وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة مصطفى اسطمبولي _ معسكر
University of Mustapha Stambouli – Mascara
كلية الآداب واللغات
Faculty of Letters and Languages

رئاسة الجمهورية
المجمع الجزائري للغة العربية

Proceedings of the National Conference

The Digital Shift and Translation in the Service of the Arabic Language

Part Two

منشورات المجمع الجزائري للغة العربية
2026

** Any trade names or product names mentioned in this book may be subject to trademarks, intellectual property, or patent protection, and remain the registered trademarks or intellectual property of their respective owners, or product descriptions, etc. Even where a given mark is not specifically indicated in this work, this can in no way be taken to imply that such names may be considered unrestricted under intellectual property and trademark protection legislation, and therefore free for use by anyone.

Cover design : *Hemza Saadi* - Algerian Academy of the Arabic Language/ **2026

Legal deposit: **July 2026.

ISBN : 5-6-9672-9969-978

06 Colonel Mohamed Bouguerra Street - El Biar - Algiers.

*** Bibliographic Description

The Digital Shift and Translation in the Service of the Arabic Language: Part Two -Joint national conference between the Algerian Academy of the Arabic Language and the University of Mascara on the occasion of International Translation Day - 30 September 2026 / Mohamed Babchikh, Kawthar Farah, Chahrazad Khelfi... [et al]. - Algiers: Algerian Academy of the Arabic Language, 2026. -262 p.; 16×24 cm.

ISBN: **5-6-9672-9969-978**

Keywords: Scientific translation - research and conferences; Arabic language - terminology; foreign languages; digital shift; artificial intelligence; dictionaries; Arabic rhetoric.

Dewey Decimal Classification: 418.02

Honorary Presidency

Prof. Cherif Meribai,

President of the Algerian Academy of the Arabic Language

Prof. Younes Mouschaal,

Rector of Mustapha Stambouli University, Mascara.

Conference Supervisor:

Prof. Habib Bouzouada

Dean of the Faculty of Letters and Languages.

Scientific Committee of the conference

Presidency of the Committee

Prof. Saida Kohil, Head of the Translation Committee, Algerian Academy of the Arabic language

Prof. Rahma Boushaba, University of Mascara

Members:

- Prof. Salah Khennour, Algerian Academy of the Arabic Language
- Prof. Amor Lahcen, Algerian Academy of the Arabic Language
- Prof. Driss Mohamed Amine, University of Mascara
- Prof. Salem Mohamed, University of Mascara
- Prof. Fekir Abdelkader, University of Mascara
- Prof. Reda Baba Ahmed, University of Mascara
- Prof. Faiza Boukhalf, University of Chlef
- Prof. Souhila Meribai, University of Algiers 2
- Dr. Fathallah Belbia, University of Mascara
- Dr. Chahrazed Khelfi, University of Mascara
- Dr. Ahlam Hal, University of Mascara
- Dr. Abdel Fettah Ben Ahmed, University of Mascara
- Dr. Lynda Kazi Tani, University of Mascara
- Dr. Mansouri Abdelrahmane, University of Mascara
- Dr. Sofiane Djeffal, University of Mascara
- Dr. Mohamed Babchikh, Algerian Academy of the Arabic Language
- Dr. Karima Bouras, University of Mostaganem
- Dr. Naima Bougherira, University of Annaba
- Dr. Chawki Bounaas, Algerian Academy of the Arabic Language
- Dr. Zine El Abidine Benhaddi, University of Mascara
- Dr. Alaa Ternifi, University of Mascara
- Dr. Mustafa Hanifi, University of Mascara
- Dr. Rayhan Khouissat, University of Blida
- Dr. Fatima Zahra Nebbati; University of Tlemcen
- Dr. Malika Bacha, University of Relizane

Table of Contents

Preamble

This conference addresses a contemporary topic within the field of interdisciplinary collaboration among linguistics, translation, and artificial intelligence. It focuses on recent developments in the application of scientific research findings within this collaborative framework to support languages in a knowledge production environment. This transformation has led to a redefinition of key concepts such as translation, the boundaries of the translational act, and the roles of translators within augmented translation, alongside the integration of disciplines that support language use and shift toward AI-driven translation professions. It also introduces new scientific and methodological challenges for researchers in Arabic translation studies and across the broader academic sphere.

