جامعة الشارقة
UNIVERSITY OF SHARJAH

University of Sharjah Journal of Humanities & Social Sciences

A Refereed Scientific journal

Vol. 22, No. 4
Jumada al-Akhirah 1446 A.H. / Decmber 2025 A.D.

ISSN : 1996-2339

https//:doi.org/10.36394/jhss/22/4/23

Specialized Translation Quality
A Comparative Study of Google Translate, DeepL, and ChatGPT in Translating Food Safety Regulations from English to Arabic

Rim Altakarli(1)

Rania Al-Sabbagh(2)

Received on: 30-09-2024 **Accepted on: 16-12-2024**

Abstract:

This study quantitatively compares the translation performance of two neural machine translation models, Google Translate and DeepL, alongside the AI chatbot ChatGPT, in translating food safety regulations from the Codex Alimentarius Commission. These regulations contain domain-specific terminology related to nutrition, food processing, non-sentence-list items, and polysemous words. The evaluation uses BLEU, chrF++, TER, and COMET metrics. While Google Translate achieved the highest BLEU score (0.197), DeepL, which added Arabic support in 2024, outperformed Google Translate and ChatGPT across all other metrics, leading to its selection for manual analysis using the Multidimensional Quality Metrics error typology. DeepL's output revealed four types of errors: fluency, accuracy, style, and terminology, but it also demonstrated a level of eloquence and readability comparable to human translations. The findings

(1) College of Arts - Humanities and Social Sciences (Sharjah - University of Sharjah)
u20104360@sharjah.ac.ae

(2) College of Arts - Humanities and Social Sciences (Sharjah - University of Sharjah)

suggest that DeepL should be considered an alternative to Google Translate and can serve as a new baseline for Arabic machine translation research. Moreover, despite its popularity, ChatGPT lags in specialized translation tasks. Finally, professional translators should consider integrating machine translation systems into their workflows to enhance efficiency in general domains and specialized fields.

Keywords: machine translation, specialized translation, English–Arabic translation, Google Translate, DeepL, ChatGPT

1. Introduction

Many machine translation models are available today for translation professionals and businesses. Some, like Google Translate and DeepL, are neural models specifically designed for machine translation. Other models, such as GPT (Generative Pre-Trained Transformer), the technology behind the widely used chatbot ChatGPT, are large language models designed to perform various natural language processing tasks, including machine translation. Choosing the best model can be a challenging task. These models are primarily trained for general-purpose translation, such as newspaper articles, and their performance in specialized fields, including medical, legal, or technical translation remains uncertain. Therefore, before investing in any of these models, stakeholders must assess which produces the most understandable output for their specialized translation needs, what types of errors it generates, and whether those errors follow a pattern that can be easily corrected in post-editing or require in-depth evaluation. This is why many studies in the machine translation literature focus on evaluating these models (Freitag et al., 2021; Maruf et al., 2021; Ranathunga et al., 2023; Rivera-Trigueros, 2022).

This study compares the performance of three machine translation models in translating the Codex Alimentarius food safety regulations from English into Arabic. These regulations include a comprehensive set of standards, guidelines, and codes of practice that address various food safety concerns, including microbiological hazards, pesticide residues, food additives, and contaminants such as heavy metals. They are crucial in minimizing foodborne diseases and protecting public health globally. The comparison involves three models: Google Translate, one of the most widely used machine translation systems, which has been translating to

and from Arabic for decades and served about 1 billion users in 2021 (Pitman, 2021); DeepL, which added Arabic to its supported languages in January 2024 (DeepL, 2024); and ChatGPT-4.0, a widely used chatbot with 200 million weekly users worldwide (Arab Times Kuwait, 2024). The comparison begins with a quantitative evaluation, measuring each model's output against reference translations from the Codex Alimentarius using four metrics: BLEU (Papineni et al. 2002), TER (Snover et al., 2006), chrF++ (Popović, 2015), and COMET (Rei et al., 2020), the latter being a recently introduced metric known for its strong correlations with human evaluations. Following the quantitative analysis, the output of the best-performing model is manually inspected following the error typology of the Multidimensional Quality Metrics (Lommel et al., 2014; Lommel, 2018).

This evaluative study contributes to the existing body of literature in three significant ways. First, it highlights the performance of DeepL. This model has been less explored in Arabic machine translation because Arabic was only recently incorporated into its list of supported languages. Understanding how DeepL compares to other models is important, especially given its reported superiority over Google Translate in translating Indo-European languages such as French and Polish (Esperança-Rodier & Frankowski, 2021), Spanish and German (Salinas & Burbataa, 2023), and Spanish and English (Peña Aguilar, 2023). Second, the study addresses a less explored textual genre—food safety regulations. These regulations, such as clinical discharge reports (Khoong et al., 2019) and patient information leaflets and health guidelines (Al-Sabbagh, 2024a), are rich in domain-specific terminology and stylistic features such as non-sentence list items and context-specific polysemous words. However, food safety regulations differ in their focus on highly technical terminology related

to the food industry rather than the medical field and include significant legal jargon outlining the consequences of non-compliance. Finally, this study's scope is broader than most evaluative studies, as it compares three machine translation models. In contrast, most studies evaluate a single model, typically Google Translate, as seen in works like Das et al. (2019) and Khoong et al. (2019). Some studies focus exclusively on ChatGPT, like Calvo-Ferrer (2023) and Hendy et al. (2023), or compare two models, Google Translate, with either ChatGPT or DeepL (Kubińska and Kubiński 2020; Piazzolla et al. 2023; Rescigno and Johanna Monti 2024) or ChatGPT to DeepL (Li, 2024; Noll et al., 2024).

2. Related Work

A substantial body of literature evaluates machine translation models across various languages and textual genres. For a broader overview, readers may refer to Al Shamsi et al. (2020) and Zappatore and Ruggieri (2024). However, despite this extensive literature, studies comparing multiple machine translation models for English and Arabic in a health-related context are lacking. Most studies focus on evaluating a single model, typically Google Translate (Almahasees et al., 2021; Al-Sabbagh, 2024a; Cornelison et al., 2021; Delfani et al., 2024; Ehab et al., 2018), and those that compare two models are generally conducted in non-Arabic contexts (Brewster et al., 2024; Noll et al., 2024; Rao et al., 2024). In this section, we will review such studies to provide a comprehensive overview of the current state of machine translation models in health-related texts, with particular emphasis on the English-Arabic language pair.

Almahasees et al. (2021) assessed Google Translate's performance in translating COVID-19 documents from international organizations

such as the World Health Organization, the United States Food and Agriculture Organization, and the European Center for Disease Prevention and Control. The researchers noted semantic, grammatical, lexical, and punctuation errors in the translated texts, which they claimed inhibited their intelligibility. However, their analysis lacked standard quantitative evaluation metrics like BLEU, chrF++, or TER, and they did not substantiate their claims with surveys or interviews to test the intelligibility among end-users. Furthermore, they did not provide examples indicating significant alterations in meaning that could impact end-users' understanding or safety, contrasting with Khoong et al. (2019), who noted such issues in English-to-Spanish and English-to-Arabic medical translations.

Al-Sabbagh (2024a) evaluated the efficacy of Google Translate by converting 50 English patient information leaflets into Arabic, using translations from the Saudi Food and Drug Authority's website as a benchmark. Google Translate achieved a moderate BLEU score of 0.255, a moderate chrF++ score of 51.13, and a high TER score of 59.7. Such results are much worse than those achieved by Google Translate for non-specialized texts such as news articles, as Kadaoui et al. (2023) reported a BLEU score of 66, a chrF++ score of 78.97, and a TER score of 28.6. Likewise, Moslem et al. (2023) reported a BLEU score of 44, a chrF++ score of 62, and a TER score of 0.58. Al-Sabbagh attributed the fact that the scores of Google Translate for patient information leaflets are much lower than the state-of-the-art results because of the abundance of English-Arabic parallel corpora with billions of words featuring general-purpose textual genres, such as news articles in contrast to the lack of large, specialized English-Arabic corpora that cover the medical domain.

