Journal Of the Iraqia University (76-4) 31 July (2026)

ISSN(Print): 1813-4521 OnLine ISSN:2663-7502

Journal Of the Iraqia University

available online at

https://iasj.rdd.edu.iq/journals/journal/view/247

Machine Translation of Gynecological Texts from English into Arabic: A Comparative Evaluation of Google Translate and DeepL

Osamah Mohammed Abdulla

Affiliation: University of Tikrit, College of Arts, Department of Translation

Email: Ossama.m.a23@tu.edu.iq

Abstract

This research evaluates the efficacy of Google Translate and DeepL in the translation of gynecological texts from English to Arabic. It concentrates on two variables: the precision of technical-term translation and the conveyance of meaning at the entire-text level. The corpus comprises 25 gynecological texts curated from two specialized volumes in obstetrics and gynecology. A comparative analytical methodology was employed to evaluate the outputs generated by both machine translation systems in relation to the English source texts and a human reference translation. The analysis looked at how many technical terms were in each text, how many were accurate and how many were wrong, and how well each translated text did overall. The results show that Google Translate got 177 technical terms right out of 194, or 91.24%. DeepL got 172 technical terms right out of 194, or 88.66%. At the whole-text level, Google Translate passed 19 texts and failed 6, with a pass rate of 76%, whereas DeepL passed 20 texts and failed 5, with a pass rate of 80%.The results indicate only a limited difference between the two systems. Google Translate shows a marginal advantage in technical-term accuracy, whereas DeepL shows a slight advantage in whole-text pass rate. The study concludes that machine translation can provide useful initial support in translating gynaecological discourse, but its output still requires expert human revision in specialized medical contexts.

Keywords: machine translation, Google Translate, DeepL, gynaecological texts, medical translation, technical terms, English-Arabic translation

1. Research Framework

1.1 Problem of the Study

Machine translation has seen tremendous progress in recent years, but the effectiveness of machine translation in specialized domains such as medicine is still under debate, especially when translating clinical content for patients. This study evaluates the quality of translation of English gynaecological content to Arabic. The translation quality to Arabic of two well-known translation tools, Google Translate and DeepL, is assessed at two levels: 1) the translation quality of gynaecological terms and 2) the translation quality in terms of meaning of clinical content.

1.2 Aim of the Study

This study aims to evaluate the performance of Google Translate and DeepL in translating specialized English gynaecological texts into Arabic. It tests whether both systems are able to accurately translate technical English gynaecological and medical vocabulary and whether the translated output conveys a meaning equivalent to that of the source text. The study compares both systems' approaches to rendering specialized gynaecological discourse in the target language. The findings do not attempt to declare a superior system, but rather highlight the similarities and differences in tackling such complex language data.

1.3 Objectives of the Study

The present study seeks to achieve the following objectives:

1. To examine how accurately Google Translate and DeepL translate gynaecological texts from English into Arabic
2. To identify the extent to which each system succeeds in rendering technical gynaecological and medical terms accurately
3. To determine whether the translated texts convey the intended meaning of the source text as a whole
4. To compare the performance of Google Translate and DeepL in this specialized field and to assess whether any observed difference remains limited to the selected corpus

5. To calculate the percentages of technical-term accuracy achieved by each system, as well as the rate of Pass and Fail judgments across the selected texts

1.4 Research Questions

The study attempts to answer the following questions:

1. How accurate are Google Translate and DeepL in translating gynaecological texts from English into Arabic
2. To what extent does each system succeed in translating technical gynaecological and medical terms accurately
3. Do the translated texts convey the intended meaning of the source text as a whole
4. To what extent do Google Translate and DeepL differ in translating this type of specialized medical text

1.5 Hypotheses of the Study

The study is based on the following hypotheses:

1. There is a difference in the accuracy of translations from English to Arabic in gynaecology using Google Translate and the DeepL translator.
2. Translation software such as Google Translate or DeepL has different capabilities to translate technical terms of gynaecology and medical issues correctly.
3. An assessment of the translated texts produced by the two systems reveals varying levels of success in capturing the message contained in the original text.
4. Differences in the translation of specialized gynaecological texts are expected to be limited and largely corpus-bound.

1.6 Limits of the Study

The study does not cover all translations or language pairs. Instead it only deals with comparing the quality of translations by Google Translate on the one hand and by DeepL on the other, for translations from English into Arabic. The study's data base consists of 25 texts from two English books dealing with obstetrics and gynaecology. Two variables have been studied in the study, the first of them is concerned with the correct translation of technical terms in the gynaecology field and the second variable is concerned with the accurate translation of the whole text. Based on the scope of the study, the results only refer to the selected corpus of the study and cannot be generalised to other languages or other medical texts.

1.7 Significance of the Study

This study deals with the evaluation of machine translation (MT) in a highly specialized field of the medical discourse, i.e. in the field of gynaecology. It examines and compares the two prominent online machine translation tools, Google Translate and DeepL, assessing their ability to accurately translate English technical gynaecological terms into Arabic and the overall translation quality achieved by the two MT systems under study. This study will be useful to scholars of translation studies, medical translation and machine translation as well as to translators and translation students at different stages during their studies. They will be able to see the constraints and potential of MT in dealing with specialized aspects of human languages.