The current digital transformation constitutes one of the most significant developments accompanying the project of translation from foreign languages into Arabic, at both theoretical and practical levels. It has profoundly reconfigured the knowledge environment surrounding the field's foundational concepts, most notably translation tools, the role of the translator, and the nature of the translation process itself. In particular, the rapid evolution of computer-assisted translation technologies, the emergence of neural machine translation systems, and the advancement of artificial intelligence-driven language models have collectively resulted in a deep restructuring of translation work environments. These environments are now embedded within integrated digital ecosystems built on large-scale databases, sophisticated algorithms, and cloud-based platforms used for project management and quality assurance.

This transformation extends beyond mere tool enhancement to a fundamental redefinition of the relationship between humans and machines within the translation process. On the one hand, modern technologies have

increased productivity, accelerated workflows, and expanded access to multilingual digital content, thereby facilitating integration into the global digital space. On the other hand, they raise complex issues related to output quality, the reliability of algorithmic models, data ethics, privacy protection, and intellectual property, in addition to their impact on linguistic identity, the status of augmented translation vis-à-vis human translation, and the translator's evolving role in the labor market.

This transformation extends beyond mere tool enhancement to a fundamental redefinition of the relationship between humans and machines within the translation process. On the one hand, modern technologies have increased productivity, accelerated workflows, and expanded access to multilingual digital content, thereby facilitating integration into the global digital space. On the other hand, they raise complex issues related to output quality, the reliability of algorithmic models, data ethics, privacy protection, and intellectual property, in addition to their impact on linguistic identity, the status of augmented translation vis-à-vis human translation, and the translator's evolving role in the labor market.

Despite notable advances in machine translation technologies, Arabic continues to face specific challenges linked to the complexity of its morphological and syntactic systems and its semantic richness, alongside a lack of high-quality specialized data, limited digital resources, and insufficient terminological representation. These factors directly affect the performance and output quality of intelligent models in Arabic. Recent studies also indicate that the performance of language models in Arabic remains lower compared to digitally dominant languages such as English and French, leading to issues such as distortion, terminological inconsistency, and hallucinations.

In this context, several key research questions arise: How can Arabic benefit from digital transformation and augmented translation? How can terminological

translation into Arabic be addressed through strategies such as borrowing, localization, literal translation, adaptation, equivalence, modulation, transposition, and calque? Can the training of specialized language models effectively resolve challenges related to the representation of scientific knowledge in Arabic? What are the prospects for developing linguistic models at the morphological, syntactic, semantic, and stylistic levels for translation into Arabic? Can institutional translation projects succeed within integrated digital environments that combine human translators and intelligent systems? How can post-editing and revision contribute to improving the quality of Arabic digital content and support AI systems? And what should be the nature of academic training in augmented translation into Arabic?

Within this framework, the conference aims to bring together researchers - linguists, translators, computer engineers, and AI experts - within an interdisciplinary approach to discuss the latest scientific methodologies for training large language models for Arabic translation. It also seeks to explore the associated cognitive, ethical, and technical challenges, strengthen the interface between translation studies and computational sciences, and showcase Arab experiences in augmented translation, prompt engineering, post-editing practices, and translation revision, with the aim of enhancing the quality of knowledge transfer into Arabic.

Objectives of the Conference

The conference aims to:

- Analyze the current state of augmented translation into Arabic and critically assess its performance.
- Examine state-of-the-art methodologies for training intelligent models for low-resource languages, with a particular focus on Arabic.
- Investigate the role of Arabic linguistic resources, including text corpora and databases, in enhancing translation quality.

- Foster interdisciplinary integration between translation studies, artificial intelligence, and machine translation in support of the Arabic language.
- Strengthen the presence of Arabic in the global digital space and promote institutional translation into and from Arabic.
- Address the ethical and professional implications of artificial intelligence in Arabic translation.
- Anticipate the future of the Arabic translator in the era of generative AI and identify pathways for improving the quality of Arabic translation output.

Conference Themes

1- Theoretical Foundations and Interdisciplinary Integration in Translation from Foreign Languages into Arabic

- The digital shift in translation into and from Arabic.
- Theoretical foundations and interdisciplinary integration between translation theories and artificial intelligence.

2- Training Large Language Models for Arabic Translation: From Development to Deployment

• Academic training frameworks in augmented translation in service of the Arabic language.

• Strategies and methodologies for designing large language models for augmented Arabic translation projects.

• Building parallel corpora and developing linguistic databases to enhance Arabic digital content: achievements and case studies.