Al-Sabbagh (2024a) further analyzed 760 sentences manually and found that 78.4% were error-free; they were different from the reference translations but they are not wrong like "If you are taking a medicine containing nelfinavir (used for HIV infection)" that was translated by Google Translate as "إذا كنت تتناول دواء يحتوي على نلفينافير (المستخدم لعلاج الإصابة بفيروس نقص المناعة البشرية)" and in the reference translation as "إذا كنت تتناول دواء يحتوي على نيلفيناڤير (يستعمل لعلاج التهاب ڤيروس نقص المناعة المكتسبة)" The terminology translation is different so the BLEU score is not perfect but both translations are valid. Of the sentences with errors, 29.7% contained significant inaccuracies, including mistranslations and incorrect medical terminology. Errors such as translating 'aggression' as 'عدوان' instead of 'عدوانية' and 'hives' as 'خلايا النحل' could seriously mislead patients about side effects. Google Translate also exhibited issues with stylistic and lexical fluency, producing awkward expressions such as 'تسحق الكبسولات' for 'crush the capsules' instead of the more natural 'تطحن الكبسولة'. Therefore, Al-Sabbagh concluded that Google Translate can generate helpful, reliable, comprehensible translations that should not be stereotyped as an inherently risky tool. However, it still requires improvement, either through rigorous human post-editing or by integrating extensive health-related parallel corpora that it can use to leverage its performance.

Cornelison et al. (2021) assessed the accuracy of Google Translate in translating usage directions and counseling points for the top 100 drugs in the United States into Arabic, Chinese (simplified), and Spanish. Two clinicians identified directions and counseling points and translated them using Google Translate. To evaluate the accuracy, two bilingual native speakers of each language, both non-clinicians, performed back-

translations into English, to check their clarity and accuracy in their native languages. Clinicians then reviewed these back-translations to determine the clinical significance of any inaccuracies. The results showed that out of 38 directions for use, 76.3% were accurately translated into Arabic, 89.5% into Chinese, and 71% into Spanish. The accuracy for the 170 counseling points was 54.1% for Arabic, 76.5% for Chinese, and 38.2% for Spanish. Importantly, 29.1% of the 247 inaccurate translations had significant clinical implications, including potential life-threatening errors. The study concluded that certified translators are essential for translating medical documents, and clinicians should be aware of the risks associated with using Google Translate.

Delfani et al. (2024) focused on assessing the performance of Google Translate in translating mental health information from English into Arabic, among other languages, using two primary datasets: the United Kingdom National Health Service dataset and the Royal College of Psychiatrists dataset. The translations were manually analyzed by native speakers for accuracy and comprehensibility, categorizing errors into terminology inaccuracies, syntactic/semantic errors, and other critical errors impacting patient safety. The National Health Service dataset revealed a high error rate across various categories in Arabic translations, especially inaccuracies in medical terminology, fluency issues, and critical errors that could mislead patients. For instance, ‘mantras’ were incorrectly translated as singing, and advice on practicing yoga and meditation was reversed to avoid these activities. In contrast, the translation of the Royal College of Psychiatrists dataset, which involved longer text segments, showed improved accuracy and fluency. This suggests that Google Translate performs better with longer texts, providing better contextual understanding.

Ehab et al. (2018) 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 احتقـان الرئـة (lung congestion) and خلـل لوظائـف الكلـى (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). However, 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.

Rao et al. (2024) compared ChatGPT-3.5 and Google Translate in translating two English-language documents: postoperative care instructions following circumcision and patient information on undescended testicles. These texts were translated into Russian, Vietnamese, and Spanish. Language experts evaluated the translations based on accuracy in conveying meaning, expression, and technical precision. Errors were categorized by type (meaning, flow, language, form, style guide, or terminology) and severity (minor, major, or critical). The primary focus of the evaluation was the accuracy of meaning retention in the translated texts. The analysis revealed that, out of 132 sentences, ChatGPT made errors in 3.8% of the Spanish translations, whereas GT had an error rate of 18.1%. ChatGPT and Google Translate had error rates of 35.6% and 41.6% for Russian

translations, respectively. In Vietnamese, ChatGPT incorrectly translated 24.2% of sentences, while GT made errors in 10.6%. In summary, ChatGPT outperformed Google Translate in Spanish translation but was less effective in Vietnamese.

Similar to the study by Rao et al. (2024), Brewster et al. (2024) compared ChatGPT-4.0 (a more recent version than that used by Rao et al.) with Google Translate in translating twenty pediatric discharge instructions from English into Spanish, Brazilian Portuguese, and Haitian Creole. The translations were evaluated based on adequacy (preservation of information), fluency (grammatical correctness), meaning (preservation of connotation), and severity (potential clinical harm). Both Google Translate and ChatGPT showed similar domain-level ratings compared. However, ChatGPT created more clinically severe errors than Google Translate.

Noll et al. (2024) compared GPT-3.5 and DeepL in translating medical terminology, with a special focus on Human Phenotype Ontology (HPO). Human medical experts evaluated 120 translated HPO terms using a 4-point Likert scale (1 = strongly agree, 4 = strongly disagree), with an independent reference translation from the HeTOP database serving as validation. GPT-3.5 received an average Likert rating of 1.29, while DeepL scored 1.37, indicating a high level of similarity between both models and the reference translations. Statistical analysis showed no significant differences between the models, confirming their comparable effectiveness in translating HPO terms from English to German.

3. Methods

3.1. Corpus

Codex food safety regulations are the standards, guidelines, and codes of practice developed by the Codex Alimentarius Commission, a joint program of the Food and Agriculture Organization of the United Nations and the World Health Organization. These regulations promote international food safety and quality, covering various aspects of the food supply chain, including food labeling, additives, contaminants, hygiene practices, and food safety management systems (Lee et al., 2021).

Similar to other health-related texts, food safety regulations are dense with terminology related to chemicals, food processing procedures, and additives. These regulations frequently utilize enumerated lists to outline requirements, standards, or banned substances, which may consist of phrases, complete sentences, or segments of complex sentences like those shown in Table 1. What distinguishes food safety regulations from other health texts is their inclusion of legal language—detailing enforcement mechanisms and penalties for non-compliance—and the prevalence of polysemous words. These words often carry one meaning in everyday usage but adopt specialized meanings in food processing. For instance, 'express' commonly describes conveying thoughts or actions directly and swiftly. However, food processing refers to the mechanical or chemical extraction of juices or oils from fruits, vegetables, or seeds. Similarly, 'agent' in common parlance denotes an individual acting on behalf of another, but in food processing, it refers to substances that produce specific effects.

Table 1: A corpus excerpt.

