Google translate
Seven	0.83		Seven	0.83
Ten	0.83		Ten	0.83
Five	0.82		Five	0.82
Six	0.69		Six	0.53
Four	0.43		Four	0.43
Eight	0.27		Nine	0.35
Nine	0.24		One	0.33
Two	0.4		Eight	0.29
One	0		Two	0
Three	0		Three	0

Expert committee inputs

In seven out of the ten items evaluated, the expert committee identified either semantic or grammatical errors made by GT that could have altered the meaning and understanding for the target population. These are listed below (**see table 1**):

Item number one (Did your teeth hurt when they were coming through?):

- The term "hurt" was chosen to depict tooth pain in E-SOHO-5 based on parental input during the development process (Tsakos et al., 2012). GT provided the translation as "أصيبت," a literal rendering of "hurt" into Arabic, implying "injured" when translated back into English. The more appropriate Arabic term for describing tooth pain is "ألم," and this was selected by the HT. The expert committee concurred on using this term in the draft version of A-SOHO 5.
- The term "coming through" was inaccurately translated by GT into the Arabic word "تتأذى," which means "injured." The HT correctly translated it as (خروج).

Item number two: "Do your teeth hurt now (other than when they were coming through)?

- The term "coming through" was incorrectly translated by GT into the Arabic word "تتألم," which means "in pain." The HT accurately translated it as (خروج).
- GT rendered "your teeth hurt" as the Arabic phrase "تتألم أسنانك," implying that the teeth themselves are in pain, not the child because of their teeth. The expert committee revised this to "تؤلمك اسنانك," which, when translated back into English, conveys the meaning of "causing you pain" The HT, adopted this refined translation.

Item number three: Have your teeth ever hurt you (other than when they were coming through)?

- The term "hurt," employed to describe pain or ache in E-SOHO 5, was translated by GT into the Arabic word "أضرت," representing another literal translation of the word "hurt" or "harm." The expert committee opted for the term "آلمتك," signifying "caused pain or ache" in Arabic. The HT appropriately incorporated this term.
- GT incorrectly rendered the phrase "coming through" as "بخلاف ما كانت عليه في السابق," which means "Unlike what it was before" in English. The HT accurately translated the phrase "coming through" to the Arabic word "خروج."

Item number four: Has it ever been hard for you to eat because of your teeth?

- GT translated the phrase "to eat" into the Arabic phrase "تناول الطعام." This is a formal way of describing the act of eating and could be challenging for young children to understand. The expert committee opted to use the word "تأكل," which is a more informal way of expressing eating. The HT appropriately used this later word.

Item number six: Has it ever been hard for you to speak because of your teeth?

- GT translated the word "speak" into the Arabic word "الحديث." However, the expert committee members reached a consensus that the term "الحديث" is more formal for use in dialogue. Instead, the synonym "الكلام" is more common, understandable, and suitable for general use and circulation among the public. Therefore, it was decided to use "الكلام" instead. The HT appropriately used this word.

Item number nine: Have ever not smiled because your teeth were hurting?
Item number ten: Have ever not smiled because of how your teeth look?

- In both these items, the phrase "Have you ever not smiled," achieves explicit negation in the English version by adding "not" before the verbs. GT translated it into Arabic using explicit negation as well. However, to ensure clarity in the Arabic translation, the expert committee decided to employ implicit negation by using the word "منعك" (meaning "prevented you") before the word smile in Arabic. The HT has also translated the phrase while keeping the explicit negation in Arabic by adding the phrase (لم تستطع) meaning (could not).

In the nine errors identified above and attributed to GT, four were due to culturally inappropriate use of words or phrases, three were errors of word sense (where the meaning of the word was incorrectly translated), and two were grammatical errors.