3- Technical and Linguistic Challenges in Arabic Translation

• Analysis of the technical and linguistic causes of difficulties in augmented translation into Arabic.

• Challenges of Human–machine interaction in translation within the framework of translation ethics.

• Proposing practical technical and linguistic solutions through corpus-based

approaches.

4- Quality Assessment of Augmented Arabic Translation

• Linguistic hallucination in augmented Arabic translation: analysis of applied models.

• Terminological, morphological, syntactic, contextual, and stylistic quality in augmented translation into Arabic: model-based studies.

• The role of prompt engineering, post-editing, and revision in improving the quality of Arabic translation output: case studies in scientific translation.

5- Arab Initiatives in Augmented Translation

• Presentation of scientific translation projects at the Algerian Academy of the Arabic Language (open-source initiatives).

• Overview of institutional projects from Arabic intelligent translation platforms in the era of digital transformation.

The Scientific Committee of the Conference

Quality Assessment of LLM-based Machine Translation of Academic Texts into Arabic: A Multidimensional Composite Framework

Mohamed Babchikh

Laboratory TRADIL, Annaba, University of Medea

Abstract

Artificial Intelligence is transforming the field of translation, with Large Language Models (LLMs) further reshaping workflows by accelerating processing and reducing costs. In the context of Arabic, LLMs have significantly enhanced translation quality compared to earlier machine translation approaches, leading to increased adoption by translators, public institutions, and academics. This study evaluates Arabic translations of three journal articles generated by ChatGPT Plus using three complementary quality assessment methods. The translations achieved a mean COMET-QE score of 0.773, indicating generally high semantic adequacy. This result, however, contrasts with the MQM analysis, which identified 228 errors, mainly involving style, accuracy, and additions. The post-editing data also showed that 67.0% of the segments underwent at least one word-level change, with a mean normalised edit rate of 0.222. The three assessment methods did not always produce consistent results. Some translations received relatively high COMET-QE scores but still needed considerable post-editing before reaching scholarly publication standards.

The lack of agreement among the measures suggests that no single quality dimension can adequately represent overall translation quality, making their combined use more informative.

Key Terms: AI-based Machine Translation, Arabic, TQA Composite Framework, Hallucination, Terminology Inconsistency, *Translationese*.

Introduction

1.1 Background

The rapid development and widespread adoption of LLMs have transformed the translation industry by enhancing machine translation quality and broadening its application to diverse user groups, including academics. In Algeria, the pressing need for increased visibility among universities and researchers has prompted the use of AI-driven machine translation tools for drafting, revising, and translating academic publications into English and other languages, including Arabic.

An additional application of generative AI in academic translation is the "*Arwiqat Al-Ouloum*" ("Corridors of Science") project initiated by the Algerian Academy of the Arabic Language. This project aims to translate open-source academic research papers from a range of disciplines, originally written in French and English, into Arabic using ChatGPT Plus. The resulting translations are subsequently reviewed by members of the Academy's translation committee.

Accordingly, the workflow for producing these translations integrates AI-driven machine translation, specifically LLMs such as ChatGPT Plus, with

human expertise for post-editing and proofreading. This approach addresses terminological inconsistencies, stylistic issues, and unsupported additions.

Translating academic texts involves more than accurately transferring meaning. Terminology must remain consistent, concepts must be conveyed precisely, and the translation must follow the conventions of academic writing in the target language. This requires both linguistic competence and knowledge of the subject matter. LLM-generated translations may nevertheless retain features of the source language, resulting in literal syntax, unnatural phrasing, and other forms of "*translationese*" (Li et al., 2025). Károly (2022) similarly stresses the need to meet the expectations of the target academic community while preserving the content and purpose of the source text.

Recent work has shown the possibilities and limitations of integrating LLMs into professional translation workflows. It has also stressed the need for human participation during pre- and post-editing. Previous studies have shown that LLMs can benefit from document-level context in translation but continue to produce errors needing human review (Karpinska & Iyyer, 2023).

Recent studies have reported persistent terminology and stylistic problems in LLM-generated translations, raising questions about how well existing evaluation methods capture these aspects of quality (Zhang et al., 2025).

The drawbacks of relying on individual evaluation measures have also been documented. Kumar et al. (2026) found divergences between ChrF++ and BLEU, while Al Swuee Ali Ali and Babchikh (2024) showed that MQM offered more detailed information on error types and severity than an automated metric when evaluating English–Arabic legal machine translation. The results support the use of complementary evaluation methods.