English Source Text	Arabic Reference Translation
Notwithstanding the provision in Section 4.2.3.1, pork fat, lard, and beef fat shall always be declared by their specific names.	مع مراعاة الأحكام الواردة في القسم 4-2-3-1 ، ينبغي دائما ذكر دهون الخنزير وشحم الخنزير وشحم البقر بأسمائها المحددة
For food additives falling in the respective classes and appearing in lists of food additives permitted for use in foods, the following functional classes shall be used together with the specific name or recognized numerical identification, such as the Class Names and the International Numbering System for Food Additives (CXG 36- 1989) as required by national legislation.	فيما يتعلق بالمواد المضافة إلى الأغذية التي تندرج في إطار مختلف الفئات، والتي تظهر في قوائم المواد المضافة إلى الأغذية المسموح باستخدامها في المواد الغذائية، ينبغي استخدام الفئات الوظيفية المذكورة فيما يلي بالاقتران مع اسم محدد أو رقم التعريف المعترف به مثل أسماء الفئات والنظام الدولي لترقيم المواد المضافة إلى الأغذية (CXG 36-1989) على النحو الذي تقتضيه التشريعات الوطنية
Acidity Regulator	منظمات الحموضة
Bulking Agent	عامل مضخم

The English and Arabic food safety regulations were sourced from the Commission's website in PDF format. The corpus encompasses four general food standards: the general standard for the labeling of pre-packaged food (CXS 1-1985), the general standard for fruit juices and nectars (Codex Stan 247-2005), the general standard for bottled/packaged drinking waters (excluding Natural Mineral Waters) (CXS 227-2001), and the general standard for the labeling of food additives when sold as such (Codex Stan 107-1981). Table 2 displays the statistics of the collected corpus.

Table 2 Corpus statistics.

Language	Word Tokens	Word Types	Average Sentence Length
English	10,488	2,111	words 16.9
Arabic	9,462	2,820	words 19.5
Total number of sentences: 535			

3.2. Quantitative Metrics

Four quantitative metrics are used for comparison: BLEU (Bilingual Evaluation Understudy; Papineni et al., 2002), chrF++ (character n-gram F-score; Popović, 2015), TER (Translation Edit Rate; Snover et al., 2006), and COMET (Crosslingual Optimized Metric for Evaluation of Translation; Rei et al., 2020). BLEU is the most widely used evaluation metric in machine translation literature, and therefore, it will facilitate comparing the results obtained in this study to those obtained elsewhere. It measures the similarity between machine and reference translations based on n-gram sequences while accounting for brevity. The scores range from 0 to 1, with 1 showing a perfect resemblance between machine and reference translations. BLEU is the most used quantitative evaluation metric in machine translation literature; therefore, it is important to compare the findings of this study with others. However, one drawback of BLEU is that it does not consider synonyms and word-order alternations. This poses a particular challenge 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++, TER, and COMET.

chrF++ (character n-gram F-score; Popović, 2015) operates at both word and character levels, offering advantages over BLEU. Unlike BLEU, it does not penalize translations for different word orders and gives partial credit for partially matching words, which is beneficial for morphologically rich languages like Arabic—for instance, translating 'must' as تجب (wajib; must) instead of يجب (yajib) still earns some credit. chrF++ computes the harmonic mean of precision and recall of character n-grams. It also introduces a penalty for excessively long translations, thus favoring longer but more accurate translations (Maučec & Donaj, 2019).

TER (Translation Edit Rate; Snover et al., 2006) is a proxy of the effort needed to transform the machine translation output into the reference translation. It quantifies the number of edit operations (insertions, deletions, substitutions, shifts) necessary to align the machine translation with the human reference. This metric estimates the post-editing effort, as these editing operations can be manually performed with a keyboard and mouse. The TER score ranges from 0 to 1, with lower scores indicating better quality (i.e., fewer edits required).

Finally, COMET(1) (Crosslingual Optimized Metric for Evaluation of Translation; Rei et al., 2020) is based on pre-trained transformer models, like BERT, which are trained on massive amounts of multilingual data. These models learn deep representations of language, capturing not just word-level patterns but also the broader context and meaning of sentences.

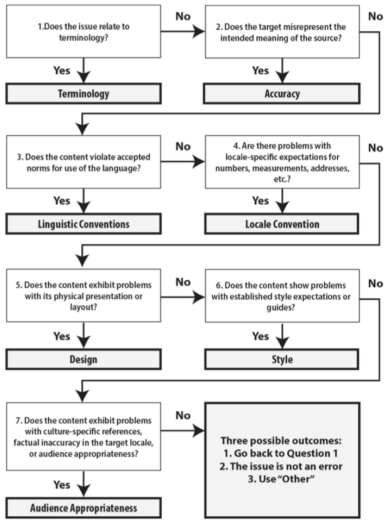
(1) https://github.com/Unbabel/COMET/tree/master

This allows COMET to understand synonyms, paraphrases, and different grammatical structures that convey the same meaning. Compared to other metrics, COMET's main advantage is its high correlation with human direct assessments.

3.3. Manual Inspection for Error Types

The output of the best-performing model is manually analyzed to identify the types of errors. For that purpose, we adopted the translation-error typology known as Multidimensional Quality Metrics (MQM) developed by Lommel et al. (2014) and Lommel (2018). We opted for MQM over alternative typologies due to the accessibility and practicality provided by its online platform(1). This website is a centralized repository furnishing clear definitions, well-defined annotation guidelines, and decision trees (as illustrated in Figure 1). MQM offers these resources to streamline the annotation process, guaranteeing clarity and consistency in our evaluation.

Figure 1 MQM decision tree



(1) https://themgqm.org/

Table 3: Definitions of the MQM error types as defined by Lommel et al. (2014), Lommel (2018), and Mariana (2014).

Dimension	Error Types	Specific Error Definitions
Terminology – Use of terms that are not correct, not adequate for the context, or not consistent with preferred terminology assets in the study.	[illegible]	[illegible]
	Inconsistency	Use inconsistent use of the same term, where consistency is required by the client, domain or internal guidelines.
Accuracy – The extent to which the target text accurately renders the meaning of the source text. Any deviation in translation from the source, whether omission, addition, or mistranslation, that affects the meaning.	Addition	Text in the target that does not exist in the source. Includes information that is not present in the source and is not justified by context.
	Omission	Content in the source that is missing from the target. Includes omitted words, phrases, sentences, or entire paragraphs that change the meaning.
	[illegible]	[illegible]
Linguistic Quality – Issues related to the correctness of the target text in terms of its language rules.	grammar	Errors in grammar, including: • Verb tense or aspect: The wrong verb form is used, such as using the past tense when the present tense is required. • Subject-verb agreement: The verb does not agree with the subject in number or person, such as using a singular verb with a plural subject. • Word order: Words are in the wrong order, such as placing an adjective after the noun it modifies in English. • Pronoun use: Incorrect use of pronouns, such as using the wrong pronoun case or failing to use a pronoun when required. • Preposition use: Incorrect use of prepositions, such as using the wrong preposition or omitting a preposition when required.
	spelling	Incorrect spelling of words.
	[illegible]	[illegible]
Style – Issues related to the overall quality of the writing style, including its appropriateness for the target audience and purpose. Style errors do not affect the meaning of the text but can make it awkward, unnatural, or difficult to read.	awkward	Text that is grammatically correct but reads awkwardly or unnaturally. This may include overly complex sentences, poor word choice, or inappropriate tone for the target audience.	
	inconsistent	Using inconsistent terminology, formatting, or writing style throughout the document.
Locale Convention – Issues related to the use of locale-specific conventions, such as formatting of dates, times, numbers, and currencies. Ensuring that the target text follows the conventions of the target locale is important for clarity and professionalism.	–	Incorrect number, currency, measurement, time, address, telephone, or other locale format for the locale.
Audience Appropriateness – Issues related to the appropriateness of the target text for the intended audience, including factors such as reading level, background knowledge, and cultural sensitivities. Ensuring that the target text is appropriate for the target audience is important for effective communication.	cultural reference	Text that is inappropriate, unclear, or confusing because it contains references that are culturally specific to the source culture and may not be understood by the target audience.
Design and Markup – Issues related to the visual design and presentation of translated content, including formatting, typography, and layout. Design errors can affect the readability and usability of the text and can detract from the overall quality of the translation.	Character formatting	Incorrect character formatting, such as font, size, color, bold, italic, underline, superscript, subscript, or other styles that are not consistent with the source text or the style guide.
	Tagging	Incorrect or missing tagging, such as HTML tags, that affects the structure or functionality of the text.
	Markup	Incorrect or missing markup, such as Markdown or LaTeX.
	Truncation/Text Expansion	Text that is too long or too short to fit within the available space, such as text that is cut off or has excessive white space.
	Missing Text	Text that is missing from the translation, such as a sentence or paragraph that was not translated.
	Layout	Incorrect or inconsistent layout, such as incorrect margins, line breaks, or page breaks.