2. Methodology

2.1 Data of the Study

The data of the study consist of 25 selected gynaecological texts translated from English into Arabic by Google Translate and DeepL. The texts were taken from two specialized books in obstetrics and gynaecology and were chosen purposively to represent different types of gynaecological discourse. A human reference translation was also prepared for each text in order to support the comparative analysis.

2.2 Data Selection Criteria

The 25 selected texts were chosen according to the criteria which are connected to the research questions. First, the selected texts deal with gynaecology and contain specialised information. Second, the selected texts include enough technical vocabulary to be able to analyse it in the context of the research aims. Third, the length of the selected texts had to be manageable enough to make a detailed comparison between corresponding parts of both systems possible. The fourth criterion requires that different topics of the gynaecological discourse should be represented in the selected texts. All words identified as being technical have been tagged and checked against the reference translation. The 25 chosen texts vary in length, and have been chosen to be as evenly technical as possible, although as the project has a specific subject, the texts are not comparable in all respects.

2.3 Procedures of Analysis

To carry out the analysis, a clear sequence of steps was followed. First, 25 gynaecological texts were selected from the two books adopted in the study. Second, each text was translated from English into Arabic by means of Google Translate and DeepL. Third, a human reference translation was prepared and used as a benchmark for

comparison. These reference translations were prepared by the authors of the study, who specialize in translation, under the supervision of a specialist gynecologist from the University of Tikrit in order to ensure medical accuracy and terminological appropriateness.After that, the technical gynaecological and medical terms in each source text were identified and counted. The renderings of these terms in the two machine-translated versions were then examined and classified as either accurate or problematic, depending on whether they preserved the intended medical meaning in context. Finally, each translated text was evaluated as a whole and was given a Pass or Fail judgment according to its success in conveying the intended medical meaning of the source text. A translation was marked Fail when one or more errors distorted a clinically relevant meaning, affected diagnosis or management, or produced medically unsafe wording.

2.4 Statistical Treatment of Data

The data was analyzed quantitatively by calculating total technical terms per test set, correct and incorrect technical terms per system, and accuracy of technical terms per system. The number of passed or failed test texts per system was also counted and used as separate data to compare whole-text processing in addition to term-level processing.

3. Theoretical Background

3.1 Specialized Translation

Specialized translation is used to refer to the translation of texts originating in professional, scientific/technical, legal or institutional language use. Such texts contain specialised language, which requires more than simple bilingualism and involves reference to subject matter knowledge, generic variation and disciplinary variation as well as the discourse goals and strategies of the target language and communication event. This volume presents current research on specialised translation (Munday et al., 2022; Prieto Ramos et al., 2024).In specialized translation, every word forms part of the technical vocabulary of a particular discipline(s) and hence functions as technical units within the communicative purpose of the translation. Specialized translation is therefore necessarily a high quality activity which requires consistent and accurate use of terminology and an appropriate use of context (Colina, 2025; Prieto Ramos et al., 2024).This framework is relevant to our study because gynaecological texts are found in a specialized medical domain where terms, content and functionality are crucial. The framework of specialised translation is therefore relevant when approaching the translation of such texts. Viewed from a broader perspective, the sub-domain of medical translation is particularly relevant to our study.

3.2 Translation of Medical Terminology

Medical terminology constitutes one of the principal elements of medical discourse and, as such, forms a pivotal area of focus in medical translation. Specialist terms in the medical field not only convey conceptual meaning but also clinical meaning, which is often of significant importance to the treatments and diagnoses of a variety of medical conditions. Translation of medical terminology, therefore, goes beyond the quest for simple equivalent words in target language; it also entails consideration of meaning, purpose and audience needs of the source language and further communicative aspects (Montalt-Resurreccio et al., 2025).Machine translation in healthcare brings many benefits to doctors, patients and health professionals. In this paper we report on an experiment in medical MT and show that testing and evaluating of such a challenging application is fundamental. Unlike other MT applications, simply checking how well the output is translated is not sufficient. The Fluency Evaluation Criterion cannot be fully sufficient for measuring the quality of translations, especially when clinical terms are involved. In fact, the problem of terminological accuracy is crucial and thus the accuracy of translated clinical terms such as being too vague, too inconsistent, too imprecise or even wrong, needs to be assessed (Zappatore and Ruggieri, 2024)In their study on the performance of GPT and also DeepL when translating medical terminology, Noll et al. (2025) choose the vast Human Phenotype Ontology as their specific domain framework. As argued above, medical terminology can and should be studied as a separate category of language and not only as one of the aspects of language that affect the quality of the translation of also other types of text, such as the gynaecological texts dealt with in this volume. Therefore, terminological accuracy deserves special attention when machine-translated gynaecological texts are evaluated.Although the issue is relevant to any language pair, it has special importance in English-Arabic. While English-Arabic translation and its assessment have been the subject of some research, most of it has been biased towards the evaluation of machine translation and postediting. More domain-sensitive research is required as well as a focus on translating real applications of translational competence. The paper explores this issue with reference to medical translation, highlighting the acceptability of a term in isolation as opposed to its inappropriateness due to vagueness, insufficiency of use, or

unprofessional usage in translation practice (Omar and Salih, 2024)This study is significant for the current research on translation because the texts in the field of gynaecology include terms of anatomy and organs as well as terms relating to disease, diagnosis, treatment and risk. From the point of view of the evaluation criteria, priority will be given to the accuracy of terminology use. A critical evaluation will be made regarding the quality of the translated output provided by both DeepL and Google Translate(Zappatore & Ruggieri, 2024). In the context of English to Arabic MT postediting research, it is significant to note that such an assessment must be domain-specific and requires a corresponding translation competence that is also reflected in translation practice (Omar & Salih, 2024).