Collectively, these studies indicate that no single evaluation method can fully capture translation quality and support the use of a multidimensional framework uniting automated quality estimation, human error analysis, and post-editing effort.

1.2 Problem Statement

Although LLMs have improved machine translation considerably, assessing the quality of their output remains difficult, particularly for academic translation into Arabic. Existing assessment methods generally rely on automatic metrics, human evaluation, or a combination of both, with each capturing different aspects of translation quality. Against this background, the present study examines Arabic translations of academic texts generated by ChatGPT Plus.

1.3 Objectives

This study assesses the quality of ChatGPT Plus-generated Arabic translations of academic journal articles from three complementary

perspectives. COMET-QE is used to evaluate semantic adequacy, MQM to identify and classify translation errors, and normalised edit rate to quantify the technical post-editing effort required to produce the final translations. By bringing these three measures together, the study examines whether a multidimensional approach can provide a more comprehensive assessment of LLM-generated academic translations into Arabic.

1.4 Research Questions

The study addresses the following research questions:

RQ1. What level of semantic adequacy is achieved by ChatGPT Plus-generated Arabic translations of academic texts, as measured by COMET-QE?

RQ2. What are the most frequent types and severity levels of translation errors identified through MQM 2.0?

RQ3. What level of technical post-editing effort is required to transform the ChatGPT Plus-generated translations into their final human post-edited versions?

RQ4. To what extent do COMET-QE, MQM 2.0, and technical post-editing effort converge or diverge in their assessment of translation quality?

1.5 Significance

This study contributes to the field of LLM-based machine translation by introducing a multidimensional composite system for assessing the quality of academic texts translated into Arabic. By employing COMET-QE metrics,

MQM measures, and post-editing effort, the study demonstrates the advantages of integrating automated and human-centred approaches to translation quality assessment. Empirically, the research expands the limited body of literature on LLM-generated academic translations into Arabic by analysing a corpus of AI-powered translations produced in Algeria. The findings may be useful to translation scholars, professional translators, and educational organisations involved in evaluating and integrating LLMs into academic translation workflows.

2. Literature Review

2.1 Theoretical Background

Machine Translation Quality Assessment (MTQA) is still a major challenge in translation technology, particularly as advances in NMT and LLM-based systems make translation errors increasingly difficult to detect using traditional evaluation methods (Castilho, 2020; Specia et al., 2018; Rei et al., 2022).

Traditional automatic metrics deliver rapid quantitative evaluation but have shortcomings in assessing semantic adequacy and contextual appropriateness. These limitations have contributed to the development of neural evaluation models such as COMET and COMET-QE, alongside human-centred frameworks such as Multidimensional Quality Metrics (MQM) (Rei et al., 2020, 2022).

2.1.1 Evolution of Machine Translation Evaluation

Early automatic MT evaluation was largely based on lexical similarity between machine output and human reference translations. One of the best-known reference-based metrics, BLEU, calculates n-gram overlap between a machine translation and one or more references (Papineni et al., 2002). However, its reliance on lexical matching means that valid paraphrases and alternative renderings may receive lower scores.

Subsequent metrics attempted to address some of these limitations. METEOR incorporated more flexible forms of lexical matching (Banerjee & Lavie, 2005), TER quantified the edits required to transform machine output into a reference translation (Snover et al., 2006), and chrF introduced character-level comparison, rendering it especially useful for morphologically rich languages (Popović, 2015).

2.1.2 From Automatic Evaluation to Hybrid Quality Assessment

Human evaluation examines aspects of translation quality such as accuracy, fluency, style, and terminology, while also revealing specific errors that may not be reflected in aggregate automatic scores (Callison-Burch et al., 2007).

Human evaluation methods include Direct Assessment, in which evaluators rate translation quality on a scale (Graham et al., 2013, 2015), as well as pairwise ranking, in which they select the better of two translations (Callison-Burch et al., 2007). Although these methods deliver overall

quality judgements, they offer limited diagnostic information about specific errors.

The Multidimensional Quality Metrics (MQM) framework tackles this limitation through a thorough error taxonomy that enables errors to be classified by type and severity. It has become a broadly adopted framework for human evaluation of machine translation, including neural and LLM-based output (Lommel et al., 2014).