4. Results

4.1. Comparative Results

All the translation outputs were generated between January and September 2024. Table 4 presents the quantitative evaluation scores for Google Translate, DeepL, and ChatGPT. While Google Translate achieved the highest BLEU score, the margin between its score and DeepL's was minimal, with a difference of just 0.008. However, DeepL outperformed the other models across the remaining three metrics. ChatGPT, on the other hand, had the lowest scores in all evaluation metrics. This outcome is noteworthy, especially considering that DeepL only added Arabic to its supported languages in 2024, making it a potential alternative to the widely used Google Translate for translators moving forward.

Table 4: Quantitative evaluation scores.

	BLEU	chrF++	TER	COMET
Google Translate	**0.197**	58.1	0.731	0.411
DeepL	0.189	**59.4**	**0.721**	**0.519**
ChatGPT	0.151	54.6	0.797	0.418

4.2. Manual Inspection Results

Four types of errors were identified in DeepL's output: accuracy, linguistic conventions, style, and terminology. Some categories from the MQM error typology did not apply to food safety regulations. For example, these regulations are scientific texts with minimal culture-specific references designed to serve as international standards. Additionally, the "design and markup" category was irrelevant, as the regulations were provided in PDF format, and DeepL does not yet support Arabic PDF translations. Instead,

translations into Arabic had to be done by copying and pasting the source text into the DeepL interface. ChatGPT also faced challenges in generating properly encoded Arabic PDFs, making it impossible to evaluate the design and layout of the translated documents. Furthermore, no significant errors in locale conventions were observed; numbers, measurements, and dates were generally translated correctly, except for two instances where DeepL used five "Y"s (YYYYY) for the year format.

4.2.1. Accuracy Errors

In addition to acronyms like GSFA (Codex General Standard for Food Additives), some words were transliterated instead of translated by DeepL. In Example 1, the system of measurement 'avoirdupois,' which is based on a pound of 16 ounces or 7,000 grains and widely used in English-speaking countries, was transliterated into الوحـدات الأفرو-أفوردبـوازي, which might be hard to understand for the Arabic speaker, so the reference translation did not just transliterate the word like DeepL. However, it also explained it between brackets as الرطـل (by bound). Google Translate also occasionally missed the translation of words, sometimes leaving them in English and other times transliterating them inconsistently. For instance, "Codex" was sometimes left in English, transliterated as كوديس or كودكس, and occasionally translated correctly as الدسـتور الغذائـي.

1	The net contents shall be declared in either (a) metric units or "Système International" units or (b) avoirdupois unless both systems of measurement are required explicitly by the country in which the food additive is sold.

.Ref	يتعين الإعلان عن المحتوى الصافي باستخدام النظام المتري الدولي أو النظام الافردوبويز (الرطل) أو باستخدام النظامين ما لم تقتض متطلبات البلد الذي يباع فيه الغذاء، على وجه التحديد، استخدام النظامين معا
DeepL	يجب الإعلان عن المحتويات الصافية إما (أ) بالوحدات المترية أو بوحدات "النظام الدولي" أو (ب) بالوحدات المترية أو بوحدات "النظام الدولي" أو (ب) بالوحدات الأفرو-أفوردبوازي، ما لم يكن كلا نظامي القياس مطلوباً على وجه التحديد من قبل البلد الذي تباع فيه المادة المضافة الغذائية

No additions were found in DeepL's output, but a few examples of omission and mistranslations were noted. One example of omission is "ingoing" in Example 2. As for Mistranslations, they occurred with polysemous words when the wrong sense of the word is selected. For instance, "claim" in Example 2 is translated as الإدعاء: although this is a generally accepted translation, in this context, it is better translated as التنويه (disclaimer). In the same example, "quality" was translated as نوعية (type) rather than a "feature."

2	The ingoing percentage of an ingredient (including compound ingredients or categories of ingredients), by weight or volume as appropriate, at the time of manufacture, shall be disclosed for foods sold as a mixture or combination where :the ingredient
.Ref	ينبغي، بالنسبة إلى المنتجات الغذائية التي تباع بشكل مزيج أو مجموعة من المكونات، أن تذكر النسبة المئوية الأولية من الحجم أو الوزن لكل مكون من مكوناتها (بما في ذلك المكونات المركبة أو فئات المكونات)، بحسب الاقتضاء عند التصنيع، وذلك في الحالات التالية:

DeepL	يجب الإفصاح عن النسبة المئوية للمكونات (بما في ذلك المكونات المركبة أو فئات المكونات)، بالوزن أو الحجم حسب الاقتضاء، وقت التصنيع، للأغذية التي تباع كمخلوط أو مزيج حيثما يكون المكون.

3	Claim" means any representation which states, suggests" or implies that a food has particular qualities relating to its origin, nutritional properties, nature, processing, composition .or any other quality
.Ref	يقصد بمصطلح «التنويه» بيان يذكر أو يشير أو يفيد ضمنيا بأن للمنتج الغذائي مزايا معينة ترتبط بمصدره أو خصائصه الغذائية أو طبيعته أو تجهيزه أو تركيبته أو أي ميزة أخرى.
DeepL	"الادعاء" يعني أي تمثيل ينص أو يوحي أو يوحي ضمناً بأن الغذاء له صفات معينة تتعلق بمنشأه أو خصائصه الغذائية أو طبيعته أو معالجته أو تركيبته أو أي نوعية أخرى.

Table 5 provides additional examples of mistranslation errors. For instance, "immediate" was translated as فـوري (instant) instead of مباشـر (direct), which would have been more accurate in this context. Additionally, "coined" and "fanciful" were translated as مصـاغ (made-up) and خيالـي (imaginative), neither of which fits the context or conveys the intended meaning. A common mistranslation involved rendering "foods" as طعـام or أغذيـة(food) instead of منتجـات غذائيـة (a food product), as in the reference translation. This can be considered an under-translation, as the former refers to all types of edible substances—whether raw, processed, or packaged. In contrast, المنتجـات الغذائيـة specifically refers to processed and packaged foods

prepared for sale.

Table 5: Examples of mistranslation errors.

Source text	Reference	DeepL
Foods for Cater-" ing Purposes" means those foods for use in restaurants, canteens, schools, hospitals and similar institutions where food is offered for immediate con- .sumption	ويقصد بعبارة «المنتجات الغذائية المخصصة لخدمات المطاعم» المنتجات الغذائية التي تستخدم في المطاعم والمقاصف والمدارس والمستشفيات وغيرها من المؤسسات المماثلة حيث تقدم المنتجات الغذائية للاستهلاك المباشر.	"أغذية لأغراض تقديم الطعام" تعني تلك الأطعمة المستخدمة في المطاعم والمقاصف والمدارس والمستشفيات والمؤسسات المماثلة التي يتم فيها تقديم الطعام للاستهلاك الفوري.
A "coined", "fanci- ful", "brand" name or "trademark" may be used, provided it accompanies one of the names provided in Subsections 4.1.1.1 to .4.1.1.3	يجوز استخدام اسم «مبتدع» أو «مستحدث" أو اسم "علامة مسجلة" أو "علامة تجارية" بشرط اقترانه بأحد الأسماء المذكورة في الأقسام الفرعية من 4-1-1-1 إلى 4-1-1-3.	ويجوز استخدام اسم "مصاغ" أو "خيالي" أو "علامة تجارية" أو "علامة تجارية" شريطة أن يكون مصحوباً بأحد الأسماء الواردة في الأقسام الفرعية من 4-1-1-1 إلى 4-1-1-3.