3.3 English-Arabic Challenges in Medical Translation

English-Arabic medical translation is complicated by two factors: the specialized nature of the language used in medical texts, and the complexities of the Arabic language. Research on English-Arabic machine translation post-editing has exposed significant gaps between current research in training, evaluation, and translation competence for this language pair and what is actually needed to attain translation competence for this language pair, which is clearly not a simple transfer problem (Omar & Salih, 2024). Thus, translation of medical texts requires not only in-depth knowledge of the target language, but also in-depth knowledge of the source language in order to achieve the correct medical meaning, and to produce a precise, readable, and adequate translation for medical professionals.In addition to the above difficulties, the language of the medical texts has its own problems as highlighted in several recent reviews on Arabic Natural Language Processing (NLP). These include complex morphology, orthography and diacritics, and ambiguous and variant words. Modern Standard Arabic (MSA) and the dialects, for example, form two different language systems in the Arab world and thus require separate resources. Unfortunately, there is a limited number of available Arabic NLP resources. Another challenge faced by Arabic NLP systems is the language's vast derivational and inflectional variations, and significant syntactical differences between Arabic and English. As a result, English medical terms and expressions are often translated into Arabic in an inappropriate, ambiguous, or inconsistent manner by current computer systems.Medical translation also has a lot of responsibility when it comes to communication. Recent studies indicate that specialized healthcare texts continue to pose challenges for machine translation, and even proficient neural output may still present safety concerns in medical settings. This problem is especially important when translating from English to Arabic, where language complexity and domain sensitivity come into play. In these situations, a translation may seem natural but still make a diagnosis less accurate, misinterpret a medical term, or use formal Arabic that isn't appropriate.This paper argues that English-Arabic medical translation is a separate problem and deals with an example of that problem through translating gynaecological texts. These texts contain very specialized vocabulary and content; hence, their translation into Arabic may encounter some challenges posed by Arabic morphology and syntax besides lexical ambiguity and lack of sufficient resources and materials in the medical field. This paper tests the output of two transformer based neural machine learning models, namely, DeepL and Google Translate, on English gynaecological texts and evaluates their performance in English-Arabic translation.

3.4 Machine Translation and Neural Machine Translation

The process of machine translation (MT) refers to the use of software to translate text automatically. Recent MT research and applications have largely adopted a neural machine translation (NMT) approach, which has reached human proficiency and produces more natural human communication than earlier MT versions. However, critical issues remain in the medical application of such systems, particularly with respect to the correct use of terms and their semantic meaning, especially during critical scenarios (Rivera-Trigueros, 2022; Zappatore & Ruggieri, 2024).To tackle this task, we first need to know what machine translation and neural machine translation are. After that we have a closer look on Google Translate and – especially impressive – on DeepL. Both are able to generate really fluent neural output for medical gynaecological texts. But, just fluency is not enough for good results, thus an evaluation by humans is also necessary.

3.5 Machine Translation in Medical Discourse

Machine translation is an important technology in medicine, as more and more medical communication is done in more than one language for clinical, educational, and public health purposes. In addition to these routine uses of machine translation, there is a growing interest in its use in health care settings where speed and accessibility of translation are paramount and instant human translation is not possible. (Herrera-Espejel & Rach, 2023; Merx et al., 2024).MT in medical discourse is still a challenging area of research that requires more investigation into the potential shortcomings of the current MT systems. Even though MT can process massive amounts of text in

the medical field, it is difficult for the MT to generate output in specialized discourse, because MT requires not only accurate use of biomedical terms but also semantic accuracy and sophisticated choices of words and phrases which take context into consideration. Therefore, research in healthcare translation is shifting its focus from human evaluation of fluent surface forms to more in-depth evaluations (Merx et al., 2024; Zappatore & Ruggieri, 2024).

3.6 Translation Quality Assessment

Quality evaluation of translation output has become an important area of research within translation studies. Within the MT research community there is a growing number of studies which indicate that the quality concept cannot be represented by a general, unique impression of MT output. Instead, MT performance is deeply influenced by a variety of independent variables such as the type of text, the domain, language pair, or the methodology employed for quality evaluation. Therefore, the notion of quality is no longer conceived as confined to fluency. This paper focuses on the quality evaluation of three web-based systems that provide automatic translation from English to Portuguese. It explores both intrinsic and extrinsic quality aspects of the systems and the results clearly indicate that significant differences are found in the quality obtained for the three systems under analysis. The quality-intrinsic dimension did not necessarily predict the quality scores achieved in the overall evaluation (Rivera-Trigueros, 2022; Zaretskaya et al., 2024).Unlike in other domains, simply stating that medical texts are “fluently” translated by several MT systems is not sufficient. Although such output may easily be mistaken for correct and fluent text by an untrained eye, a closer analysis uncovers additional errors such as insufficient term accuracy and semantic distortions as well as inadequate wording in context. Therefore, the output of machine translation in medical language requires a two-level evaluation: term accuracy for individual translated segments, and message adequacy for the entire translated document. (Zappatore & Ruggieri, 2024; Merx et al., 2024).Instead of covering the entire spectrum of assessment criteria for quality, two variables were chosen as the focal points for this assessment. These two variables focus on the accuracy of translations of technical terms and names of medical and gynaecological conditions (general and specific) and on the overall quality of the translation assessed in terms of the intended medical content and its correct communication to the readers of the target language.