Discussion

The goal of our study was to assess the effectiveness of a MT software in the context of healthcare research. We compared the accuracy of GT to human translation and found that human translation demonstrated superior accuracy. This aligns with findings from other studies, indicating that human-generated translations generally outperform raw MT output (Papineni *et al.*, 2002; Turner *et al.,* 2015). In previous research, human translation was favored over machine translation due to its enhanced word order, a higher level of professionalism in reading level, a smoother flow in the translated text, precise word choices, preservation of the original meaning, and cultural appropriateness concerning the source document.(Turner *et al.,* 2015).

In our study, the gold standard reference against which we assessed the quality of GT was the set of items agreed upon by the expert committee participating in the CCA. This multidisciplinary committee consisted of members from various backgrounds related to both the context of the SOHO-5 and the target population. Based on this gold standard, the majority of errors made by GT involved the use of culturally inappropriate words and phrases. GT relies on statistical matching for translation, deviating from a conventional approach based on dictionaries and grammar rules. This method makes it susceptible to producing translations that may not make sense(Patil and Davies, 2014).

Also, unlike human, MT systems cannot make inferences or extrapolate beyond the fixed examples in their training data(pairs of source language sentences and their corresponding translations in the target language) making it a challenge for them to adapt to different and changing contexts(Lumeras and Way, 2017). Therefore, their output is generally a literal translation of words and phrases without taking into account the underlying cultural context. In reader-focused texts such as questionnaires intended for gathering information, a literal translation approach can pose challenges as it results in readability and comprehension difficulties, particularly for monolingual individuals in the target language(Colina *et al.*, 2019). The aim of literal translation is to achieve linguistic equivalence, which is not suitable when translating a questionnaire item intended to be read and comprehended by a target respondent(Colina *et al.*, 2019). The translation approach should be functionally communicative-oriented rather than linguistically literal. This is supported by trial results of translating a medical quality-of-life questionnaire, where the communicative translations were assessed as better and easier to comprehend by respondents(Colina *et al.,* 2019).

HT have the advantage of exercising judgment in determining the appropriate level of alignment with the source text and the extent of adaptation required to meet the researchers' goals. This judgment is based on factors such as the purpose of the translation, the researchers' requirements, and the contextual considerations(Colina *et al.,* 2019).

The findings of this study also indicated inaccuracies in the translation of certain words and phrases by GT. These errors, known as word sense errors, are a common issue in MT (Turner *et al.,* 2015) and are referred to as disambiguation, wherein MT systems struggle to choose the most appropriate word or phrase (Lumeras and Way, 2017).

These types of error can be acceptable in cases where the main goal is to achieve a general understanding of the text. However, when it comes to eliciting accurate responses, a higher level of accuracy is necessary. In their study to assess the accuracy of GT to translate medical phrases and words, Patil and Davies(2014) identified numerous translations that were entirely inaccurate and advised against using GT for taking consent for medical and surgical procedures or for research(Patil and Davies, 2014).

In this study, we observed grammatical errors in the translations generated by GT. This disparity can be attributed to the inherent differences between the English and Arabic languages in terms of their nature and structure. The accuracy of translation is significantly influenced by the characteristics of the original language of the text(Balk, Chung and Chen, 2013). For instance, translations from English to other European languages tend to be of higher quality compared to translations between English and other language families (Balk *et al.*, 2013; Patil and Davies, 2014;Taylor et al., 2015). This difference in quality could be attributed to the similar sentence structures between English language and European language (Turner *et al.*, 2015), or it may be due to the greater availability of reference texts and translating algorithms for languages more commonly used in computing, such as the European languages (Patil and Davies, 2014; Taylor et al., 2015).

In addition to that, Arabic presents substantial challenges for MT due to its complex morphological features, diverse word forms, and flexible word orders, allowing for multiple sentence expressions. The existence of various dialects and differences in word order between source and target languages further contribute to multiple interpretations for a given sentence(Al-kabi *et al.*, 2013).