4.2.2. Linguistic Convention Errors

DeepL made errors in word forms, particularly in agreement between nouns and their modifying adjectives, subjects and verbs, or pronouns and their referents regarding gender, number, and definiteness. For instance, in Example 4, the noun العـائلات (the families) should be referred to using singular feminine pronouns rather than plural masculine ones. Also, the verbs should have been conjugated in the singular feminine form.

4	Consumer" means persons and families purchasing and" receiving food to meet their needs.
.Ref	ويقصد بمصطلح «المستهلك» الأشخاص والأسر التي تشتري المنتجات الغذائية أو تحصل عليها لتلبية احتياجاتها الشخصية.
DeepL	"المستهلك" يعني الأشخاص والعائلات الذين يشترون ويتلقون الأغذية لتلبية احتياجاتهم الشخصية.

DeepL's output had unnatural collocations. For instance, "crops husbandry" was translated as تربيـة المحاصيـل (literally: raising or bringing up crops). While "husbandry" can be translated as raising or bringing up, تربيـة المحاصيـل is an inappropriate collocation, as the more common phrase is زراعـة المحاصيـل (growing crops). Similarly, "acceptable for consumption" was rendered as مقبولا للاستهلاك (literally: accepted for consumption) instead of the more standard collocation قـابلا للاسـتهلاك (consumable). Additionally, translating "product durability" as متانـة المنتـج is incorrect when referring to food, as food typically collocates with صلاحيـة (expiration/shelf life) rather than "durability." A more accurate translation would be صلاحيـة المنتـج (product shelf life).

At times, DeepL followed the punctuation patterns of the source text, disregarding the differences in punctuation usage between English and Arabic, leading to awkward translations. In Example 5, using multiple commas appears unusual to Arabic readers, though it does not affect the overall comprehension. By contrast, the reference translation used only two commas, which is more appropriate in Arabic.

5	Instructions for use, including reconstitution, where applicable, shall be included on the label, as necessary, to ensure correct utilization of the food.
Ref.	تذكر تعليمات استخدام المنتج الغذائي بما يشمل تعليمات إعادة تكوينه حيثما ينطبق ذلك، على بطاقة التوسيم بحسب الاقتضاء، من أجل ضمان الاستخدام الصحيح للمنتج الغذائي.
DeepL	يجب أن تدرج تعليمات الاستخدام، بما في ذلك إعادة التكوين، عند الاقتضاء، على بطاقة الوسم، حسب الاقتضاء، لضمان الاستخدام الصحيح للغذاء.

Additionally, DeepL sometimes struggled with specific structures, such as “single ingredient product” in Example 6. This noun phrase requires a preposition like مـن (of) in Arabic, as in منتـج مـن مكـون واحـد (a product made of one ingredient). Omitting the preposition negatively affects readability and it can lead to confusion or misinterpretation. Similarly, DeepL mistranslated the phrase “water-extracted fruit juice” in Example 7, using the preposition من (from) instead of the correct prepositional phrase بواسـطة (by using), which better conveys the intended meaning.

6	When a single-ingredient product is prepared from an irradiated raw material, the label shall contain a statement indicating the treatment.
Ref.	عندما يتم إعداد منتج من مكون واحد انطلاقا من مادة خام تمت معالجتها بالإشعاع، ينبغي أن يرد في بطاقة توسيم المنتج بيان يشير إلى تلك المعالجة.
DeepL	عندما يتم تحضير منتج مكون واحد من مادة خام تم تشعيعها، يجب أن يحتوي ملصق المنتج على بيان يشير إلى المعالجة.

7	Water Extracted Fruit Juice defined under Section 2.1.3
Ref.	عصير الفاكهة المستخرج بواسطة الماء على النحو المحدد في القسم 3.1.2
DeepL	عصير الفاكهة المستخرج من الماء المحدد في القسم 2.1.3

At times, DeepL avoided the linguistic convention errors found in Google Translate. In Example 8, Google Translate mistakenly used the past-tense negation particle لـم (not), while DeepL appropriately used the present-tense negation particle لا (not). Another frequent error in Google Translate's output, which DeepL handled correctly, was skipping the definite article in mass nouns. In Arabic, collective or mass nouns should be definite. For instance, "poultry meat" should be translated as لحوم الدواجن (the poultry meat), not لحـوم دواجـن; similarly, "fish" should be الأسـماك, not أسـماك, and "starch" as النشـا, not نشـا.

8	In the absence of any such name, either a common or usual name existing by common usage as an appropriate descriptive term which was not misleading or confusing to the consumer shall be used.
DeepL	وفي حالة عدم وجود أي اسم من هذه الأسماء، يستخدم الاسم الشائع أو المعتاد الموجود حسب الاستخدام الشائع كمصطلح وصفي مناسب لا يكون مضللاً أو مربكاً للمستهلك.
Google Translate	في حالة عدم وجود أي اسم من هذا القبيل، يجب استخدام إما اسم شائع أو معتاد موجود من خلال الاستخدام الشائع كمصطلح وصفي مناسب لم يكن مضللا أو مربكا للمستهلك.

4.2.3. Style Errors

One stylistic error was the awkward use of the passive voice through the dummy verb "تـم" (or its derivatives) followed by a verbal noun. This construction is generally considered a poor style and is often rejected by translation reviewers and editors (Elsherif, 2023; Versteegh, 2014). It is preferable to use the passive voice verb directly. However, DeepL used it significantly less frequently than Google Translate, which translated most passive voice verbs in this manner. A comparison of the three models' outputs can be seen in Table 6. Another poor style was the use of prepositional phrases starting with بشـكل (in a manner). Such a style is unnecessarily wordy and would better be replaced by just the adverb; compare the reference translation with DeepL's output in Example 9.

Table 6 Passive voice translation across the three models.

Source Text	Google Translate	ChatGPT	DeepL
General standard for the labeling of prepackaged foods	المعيار العام لتوسيم المنتجات المعبأة	المعيار العام لتسمية الأطعمة المعبأة مسبقًا	المواصفة العامة لتوسيم الأغذية المعبأة مسبقاً
Adopted in 1985.	اعتمد عام 1985.	اعتمد في عام 1985.	اعتُمدت في 1985.
Amended in 1991, 1999, 2001, 2003, 2005, 2008 and 2010.	تم تعديله في أعوام 1991 و1999 و2001 و2003 و2005 و2008 و2010.	تم تعديله في السنوات 1991، 1999، 2001، 2003، 2005، 2008، و2010.	عُدلت في 1991 و1999 و2001 و2003 و2005 و2008 و2010.
Revised in 2018.	تمت مراجعته في 2018.	تم مراجعته في عام 2018.	نُقحت في 2018.