3.7 Functional Adequacy in Medical Translation

Medical translation is often equated with language translation; yet, as is the case with many areas of human endeavour, success is measured not simply by the quality of performance but by the extent to which objectives are met. Indeed, an outstanding performance is one that is suited to purpose. Medical translation is no exception to this rule. While a good translation will provide the reader with an accurate rendering of the original text in terms of language and content, in Arabic-speaking cultures, such a translation may well increase ambiguity, diminish the importance of critical information or simply fail to convey an appropriate message (Montalt-Resurreccio et al., 2025; Zappatore & Ruggieri, 2024).The study builds on established approaches to assessing functional adequacy for the purpose of ensuring adequate quality of translated health communication materials. In addition to assessing the accuracy of translation of individual technical vocabulary, the study assesses whether the translated text as a whole conveys the intended meaning in the source language text, in this case, gynaecological content. Functional adequacy is used in a limited, pragmatic sense to measure for extent to which the output, in this case the translated Arabic text, clearly, accurately and sufficiently conveys the intended message for the user.

3.8 Error Analysis in Machine Translation

Error analysis is an important aspect of evaluating machine translation systems. While an initial impression of fluency of surface quality can provide valuable first level feedback on the output of a translation system, it does not measure well meaning and terminology related weaknesses. This paper reports on a detailed error analysis which identified a set of error types and quantified their impact on the overall quality of the translated documents. The analysis detected meaning-related and terminology-related problems detected by human post-evaluation against the original content. The results of the study provide new insights in the types and frequencies of errors that current MT systems make. From a MT research perspective, the study shows that error analysis is an important tool to evaluate automatic translation systems (Rivera-Trigueros, 2022; Omar & Salih, 2024).The study is based on error analysis in a practical and limited way. It aims at identifying areas of problematic representation, pinpointing errors and categorizing them in order to explain how many errors there are and what kinds of problems these errors represent in terms of the quality of language produced by Google Translate and

DeepL. An attempt is made to measure the extent to which these errors reduce the adequacy of the generated Arabic medical translation.

4. Practical Section

4.1 Analytical Procedure

The practical analysis compares the machine-translated outputs produced by Google Translate and DeepL for selected gynaecological texts translated from English into Arabic. We compare each translation to the English source text and a human reference translation. The machine-translated texts examined in this study were produced via the public websites of Google Translate and DeepL in 2025. There are two variables that the analysis is based on. The first variable is how well technical medical and gynecological terms are translated. For each text, the technical terms are found and counted, and the translations in both systems are rated as either correct or wrong. When a rendering keeps the intended medical meaning in context, it is considered accurate. It is considered problematic when it includes semantic distortion, medically inappropriate language, imprecise terminology, or a less specialized equivalent that dilutes the technical accuracy of the original term. Terms that are problematic are listed and briefly explained. But just because there is one or more problematic terms doesn't mean that the person will automatically fail unless the problem changes the meaning of the whole text in a way that is important for diagnosis, management, or medical safety. The second variable is whether or not the translation was able to convey the meaning of the source text as a whole. A translation is marked as "Pass" when it accurately conveys the intended medical meaning of the source text without distorting it in a way that is clinically relevant at the whole-text level. When one or more mistakes change a clinically relevant meaning or affect diagnosis, management, or medical safety, it is marked as Fail. The analysis keeps track of the total number of technical terms in each text, the correct and incorrect translations in each system, the terms that are causing problems (if any), and the overall Pass/Fail rating. This is how Google Translate and DeepL are compared in terms of how well they translate technical terms and whole texts.

4.2 Data Presentation

*Due to the length of the texts included in 4.2 Data Presentation, the full materials have been uploaded to Google Drive and can be accessed through the link and QR code provided below, so that reviewers may consult the complete dataset directly.

https://docs.google.com/document/d/1tAjXwPBXG7svLEYFW0I2DiEj54Tc9zW5/edit

4.3 Data Analysis and Discussion

This subsection presents a text-by-text comparison between Google Translate and DeepL in terms of technical-term accuracy and overall Pass/Fail judgment at the whole-text level. For each text, the analysis identifies the total number of technical terms, the accurate and problematic renderings in each system, the type and justification of problematic terms when present, and the final judgment on whether the translated text succeeds in conveying the intended meaning of the source text as a whole.