Based on the aforementioned observations, it is evident that to ensure high-quality translation of research instruments in health or quality of life-related fields, MT must strike a delicate balance between excessively literal and overly pragmatic translations. Achieving this balance necessitates the involvement of human judgment. As highlighted in previous research, relying solely on MT does not guarantee effective communication of the intended message(Taylor *et al.*, 2015; Turner *et al.*, 2017). To ensure high-quality translations, human readers with domain expertise and fluency in both the source and target languages are needed to correct MT errors. This process, known as post-editing (PE), plays a crucial role in refining and improving the accuracy of translations(Turner *et al.*, 2015). Previous research has indicated that MT tools, such as GT, can be effectively combined with human post-editing to efficiently generate high-quality translations at a reduced cost(Turner *et al.*, 2015, 2017).

In our CCA, although the human translator, who was a professional but not familiar with the construct under study, outperformed GT based on the BLEU score, both required varying degrees of editing by the expert committee to achieve linguistic and cultural equivalence.

Conclusion

In our assessment of Google Translate's accuracy in translating the E-SOHO 5 questionnaire from English to Arabic, we compared it to human translation. Across most of the questionnaire items, the human translator consistently achieved higher scores than GT. The errors made by GT included culturally inappropriate use of words or phrases, word sense errors, and grammatical errors. These errors can be attributed to the tendency of MT to produce literal translations and its inability to adapt to different and evolving contexts. Additionally, the inherent differences between the English and Arabic languages in terms of their nature and structure contribute to these errors. Post-editing of translations by GT is necessary to ensure linguistic and cultural adaptation of the translated text.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Abanto J, Tsakos G, Paiva SM, Goursand D, Raggio DP, Bönecker M. 2013. "Cross-cultural adaptation and psychometric properties of the Brazilian version of the scale of oral health outcomes for 5-year-old children (SOHO-5)". *Health Qual Life Outcomes*. 11(137):1-7. doi: 10.1186/1477-7525-11-16.
2. Al-kabi, M., Hailat, T., Al-Shawakfa, E., Alsmad, I., 2013. "Evaluating English to Arabic Machine Translation Using BLEU. *International Journal of Advanced Computer Science and Applications* 4(1), 66–73. doi:10.1109/AEECT.2013.6716439.
3. Ariyavutikul, W., Jirarattanasopha, V., Duangthip, D., Gao, S. 2023. "Psychometric properties of the Thai version of the Scale of Oral Health Outcomes for 5-year-old children". *Int J Paediatr Dent*. 33(2), 113-123. doi: 10.1111/ipd.13026. Epub 2022 Jul 28.
4. Asgari, I. and Kazemi, E. 2017. "Cross-Cultural Adaptation of Persian Version of Scale of Oral Health Outcomes for 5-Year-Old Children". *J Dent (Tehran)*.14(1),48-54.
5. Balk, E., Chung, M. and Chen, M. 2013. "Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Languages". Methods Research Report. (Prepared by the Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-1.) Rockville, MD: Agency for Healthcare Research and Quality. January 2013. AHRQ Publication No. 12(13)-EHC145-EF
6. Bani, M., Akin, Y., Coşkun, A., Alacam, A. 2021. "Cross-Cultural Adaptation of the SOHO-5 and Impact of Caries and Trauma on the Quality of Life in Turkish Children". *Journal of Gazi University Health Sciences Institute*. 3(3), 105-112.