9	The following class titles may be used for food additives falling in the respective classes and appearing in lists of food additives permitted generally for use in foods:

.Ref	يجوز استخدام عناوين الفئات التالية للمواد المضافة إلى الأغذية التي تندرج ضمن هذه الفئات والواردة في قوائم المواد المضافة إلى الأغذية المسموح باستخدامها في المنتجات الغذائية عموما:
DeepL	يمكن استخدام عناوين الفئات التالية للمضافات الغذائية التي تندرج ضمن الفئات المعنية، وتظهر في قوائم المضافات الغذائية المسموح باستخدامها في الأغذية بشكل عام:

One last thing that exacerbated the awkwardness of DeepL's style was the repetitive use of the same word instead of employing synonyms or near-synonyms. In Example 10, the source sentence contained numerous near-synonyms or words belonging to the same semantic field, such as "tag, brand, mark, pictorial, or other descriptive matter," on the one hand, and "written, printed, stenciled, marked, embossed, or impressed on, or attached to," on the other hand. In the reference translation, each word was assigned a different Arabic term to encompass the full range of possible labeling methods for the first group of words and the full range of possible attachment techniques for the second group. However, DeepL used the exact translation for all words, resulting in an awkward translation that does not fully convey the breadth of information expressed in the source segment.

10	Label" means any tag, brand, mark, pictorial, or other" descriptive matter written, printed, stenciled, marked, .embossed, or impressed on or attached to a container of food

.Ref	ويقصد بمصطلح «بطاقة التوسيم» أي إشارة أو اسم أو علامة أو بيان صوري أو وصفي آخر مكتوب أو مطبوع أو مدموغ أو مختوم أو منقوش بنقوش نافرة أو موسوم على حاوية المنتج الغذائي أو مرفق بها.
DeepL	ويشمل الوسم أي علامة أو علامة أو صورة أو أي مادة وصفية أخرى، مكتوبة أو مطلية أو مرسومة أو مرسومة أو مرسومة أو معلمة أو منقوشة أو معجونة على الحاوية أو ملحقة بها.

Alongside awkward styles, DeepL displayed some inconsistencies. For example, the term "label" was sometimes translated as ملصـق "and at other times as وسـم" (a marking) or بطاقـة "(a card)." Also, in the list of foods in Table 7, DeepL sometimes used a resumptive pronoun to refer to the head noun of the noun phrase instead of repeating it; however, other times, it repeated the head noun; the reference translation consistently used resumptive pronouns, which is considered a more eloquent style in Arabic (Badawi et al. 2016).

Table 7 Examples of DeepL's inconsistencies in using resumptive pronouns.

Source Phrase	DeepL	Reference
Crustacea and products of these;	القشريات ومنتجاتها؛	القشريات ومشتقاتها؛
Eggs and egg products;	البيض ومنتجاته؛	البيض ومشتقاته؛
Fish and fish products;	الأسماك والمنتجات السمكية؛	السمك ومشتقاته؛

Peanuts, soybeans, and products of these;	الفول السوداني وفول الصويا ومنتجاتها؛	الفول السوداني وفول الصويا ومشتقاهتما؛
Milk and milk products (lactose included);	الحليب ومنتجات الألبان (بما في ذلك اللاكتوز)؛	الحليب ومشتقاته (بما في ذلك اللاكتوز)؛
Tree nuts and nut products;	المكسرات ومنتجات الجوز؛	الثمار الجوزية ومشتقاتها؛

4.2.4. Terminology Errors

Translation errors manifested in 80 incorrectly translated terms and two inconsistent terms across the regulations. Several reasons contributed to incorrect term translations. Firstly, some terms were translated so literally that the resulting translation became a non-existent term. Examples of these are listed in Table 8.

Table 8 Examples of literally translated terms.

Source Term	DeepL	Reference
crystalized fruit	فاكهة متبلورة	فاكهة محفوظة في السكر
single strength juice	العصير أحادي القوة	العصير غير المركز
citrus reticulata	الحمضيات الشبكية	فاكهة اليوسفي
Glazing agent	عامل التزجيج	عامل تلميع
directly expressed fruit juices	عصائر الفاكهة الطازجة	عصائر الفاكهة التي يتم التعبير عنها مباشرة

Another reason for incorrect term translations is Google Translate's inability to capture the context-specific meaning of the term. For example, diffusion generally translates into انتشـار (dispersion); however, in the context of food preparation, it is better translated as نقع (soaking), as shown in Example 11.

11	Water Extracted Fruit Juice is the product obtained by diffu- :sion with water of
.Ref	عصير الفاكهة المستخرج بواسطة الماء هو المنتج الذي يتم الحصول عليه من خلال الماء الذي ينقع فيه:
DeepL	عصير الفاكهة المستخلص بالماء هو المنتج الذي يتم الحصول عليه عن طريق الانتشار بالماء من:

One last reason for incorrect term translations is that some terms were left in English without translation or transliterated into Arabic letters. While it is common for professionals in health-related domains to use English terms or transliterations when writing, translation is to provide an understandable text in the target language. Even the reference translation attempted to provide Arabic translations for these terms; compare Google Translate's output to reference translations in Table 9.

Table 9: Examples of transliterated terms.

Source Term	DeepL	Reference Translation
monohydrate	أحادي الهيدرات	أحادي الإماهة
calco-carbonate equilibrium	توازن الكالكو-كربونات	موازنة كربونات الكالسيوم

In addition to the incorrect term translations, there was frequent inconsistency in translating the terms purée and nectar. Purée was transliterated as بوريــه, while at other times, it was rendered as مهـروس. Similarly, fruit nectar was frequently and incorrectly translated as رحيـق الفاكهـة, but sometimes it was correctly translated as نكتـار الفاكهة. Ideally, رحيق collocates with flowers and roses rather than fruits, and the transliterated word نكتــار is commonly used in the Arab world.

5. Discussion

The fact that ChatGPT performed the worst compared to Google Translate and DeepL contradicts recent findings by Kadaoui et al. (2023) and Moslem et al. (2023), who demonstrated that GPT and other large language models, such as Google Translate, Microsoft Translator, and Amazon Translate. One possible explanation for this discrepancy could be the translated text type. While Kadaoui and Moslem focused on general domain texts such as newspaper articles, this study evaluates specialized translation in food safety regulations. Specialized texts often involve complex terminology, specific linguistic conventions, and a need for domain expertise, which may present more significant challenges for ChatGPT in its untuned form. Additionally, the fact that DeepL outperformed even Google Translate suggests that it should now be used as a benchmarking baseline for machine translation work involving Arabic. Currently, most machine translation research for Arabic uses one or more of the three dominant commercial systems—Google Translate, Microsoft Translator, or Amazon Translate—as their baseline.

During the manual inspection of DeepL's output, it became evident that, although the translations differed from the reference texts, they

maintained a level of eloquence and readability comparable to human translations. This observation, supported by examples in Table 10, suggests that the gap between machine and human translations is narrowing. Recent research reinforces this, showing that humans are increasingly unable to distinguish between machine-generated and human-generated translations. For instance, Calvo-Ferrer (2023) found that viewers struggled to reliably differentiate between ChatGPT-generated and human-created subtitles in English-Spanish translations. However, lower-quality subtitles were more often associated with machine translation. Similarly, Al-Sabbagh (2024a) showed that participants had difficulty distinguishing between human-generated and machine-generated English-to-Arabic translations of patient information leaflets, further blurring the lines between human and AI-produced content. These findings carry significant implications for the translation industry, raising important questions about the future role of human translators. As translation researchers and practitioners, it may be time to reconsider traditional workflows and adapt to the rapidly evolving landscape of translation technology.

Table 10: Examples of DeepL's translations that, while different from the reference translation, are still acceptable.