Table 1. Summary of Results by Text

Text	GT Total Terms	GT Accurate	GT Problematic	GT Accuracy	GT Overall Judgment	DeepL Total Terms	DeepL Accurate	DeepL Problematic	DeepL Accuracy	DeepL Overall Judgment
1	8	8	0	100%	Pass	8	8	0	100%	Pass
2	9	7	2	77.78%	Pass	9	9	0	100%	Pass
3	9	8	1	88.89%	Fail	9	8	1	88.89%	Pass
4	8	8	0	100%	Pass	8	8	0	100%	Pass
5	10	8	2	80%	Pass	10	8	2	80%	Fail
6	10	10	0	100%	Pass	10	7	3	70%	Pass
7	7	5	2	71.43%	Fail	7	6	1	85.71%	Fail
8	8	7	1	87.50%	Pass	8	6	2	75%	Pass
9	10	9	1	90%	Pass	10	8	2	80%	Pass
10	8	6	2	75%	Pass	8	6	2	75%	Pass
11	9	9	0	100%	Pass	9	8	1	88.89%	Pass
12	9	8	1	88.89%	Fail	9	8	1	88.89%	Fail
13	8	8	0	100%	Pass	8	7	1	87.50%	Pass
14	8	8	0	100%	Pass	8	7	1	87.50%	Fail
15	6	6	0	100%	Pass	6	6	0	100%	Pass
16	7	7	0	100%	Pass	7	6	1	85.71%	Pass
17	7	7	0	100%	Pass	7	7	0	100%	Pass
18	6	6	0	100%	Pass	6	6	0	100%	Pass
19	7	6	1	85.71%	Fail	7	6	1	85.71%	Fail
20	6	5	1	83.33%	Fail	6	6	0	100%	Pass
21	7	6	1	85.71%	Fail	7	7	0	100%	Pass
22	5	5	0	100%	Pass	5	5	0	100%	Pass
23	8	8	0	100%	Pass	8	7	1	87.50%	Pass
24	7	5	2	71.43%	Pass	7	6	1	85.71%	Pass
25	7	7	0	100%	Pass	7	6	1	85.71%	Pass

Text 1: Overall, both Google Translate and DeepL were judged Pass because they preserved the intended medical meaning of the source text without any problematic technical rendering that distorted the whole-text message.

Text 2: Although there were 2 terms that Google Translate got wrong, the score still reached 'Pass' as the general meaning for medical use was clearly conveyed. DeepL scored 'Pass' and did a better job of translating individual terms and the whole text.

Text 3: Overall, Google Translate was judged Fail because its rendering of *prognosis* as **التشخيص**distorts a clinically important meaning, whereas DeepL was judged Pass because its weaker rendering of *oncological pathways* narrows the meaning but does not invalidate the whole-text medical message.

Text 4: Overall, both Google Translate and DeepL were judged Pass because they conveyed the diagnostic and clinical meaning of the source text clearly and without distortion at the whole-text level.

Text 5: Overall, Google Translate was judged Pass because, despite two terminological weaknesses, it preserved the main medical meaning of the passage, whereas DeepL was judged Fail because its rendering of the opening clause distorts the central meaning of the source text by suggesting restriction to physiological changes rather than limitation by them.

Text 6: Overall, Google Translate was judged Pass because it preserved both the technical details and the overall management meaning of the text, while DeepL was also judged Pass because its problematic renderings did not ultimately prevent the source message from being understood as a whole.
Text 7: Both Google Translate and DeepL were given a Fail overall because their translations made important cancer-related terms less clear, which made the text less useful as a whole.
Text 8: Both Google Translate and DeepL were rated as "Pass" overall because they were able to successfully transfer the important diagnostic description and clinical message, even though one or two terms were not translated as accurately as in the reference translation.
Text 9: Google Translate was given a Pass overall because it kept the emergency meaning of the source text, even though one translation was wrong. DeepL was also given a Pass because its terminology issues did not change the overall life-threatening clinical message.
Text 10: Both Google Translate and DeepL were given a "Pass" because, even though each translation had some minor mistakes in terminology, both versions were able to clearly convey the follow-up instructions and the overall clinical meaning of the source text.
Text 11: Google Translate was given a Pass overall because it clearly and accurately translated the postoperative emergency scenario. DeepL was also given a Pass because its translation was still clinically understandable even though it used a less precise word for pulmonary embolism.
Text 12: Both Google Translate and DeepL were rated as Fail because each translation has a clinically important error. Google Translate makes the diagnostic category of gestational trophoblastic disease less clear, and DeepL mistranslates surgical evacuation as إجهاض جراحي, which changes the meaning of the intended management.
Text 13: The overall verdict on Google Translate was Pass because it accurately conveyed the infertility and endometriosis findings. DeepL also got a Pass because its weaker phrasing did not change the main fertility-related meaning of the text.
Text 14: Overall, Google Translate got a Pass because it clearly kept the diagnosis and the main surgical options. DeepL got a Fail because it mistranslates colposuspension as رفع عنق الرحم, which changes the procedure and affects management meaning.
Text 15: In general, both Google Translate and DeepL got a Pass because they accurately translated the obstetric emergency, the recognition of shoulder dystocia, and the necessary maneuvers at the whole-text level.
Text 16: Google Translate was given a Pass overall because it accurately preserved the descriptions of fibroids, anemia, and treatment options. DeepL was also given a Pass because its less standard rendering of the intrauterine system did not change the intended management meaning.
Text 17: Both Google Translate and DeepL were given a "Pass" because they clearly and accurately explained the endocrine effects of severe postpartum hemorrhage and the possibility of Sheehan's syndrome.
Text 18: In general, both Google Translate and DeepL were given a passing grade because they kept the contraceptive counseling scenario and the legal-ethical reference to Gillick competence without changing it in a way that affected the clinical relevance.
Text 19: Both Google Translate and DeepL were given a "Fail" because they both incorrectly translate the main procedure, uterine artery embolization, as انسداد, which changes the main meaning of the text.
Text 20: In general, Google Translate was found to be a failure because it incorrectly translates "pessary" as "التحاميل المهبلية," which changes the intended management option. DeepL, on the other hand, was found to be a success because it correctly keeps the meaning of "prolapse management."
Text 21: In general, Google Translate was found to be a failure because it misinterprets "canalization" as "ضيق," which changes the developmental mechanism described in the source text. DeepL, on the other hand, was found to be a success because its wording keeps the intended embryological meaning.
Text 22: Both Google Translate and DeepL were given a passing grade because they correctly translated the definition of chronic pelvic pain and the need for a multidisciplinary assessment without changing the meaning.
Text 23: In general, Google Translate was given a Pass because it kept the Rotterdam diagnostic logic and the metabolic effects of PCOS. DeepL was also given a Pass because its unique way of saying "polycystic ovary" is less precise but does not change the meaning of the whole text.
Text 24: Overall, Google Translate was given a "Pass" rating because the screening logic is still clear, even though the terminology in primary HPV testing and colposcopy isn't always clear. DeepL was also given a "Pass" rating because its wording keeps the overall cervical screening message clear, even though it doesn't use as many specialized terms.