7. Beaton, D.E. Bombardier, C., Guillemin, F., Ferraz, M. 2000. "Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures". *Spine (Phila Pa 1976)*. 25(24), 3186-3191. doi: 10.1097/00007632-200012150-00014.
8. Berzon, R. Hays, R., Shumaker, S. 2016. "International use , application and performance of health-related quality of life instruments". *Quality of Life Research*. 2, 367-368. doi: 10.1007/BF00422214.
9. Bullinger, M. Alonso, J., Apolone, G., Leplège, A., Sullivan, M., Wood-Dauphinee, Barbara Gandek, Anita Wagner, Neil Aaronson, Per Bech, Shunichi Fukuhara, Stein Kaasa, John E Ware. 1998. "Translating Health Status Questionnaires and Evaluating Their Quality: The IQOLA Project Approach". *J Clin Epidemiol*. 51, 913–923. doi: https://org/10.1016/S0895-4356(98)00082-1.
10. Colina, S., Marrone, N. and Ingram, M. 2019. "Translation Quality Assessment in Health Research: Functionalist Alternative to Back-Translation". *Eval Health Prof*. 40(3), 267–293. doi:10.1177/0163278716648191
11. Epstein, J., Miyuki, R. and Guillemin, F. 2015. "A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus". *J Clin Epidemiol*. 68(4),435–41. doi: 10.1016/j.jclinepi.2014.11.021.
12. Gao, S., Chen, K., Duangthip, D., Chin, E., Lo, M. 2020. "Translation and validation of the Chinese version of the scale of oral health outcomes for 5-year-old children". *Int Dent J*. 70(3),201–7. doi: 10.1111/idj.12545. Epub 2020 Jan 16.
13. Mishu, M., Watt, R., Heilmann, A., Tsakos, G. 2021. "Cross cultural adaptation and psychometric properties of the Bengali version of the Scale of Oral Health Outcomes for 5 - year - old children (SOHO - 5)". *Health Qual Life Outcomes*.1–11. doi: 10.1186/s12955-021-01681-4.
14. Min, S., Duangthip, D., Gao, S., Detsomboonrat, P. 2023. "Cross-cultural adaptation and psychometric properties of the Myanmar version of the scale of oral health outcomes for 5-year-old children". *PLoS One*. 18(3). doi: 10.1371/journal.pone.0282880.
15. Lumeras, M., and Way, A. 2017. "On the Complementarity between Human Translators and Machine Translation". *Journal of Language and Communication in Business*. (56), 21–42. doi: https://doi.org/10.7146/hjlcb.v0i56.97200.
16. Papineni, K. Roukos, S., Ward, T., Zhu W. 2002. "B LEU : a Method for Automatic Evaluation of Machine Translation". In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Stroudsburg, PA, USA, pp. 311-318. doi:10.3115/1073083.1073135.
17. Patil, S. and Davies, P. 2014. "Use of Google Translate in medical communication". *BMJ* 2014;349. doi: https://doi.org/10.1136/bmj.g7392.
18. Rachmawati, Y., Pratiwi, A., Maharani, D. 2017. "Cross-cultural Adaptation and Psychometric Properties of the Indonesia Version of the Scale of Oral Health Outcomes for 5-Year-Old Children". *J Int Soc Prevent Communit Dent*. 7,S75-81. doi: 10.4103/jispcd.JISPCD_272_17.
19. Taylor, R.M. Crichton, N., Moult, B., Gibson, F. 2015. "A prospective observational study of machine

translation software to overcome the challenge of including ethnic diversity in healthcare research". *Nursing Open*. 14-23. doi: 10.1002/nop2.13.

20. Tsakos, G., Blair, Y., Yusuf, H., Wright, W., Watt, R., Macpherson, L. 2012. "Developing a new self-reported scale of oral health outcomes for 5-year-old children". *Health Qual Life Outcomes*. 10(62),1-8. doi: 10.1186/1477-7525-10-62.
21. Turner, A., Dew, K., Desai, L., Martin, N., Kirchhoff, K. P. 2015. "Machine Translation of Public Health Materials From English to Chinese : A Feasibility Study". *JMIR Public Health Surveill*. 1(2), e17. doi: 10.2196/publichealth.4779.
22. Turner, A., Bergman, M., Brownstein, M., Cole, K., Kirchhoff, K. 2014. "A Comparison of Human and Machine Translation of Health Promotion Materials for Public Health Practice: Time, Costs, and Quality". *J Public Health Manag Pract*. 20(5), 523–529. doi: 10.1097/PHH.0b013e3182a95c87.