Source Text	Reference	DeepL
The name and address of the manufacturer, packer, distributor, importer, exporter or vendor of the food shall be declared.	ينبغي ذكر اسم وعنوان مصنع المنتج الغذائي أو المعبئ أو الموزع أو المستورد أو المصدر أو البائع.	يجب التصريح عن اسم وعنوان الشركة المصنعة أو المعبئة أو الموزعة أو المستوردة أو المصدرة أو البائعة للأغذية.
The label of a food that has been treated with ionizing radiation shall carry a written statement indicating that treatment in close proximity to the name of the food.	ينبغي أن يرِدَ على بطاقة توسيم أي منتج غذائي قد تمت معالجته بالإشعاع المؤين بيان مكتوب يشير إلى تلك المعالجة لى مقربة من اسم المنتج الغذائي.	يجب أن يحمل ملصق الغذاء الذي تمت معالجته بالإشعاع المؤين بياناً مكتوباً يشير إلى تلك المعالجة على مقربة من اسم الغذاء.

If the language on the original label is not acceptable, a supplementary label containing the mandatory information in an acceptable language may be used instead of relabeling.	وإذا كانت اللغة الأصلية لبطاقة الوسم غير مقبولة، يجوز اللجوء إلى بطاقة وسم تكميلية عوض التوجه إلى إعادة التوسيم على أن تحتوي البطاقة التكميلية على جميع المعلومات المطلوبة بلغة مقبولة.	وإذا كانت اللغة المستخدمة في البطاقة الأصلية غير مقبولة، يجوز استخدام بطاقة تكميلية تحتوي على المعلومات الإلزامية بلغة مقبولة بدلاً من إعادة وضع البطاقة.
When a food undergoes processing in a second country that changes its nature, the country in which the processing is performed shall be considered to be the country of origin for the purposes of labelling.	وعندما يخضع المنتج الغذائي لعملية تجهيز تغير من طبيعته في بلد ثان، يعد البلد الذي تجرى فيه عملية التجهيز هو بلد المنشأ لغرض وضع بطاقات التوسيم.	وعندما يخضع الغذاء للتجهيز في بلد ثانٍ يغير من طبيعته، يعد البلد الذي يتم فيه التجهيز بلد المنشأ لأغراض وضع البطاقات التعريفية.

6. Limitations of the Study

The limitation of this study is the absence of information regarding the generation process of the reference translations. The Codex website offers no insight into the translation guidelines or resources employed. It remains unclear whether these translations were created entirely by human translators, underwent post-editing following initial machine translation, or were generated using an in-house machine translation system. This lack of transparency might hamper the ability to evaluate the reliability and quality of the reference translations. However, despite this limitation, the reference translations utilized in this study are the only official translations available on the Codex website.

7. Conclusion

This study demonstrates that DeepL, despite only recently adding Arabic to its supported languages, outperformed both Google Translate and ChatGPT in translating food safety regulations from English to Arabic across multiple evaluation metrics, including chrF++, TER, and COMET. While DeepL exhibited errors related to accuracy, linguistic conventions, style, and terminology, its output occasionally showed eloquence and readability comparable to human translations. These findings suggest that users who typically default to Google Translate due to its widespread availability should consider DeepL as a viable alternative, especially before making financial investments in machine translation services. The same applies to ChatGPT despite its popularity. The study recommends using DeepL as a benchmarking baseline for evaluating newly developed machine translation models. More research is needed to assess DeepL's performance in other specialized translation domains, such as legal and

technical translations. Lastly, rather than viewing high-quality machine translation outputs as a threat, translation professionals should explore integrating these systems into their workflows, utilizing post-editing to maintain quality while benefiting from the speed and cost-efficiency of machine translations.

References

Almahasees, Z. M. (2017). Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English. International Journal of Languages, Literature, and Linguistics, 30(1), 1–4. https://doi.org/10.18178/ijlll.2017.3.1.100

Almahasees, Z. M., 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

Al-Sabbagh, R. (2024a). Google Translate for medical texts: A quantitative-qualitative analysis of English into Arabic package inserts translations. Textual Turnings.

Al-Sabbagh, R. (2024b). PEACH: A sentence-aligned parallel English-Arabic corpus for healthcare. Corpora, 19(3).

Al Shamsi, H., Almutairi, A. G., Al Mashrafi, S., & Al Kalbani, T. (2020). Implications of language barriers for healthcare: A systematic review. Oman Medical Journal, 35(2), e122. https://doi.org/10.5001/omj.2020.40

Arab Times Kuwait. (2024, September 9). ChatGPT hits 200 million weekly users worldwide. https://www.arabtimesonline.com/news/chatgpt-hits-200-million-weekly-users-worldwide/

Badawi, E.S., Carter, M., & Gully, A. (2016). Modern written Arabic: A comprehensive grammar (2nd ed.). Routledge.

Brewster, R. C.L., Gonzalez, P., Khazanchi, R., Butler, A., Selcer, R., Chu, D., Aires, B. P., Luercio, M., & Hron, J. D. (2024). Performance of ChatGPT and Google Translate for pediatric discharge instruction translation. Pediatrics, 154(1): e2023065573. https://doi.org/10.1542/peds.2023-065573

Calvo-Ferrer, J. R. (2023). Can you tell the difference? A study of human vs machine-translated subtitles. Perspectives, 1–18. https://doi.org/10.1080/0907676X.2023.2268149

Cornelison, B. R., Al-Mohaish, S., Sun, Y., & Edwards, C. J. (2021). Accuracy of Google Translate in translating the directions and counseling points for top-selling drugs from English to Arabic, Chinese, and Spanish. American Journal of Health-System Pharmacy, 78(22): 2053–2058. https://doi.org/10.1093/ajhp/zxab224

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 caregivers. Clinical Pediatrics, 58(2): 247–249. https://doi.org/10.1177/0009922818809494

DeepL. (2024, January 24). DeepL Translator welcomes Arabic. https://www.deepl.com/en/blog/deepl-welcomes-arabic

Delfani, J., Orăsan, C., Saadany, H., Temizöz, Ö., Taylor-Stilgoe, E., Kanojia, D., Braun, S., & Schouten, B. (2024). Google Translate error analysis for mental healthcare information: Evaluating accuracy, comprehensibility, and implications for Multilingual Healthcare Communication. ArXiv Preprints. https://arxiv.org/ftp/arxiv/papers/2402/2402.04023.pdf

ElSherif, M. H. (2023). English copula translation techniques into Arabic at the United Nations: A contrastive analysis study. مجلة وادي النيل للدراسات والبحوث الإنسانية والإجتماعية والتربوية, 40 (40). https://doi.org/10.21608/jwadi.2023.320755

Esperança-Rodier, E. & Frankowski, D. (2021). DeepL vs Google Translate: Who's the best at translating MWEs from French into Polish? A multidisciplinary approach to corpora creation and quality translation of MWEs. Translating and the Computer, 43. https://hal.science/hal-03779450v1

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q. & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9, 1460–1474. https://doi.org/10.1162/tacl_a_00437

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

Ehab, R., Amer, E., Gadallah, M. (2018). Example-Based English to Arabic machine translation. In Proceedings of the 7th International Conference on Software and Information Engineering, 131–135. https://doi.org/10.1145/3220267.3220294

Kadaoui, K., Magdy, S. M., Waheed, A., Md Tawkat Islam Khondaker, M. T. I., El-Shangiti, A. O., Nagoudi, E., B. I., & Abdul-Mageed, M. (2023). TARJAMAT: Evaluation of BARD and ChatGPT on machine translation of ten Arabic varieties. ArXiv (Cornell University). 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

Kubińska, O. & Kubiński, W. (2020). Feasibility of DeepL, Google, and Microsoft MT systems: Implementation into the translation process in the ENG-> PL language pair. Gdańsk University Press. https://wydawnictwo.ug.edu.pl/wp-content/uploads/2021/09/Kur-Feasibility-of-DeepL-fragm.pdf

Lee, J.-G., Lee, Y., Kim, C.-S., & Han, S. B. (2021). Codex Alimentarius Commission on ensuring food safety and promoting fair trade: Harmonization of standards between Korea and codex. Food Science and Biotechnology, 30(9), 1151–1170. https://doi.org/10.1007/s10068-021-00943-7

Li, X. (2024). Comparison of translation quality between large language models and neural machine translation systems: A case study of Chinese-English language pair. International Journal of Education and Humanities (IJEH), 4(2),121-128.