Text 25: In general, Google Translate got a Pass because it accurately conveyed the parts of a preoperative assessment. DeepL also got a Pass because its less accurate translation of electrocardiogram doesn't change the overall clinical meaning of the text.

Table 2. Problematic Technical Translations and Their Justification

Text	System	Source Term	Machine Translation	Type of Problem	Justification
1	—	—	—	—	No problematic technical rendering identified in either system
2	Google Translate	cervical malignancy	سرطان عنق الرحم	Less specialized equivalent	Medically acceptable, but less specialized than خباثة عنق الرحم
2	Google Translate	distant metastases	انتقاله إلى أعضاء أخرى	Terminological imprecision	Preserves the general meaning, but is less technical than النقائل البعيدة
3	Google Translate	prognosis	التشخيص	Semantic error	Changes prognosis to diagnosis and distorts a clinically important meaning
3	DeepL	oncological pathways	المسارات العلاجية المعتادة للأورام	Terminological imprecision	Narrows the meaning to treatment pathways only, though the general message remains recoverable
4	—	—	—	—	No problematic technical rendering identified in either system
5	Google Translate	serum concentrations	تركيزاتها في الدم	Terminological imprecision	Less precise than تراكيزها في المصل
5	Google Translate	specificity	دقتها	Terminological imprecision	Better rendered as نوعيتها in this diagnostic context
5	DeepL	interpretation is limited by physiological changes	يقتصر تفسير مؤشرات الأورام أثناء الحمل على التغيرات الفسيولوجية	Semantic error	Changes limited by into limited to and distorts the central meaning
5	DeepL	specificity	دقة هذه المؤشرات	Terminological imprecision	Better rendered as نوعية هذه الواسمات or نوعيتها
6	DeepL	adnexal masses	الأورام الملحقة بالأعضاء التناسلية	Less specialized equivalent	Weaker than the standard gynecological expression for adnexal masses

Text	System	Source Term	Machine Translation	Type of Problem	Justification
6	DeepL	serial imaging	فحوصات تصويرية متتالية	Terminological imprecision	Acceptable, but less direct than the reference term
6	DeepL	torsion	التواء الكيس	Semantic narrowing	Narrows the complication to cyst torsion rather than torsion of the adnexal mass more generally
7	Google Translate	invasive disease	انتشار المرض	Terminological imprecision	Weakens the pathological meaning of invasive disease
7	Google Translate	staging procedures	إجراءات تشخيصية أكثر شمولاً	Terminological imprecision	Omits the specific oncological sense of staging
7	DeepL	invasive disease	انتشار المرض	Terminological imprecision	Weakens the pathological meaning and contributes to whole-text inadequacy
8	Google Translate	trophoblastic proliferation	تكاثر غير نمطي للأرومة المغذية	Less specialized equivalent	Understandable, but less standard than الارومة الغاذية
8	DeepL	molar pregnancy	الحمل المولي	Less specialized equivalent	Less standard in Arabic medical usage than الحمل العنقودي
8	DeepL	trophoblastic proliferation	تكاثر غير نمطي للخلايا الغذائية	Semantic error	Does not preserve the technical trophoblastic sense accurately
9	Google Translate	internal bleeding	نزيف داخلي	Terminological imprecision	The reference prefers نزف داخلي, but the emergency meaning remains clear
9	DeepL	internal bleeding	نزيف داخلي	Terminological imprecision	The reference prefers نزف داخلي, but the meaning remains clear
9	DeepL	abdominal tenderness	حساسية البطن	Semantic error	Tenderness is not sensitivity
10	Google Translate	ectopic pregnancy	الحمل خارج الرحم	Less specialized equivalent	Less specialized than الحمل المنتبذ
10	Google Translate	serum hCG	هرمون الحمل في الدم	Terminological imprecision	Less precise than في المصل
10	DeepL	ectopic pregnancy	الحمل خارج الرحم	Less specialized equivalent	Less specialized than الحمل المنتبذ
10	DeepL	serum hCG	هرمون hCG في الدم	Terminological imprecision	Less precise than في المصل

Text	System	Source Term	Machine Translation	Type of Problem	Justification
11	DeepL	pulmonary embolism	انسداد رئوي	Terminological imprecision	Less precise than الانصمام الرئوي, but the emergency scenario remains clinically intelligible
12	Google Translate	gestational trophoblastic disease	أمراض المشيمة الحمليّة	Semantic error	Replaces trophoblastic disease with a placental expression and weakens the intended diagnosis
12	DeepL	surgical evacuation	إجهاض جراحي	Semantic error	Distorts the intended procedure; evacuation is not abortion here
13	DeepL	superficial endometriosis	مرض بطانة الرحم السطحية	Less specialized equivalent	Weaker and less standard than الانتباذ البطاني الرحمي السطحي
14	DeepL	colposuspension	رفع عنق الرحم	Semantic error	Changes the intended continence procedure and distorts the management meaning
15	—	—	—	—	No problematic technical rendering identified in either system
16	DeepL	levonorgestrel-releasing intrauterine system	النظام الرحمي الذي يفرز الليفونورجيستريل	Less specialized equivalent	Medically understandable, but less standard than the conventional gynecological rendering
17	—	—	—	—	No problematic technical rendering identified in either system
18	—	—	—	—	No problematic technical rendering identified in either system
19	Google Translate	uterine artery embolization	انسداد الشريان الرحمي	Semantic error	Distorts the name of the interventional radiological procedure; the accepted sense is انصمام

Text	System	Source Term	Machine Translation	Type of Problem	Justification
19	DeepL	uterine artery embolization	انسداد الشرايين الرحمية	Semantic error	Distorts the core procedure name and therefore affects the whole-text meaning
20	Google Translate	pessary	التحاميل المهبلية	Semantic error	A pessary is a support device, not a vaginal suppository
21	Google Translate	canalization	تضيق غير طبيعي	Semantic error	Canalization is a developmental formation process, not narrowing
22	—	—	—	—	No problematic technical rendering identified in either system
23	DeepL	polycystic ovaries	مبيض متعدد الكيسات	Terminological imprecision	Singular phrasing weakens the standard diagnostic expression, but the intended meaning remains clear
24	Google Translate	primary HPV testing	اختبار فيروس الورم الحليمي البشري	Terminological imprecision	Omits the procedural qualifier primary
24	Google Translate	colposcopy	تنظير المهبل	Less specialized equivalent	Communicates the general procedure, but is less precise in cervical screening context
24	DeepL	colposcopy	فحص المهبل بالمنظار	Less specialized equivalent	Understandable, but less specialized than the conventional Arabic rendering of colposcopy
25	DeepL	electrocardiogram	تخطيط القلب	Terminological imprecision	Less precise than تخطيط كهربية القلب, but the clinical meaning remains intact

Not all the renderings in Table 2 represent the same degree of error; some of them represent semantic errors, i.e., errors that change clinical meaning, whereas others are less accurate in terminology or less sophisticated translations.

4.4 Final Overall Results

This part shows the last quantitative comparison between Google Translate and DeepL based on how accurate they are with technical terms and how they judge the whole text as a Pass or Fail. It gives a brief overview of the total number of technical terms, the number of correct and incorrect renderings, the percentage of technical-term accuracy achieved by each system, and the number and percentage of texts that got Pass or Fail judgments at the whole-text level.Table 3. Final Overall Results

System	Total Technical Terms	Accurate Terms	Problematic Terms	Accuracy Percentage	Passed Texts	Failed Texts	Pass Percentage
Google Translate	194	177	17	91.24%	19	6	76%
DeepL	194	172	22	88.66%	20	5	80%

Table 3 shows that Google Translate got 177 of the 194 technical terms right, which is 91.24% of the time, and DeepL got 172 of the 194 technical terms right, which is 88.66% of the time. Google Translate passed 19 texts and failed 6, giving it a pass rate of 76%. DeepL passed 20 texts and failed 5, giving it a pass rate of 80%. These results show that the two systems are only slightly different in the chosen corpus. Google Translate is a little better at getting technical terms right, while DeepL is a little better at getting the whole text right. The results do not substantiate a robust assertion of distinct superiority for either system within the chosen corpus.

5. Conclusion

This study evaluated the efficacy of Google Translate and DeepL in translating 25 chosen gynecological texts from English to Arabic, focusing specifically on the precision of technical terminology and the conveyance of overall meaning. The results indicate that both systems can generate preliminary translations that may be applicable in certain contexts; however, their efficacy is inconsistent in specialized medical discourse. Google Translate was a little better at getting technical terms right, while DeepL was a little better at getting whole texts to pass. The difference between the two systems, on the other hand, is still small and based on the corpus rather than being clear-cut. The analysis also showed that not all bad translations were the same type or level of badness. Some had semantic mistakes that changed the diagnosis, treatment, or the main medical meaning of the source text. Others showed that the terms weren't always clear or that there were less specialized equivalents that didn't make the whole translation wrong. This confirms that the evaluation of machine-translated medical texts should not be limited to individual terms alone, but should also take into account the adequacy of the text as a complete medical message. The study indicates that machine translation can facilitate the preliminary translation of gynecological texts; however, it should be employed judiciously in specialized medical contexts and must not be depended upon without expert human revision. The results are confined to the chosen corpus and cannot be extrapolated to all medical texts or all machine translation outputs.