Lommel, A. (2018). Metrics for Translation Quality Assessment: A case for Standardizing error typologies. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), Translation Quality Assessment. Machine Translation: Technologies and Applications (pp. 119–127). Springer. https://doi.org/10.1007/978-3-319-91241-7_6

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. Revista Tradumàtica, 12, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Mariana, V. R. (2014). The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment. [Master's thesis, BYU University]. http://hdl.lib.byu.edu/1877/etd7404

Maruf, S., Saleh, F., & Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. ACM Computing Surveys, 54(2), 45. https://doi.org/10.1145/3441691

Maučec, M. S. & Donaj, G. (2020). Machine translation and the evaluation of its quality. In A. Sadollah & T. S. Sinha (Eds.), Recent trends in computational intelligence. IntechOpen eBooks. https://doi.org/10.5772/intechopen.89063

Miller, J. M., Harvey, E. M., Bedrick, S., Mohan, P. & Calhoun, E. (2018). Simple patient care instructions translate best: Safety guidelines for physician use of Google Translate. Journal of Science Communication, 25(1), 18–27.

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive machine translation with large language models. ACL Anthology, June 1. https://aclanthology.org/2023.eamt-1.22

Nagoudi, E. M. B., Elmadany, A. & Abdul-Mageed, M. (2021). Investigating Code-Mixed Modern Standard Arabic-Egyptian to English Machine Translation. Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching. https://doi.org/10.18653/v1/2021.calcs-1.8

Noll, R., Berger, A., Kieu, D., Mueller, T., Bohmann, F., Müller, A., Holtz, S., Stoffers, P., Hoehl, S., Guengoeze, O., Eckardt, J.-N., Storf, H., & Schaaf, J. (2024). Assessing GPT and DeepL for terminology translation in the medical Domain: A comparative study on the human phenotype ontology [Preprint]. Research Square. https://www.researchsquare.com/article/rs-4836251/v1

Papineni, K., Roukos, S. Ward, T. & Zhu, W.-J. (2002). Bleu: a method for automatic

evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). https://doi.org/10.3115/1073083.1073135

Peña Aguilar, A. (2023). Challenging machine translation engines: Some Spanish-English linguistic problems put to the test. Cadernos De Tradução, 43(1),1–26. https://doi.org/10.5007/2175-7968.2023.e85397

Piazzolla, S. A., Savoldi, B., & Bentivogli, L. (2023). Good, but not always fair: An evaluation of gender bias for three commercial Machine Translation systems. Hermes – Journal of Language and Communication in Business, 63, 209–225. https://doi.org/10.7146/hjlcb.vi63.137553

Pitman, J. (2021, April 18). Google Translate: One billion installs, one billion stories. Google Blog. https://blog.google/products/translate/one-billion-installs/

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. 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

Pym, A., Ayvazyan, N., & Prioleau, J. M. (2022). Should raw machine translation be used for public health information? Suggestions for a multilingual communication policy in Catalonia. Just. Journal of Language Rights and Minorities, 1(1-2), 71–99. https://doi.org/10.7203/just.1.24880

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, 55(11): 229. https://doi.org/10.1145/3567592

Rao, P., McGee, L. M., & Seideman, C A. (2024). A comparative assessment of ChatGPT vs. Google Translate for the translation of patient instructions. Journal of Medical Artificial Intelligence, 7, 11. https://dx.doi.org/10.21037/jmai-24-24

Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 2685–2702). https://doi.org/10.18653/v1/2020.emnlp-main.213

Rescigno, A. A. & Monti, J. (2024). Gender bias in machine translation: A statistical evaluation of Google Translate and DeepL for English, Italian and German. Proceedings of the International Conference HiT-IT 2023. https://doi.org/10.26615/issn.2683-0078.2023_001

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. Language Resources and Evaluation, 56, 593–619. https://doi.org/10.1007/s10579-021-09537-5

Salinas, M.-J. V. & Burbataa, R. (2023). Google Translate and DeepL: Breaking taboos in translator training. Observational study and analysis. Ibérica, 45, 243–266. https://doi.org/10.17398/2340-2784.45.243

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

Taira, B. R., Kreger, V., Orue, A., & Diamond, L. C. (2021). A pragmatic assessment of Google Translate for emergency department instructions. Journal of General Internal Medicine, 36(11), 3361–3365. https://doi.org/10.1007/s11606-021-06666-z

Versteegh, K. (2014). Arabic language. Edinburgh University Press.

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

جودة الترجمة المتخصصة

دراسة مقارنة بين Google Translate و DeepLو ChatGPT في ترجمة لوائح سلامة الأغذية من الإنجليزية إلى العربية

ريم صفوان[1]
رانيا الصباغ[2]

ملخص البحث:

تقارن هذه الدراسة بين نموذجين من ثلاثة نماذج للترجمة الآلية هي جوجل للترجمة وDeepL وChatGPT وذلك في سياق ترجمة لوائح سلامة الأغذية الصادرة عن لجنة الدستور الغذائي من الإنجليزية إلى العربية. وتستهدف الدراسة ترجمة لوائح سلامة الأغذية الصادرة عن لجنة الدستور الغذائي نظرًا لما تحتويه من تحديات قد يصعب على أنظمة الترجمة الآلية معالجتها، مثل المصطلحات التقنية المتعلقة بالمنتجات الغذائية وتصنيعها. في الجزء الأول من الدراسة، تعتمد المقارنة على مؤشرات أداء كمية مثل BLEU وchrF++ وTER وCOMET لتحديد أفضل النماذج أداءً. ثم في الجزء الثاني، تعتمد الدراسة على التحليل اليدوي لتصنيف نوعية الأخطاء في ترجمة أفضل النماذج طبقا لمؤشرات الأداء. وتشير النتائج إلى أن نموذج جوجل للترجمة هو الأفضل بناءً على مؤشر BLEU، بينما تفوق نموذج DeepL في جميع مؤشرات الأداء الأخرى وتخلف نموذج ChatGPT عن كليهما. ومن خلال التحليل اليدوي لترجمة DeepL، اتضح وجود أخطاء أسلوبية ونحوية ودلالية، خاصةً في المصطلحات التقنية، ومع ذلك تميزت ترجمته أحيانًا بالفصاحة وسهولة القراءة تميزًا مشابهًا للترجمات البشرية. وبهذا يمكن اعتبار DeepL بديلا قويا لنموذج جوجل للترجمة، كما يمكن استخدامه معيارا جديدا لتقييم أبحاث الترجمة الآلية للغة العربية. وأخيرًا، يجب على المترجمين المحترفين النظر في دمج أنظمة الترجمة الآلية في عملياتهم لتحسين الكفاءة ليس فقط في المجالات العامة، ولكن أيضًا في المجالات المتخصصة.

الكلمات الدالة: الترجمة الآلية، الترجمة المتخصصة، الترجمة من الإنجليزية إلى العربية، لوائح سلامة الأغذية، جوجل للترجمة، ChatGPT، DeepL

(1) كلية الآداب - العلوم الإنسانية والاجتماعية (الشارقة - جامعة الشارقة)
u20104360@sharjah.ac.ae

(2) كلية الآداب - العلوم الإنسانية والاجتماعية (الشارقة - جامعة الشارقة)