References

Alasmari, A., Alhumoud, S., & Alshammari, W. (2024). AraMed: Arabic medical question answering using pretrained transformer language models. In H. Al-Khalifa, K. Darwish, H. Mubarak, M. Ali, & T. Elsayed (Eds.), *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024* (pp. 50–56). ELRA and ICCL.

Alayba, A. M. (2025). Arabic natural language processing (NLP): A comprehensive review of challenges, techniques, and emerging trends. *Computers, 14*(11), 497. https://doi.org/10.3390/computers14110497

ALMutairi, M., AlKulaib, L., Aktas, M., Alsalamah, S., & Lu, C.-T. (2024). Synthetic Arabic medical dialogues using advanced multi-agent LLM techniques. In *Proceedings of the Second Arabic Natural Language Processing Conference* (pp. 11–26). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.arabicnlp-1.2

Chen, C. L., Dong, Y., Castillo-Zambrano, C., Bencheqroun, H., Barwise, A., Hoffman, A., Nalaie, K., Qiu, Y., Boulekbache, O., & Niven, A. S. (2025). A systematic multimodal assessment of AI machine translation tools for enhancing access to critical care education internationally. BMC Medical Education, 25(1), Article 1022. https://doi.org/10.1186/s12909-025-07452-9

Colina, S. (2025). *Fundamentals of translation* (2nd ed.). Cambridge University Press.

Herrera-Espejel, P. S., & Rach, S. (2023). The use of machine translation for outreach and health communication in epidemiology and public health: Scoping review. *JMIR Public Health and Surveillance, 9*, e50814. https://doi.org/10.2196/50814

Hu, B. (2025). *Charting translation reception: Methods and challenges*. Cambridge University Press. https://doi.org/10.1017/9781009569378

Kayano, Y., & Sugawara, S. (2025). Specification-aware machine translation and evaluation for purpose alignment. In *Proceedings of the Tenth Conference on Machine Translation* (pp. 113–141). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wmt-1.7

Lopez, I., Velasquez, D. E., Chen, J. H., & Rodriguez, J. A. (2025). Operationalizing machine-assisted translation in healthcare. *npj Digital Medicine, 8*, Article 584. https://doi.org/10.1038/s41746-025-01944-0

Merx, R., Docter, J. J., & Jongen, H. A. W. M. (2024). Machine translation technology in health: A scoping review with meta-analysis. *Medical Decision Making, 44*(8), 919–937. https://doi.org/10.1177/0272989X241281927

Montalt-Resurreccio, V., García-Izquierdo, I., & Muñoz-Miquel, A. (2025). *Patient-centred translation and communication*. Routledge. https://doi.org/10.4324/9780429299698

Munday, J., Ramos Pinto, S., & Blakesley, J. (2022). *Introducing translation studies: Theories and applications* (5th ed.). Routledge. https://doi.org/10.4324/9780429352461

Naveen, P., Gujjar, Y., Dhanesh, G. S., & Balakrishnan, V. (2024). Overview and challenges of machine translation for multilingual communication. *Heliyon, 10*(24), e40663. https://doi.org/10.1016/j.heliyon.2024.e40663

Noll, R., Berger, A., Kieu, D., Mueller, T., Bohmann, F. O., Muller, A., Holtz, S., Stoffers, P., Hoehl, S., Guengoeze, O., Eckardt, J.-N., Storf, H., & Schaaf, J. (2025). Assessing GPT and DeepL for terminology translation in the medical domain: A comparative study on the human phenotype ontology. BMC Medical Informatics and Decision Making, 25, Article 237. https://doi.org/10.1186/s12911-025-03075-8

Omar, L. I., & Salih, A. A. (2024). Systematic review of English/Arabic machine translation postediting: Implications for AI application in translation research and pedagogy. *Informatics, 11*(2), 23. https://doi.org/10.3390/informatics11020023

Prieto Ramos, F., Guzman, D., Candel-Mora, M. A., Way, C., Garcia, I., & Abdallah, K. (2024). The impact of specialised translator training and professional experience on legal translation quality assurance: An empirical study of revision performance. *The Interpreter and Translator Trainer, 18*(2), 313–337. https://doi.org/10.1080/1750399X.2024.2344948

Pym, A. (2025). *Risk management in translation*. Cambridge University Press. https://doi.org/10.1017/9781009546836

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation, 56*(2), 593–619. https://doi.org/10.1007/s10579-021-09537-5

Sebo, P., & de Lucia, S. (2024). Performance of machine translators in translating French medical research abstracts to English: A comparative study of DeepL, Google Translate, and CUBBITT. *PLOS ONE, 19*(2), e0297183. https://doi.org/10.1371/journal.pone.0297183

Taibi, M., Kenawy, A., & Antoniou, M. (2025). Healthcare translations for Arabic speakers in Australia: Impact of language variety and dissemination medium. *Translation and Interpreting Studies, 20*(1), 105–128. https://doi.org/10.1075/tis.24022.tai

Valdez, S. (2022). On the reception of biomedical translation: Comparing and contrasting health professionals' evaluation of translation options and expectations about the safe use of medical devices in Portuguese. *The Translator, 28*(4), 538–554. https://doi.org/10.1080/13556509.2022.2159297

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