2.1.3 From Reference-Based to Reference-Free Evaluation

Traditional automatic metrics frequently depend on reference translations, whose production is costly and time-consuming. In many real-world settings, however, translation quality must be assessed when no reference translation is available. These limitations are why people began working on Quality Estimation, a part of machine translation that assesses translation quality without a reference translation (Specia et al., 2018). Unlike comparing a translation to a perfect one, Quality Estimation models look at the source sentence and the machine-generated translation to estimate its quality. At first, people used rule-based methods to assess quality, such as sentence syntax, word co-occurrence, sentence complexity, word ambiguity, and the degree to which words in the source and the translation matched. Although these methods worked well, they did not perform well for all types of translations and languages. Now, as deep learning advances rapidly and multilingual language models become highly effective, Quality

Estimation has changed significantly (Specia et al., 2018). Models can now learn from collections of texts in many languages and better understand what makes a good translation.

2.1.4 Neural Evaluation Metrics

With the advancement of language models such as BERT and XLM-R, there has been a paradigm shift in assessing machine translations (Conneau et al., 2020; Devlin et al., 2019). Language models have a high capability to understand word semantics and to compare the source text with the translation to determine whether they share the same semantic meaning.

COMET is a neural metric for assessing machine translation quality. It looks at the translated sentence and sometimes a reference sentence, then uses this information to estimate translation quality. It is notable because it can learn from human evaluations and use them to improve its own judgements. (Rei et al., 2020)

COMET can learn from various kinds of evaluations. For example, it can learn from Direct Assessment, in which people evaluate translations directly (Rei et al., 2020). It can also learn from the Human-Mediated Translation Edit Rate, which is the rate at which people edit translations to improve them. COMET can even learn from MQM annotations, which are notes that explain what is good or bad about a translation.

Because COMET is so flexible, people have developed versions of it such as COMET-QE and COMETKiwi (Rei et al., 2022). Unlike COMET, COMET-QE

does not require a reference translation, making it especially suitable for genuine translation cases where reference translations are unavailable. These recent versions can evaluate translations without looking at a reference sentence. They only need the source sentence and the translated sentence to estimate translation quality, which makes evaluation easier.

2.1.5 Challenges in Automatic Translation Evaluation

Recent research has shown that neural measures outperform lexical measures, but they have serious limitations (Freitag et al., 2022). One of the most worrying issues is that the measures are biased, meaning machine translation systems can be optimised to achieve high scores rather than better translation itself. (Freitag et al., 2022)

Another major issue is that it is hard to understand why a translation received a score (Kocmi et al., 2024). Older metrics do not explain much, and even newer neural metrics usually give only a number without showing what the mistakes are or where they occur in the translation. This makes it difficult for translators and researchers to use these metrics to identify problems and fix them. Machine translation is still not perfect, so we need to evaluate it and help it improve.

Modern research shows that while the correlation between COMET and humans is high, the tool's reliability is lower in specialised domains where the accuracy of terminology, law, or medicine is critical. In addition, automated metrics do not account for terminology errors or semantically

incorrect use of the domain's terms that an expert human MQM annotator can easily identify. The necessity of human-oriented neural evaluation is thus evident since it involves both semantic sensitivity and thorough error detection. This demonstrates the importance of developing evaluation models that combine the predictive power of neural metrics with human-centred evaluation, thereby establishing the foundation for studying COMET-QE and MQM.

2.1.6 Towards Human-Centred Neural Evaluation

Current MT assessment depends on cooperation rather than competition between automated models and human evaluators. The purpose of neural assessment is to imitate human judgement but be fast and reliable. The tool is not meant to replace human judgement regarding quality. This is why models have been used in human assessments such as the MQM. The recent advances in xCOMET (Guerreiro et al., 2024), COMET-22 (Rei et al., 2022), and document-level COMET (Vernicos et al., 2022) illustrate this trend. The method predicts a sentence's quality and helps identify mistakes. MQM offers human judgement regarding the type of error, its seriousness, and the locations where the errors arise. MT evaluation has therefore become complicated and increasingly relies on neural evaluation.

Hence, studies on the evaluation of machine translation output increasingly recommend combining neural measures with human evaluations to achieve more reliable, transparent, and informative

assessments of machine translation. This serves as the basis for the combined use of COMET-QE and MQM in assessing machine translation quality.

2.2 Previous Studies

Recent studies have examined the effectiveness of COMET-QE and related reference-free metrics for MT quality estimation. Rei et al. (2022) introduced COMETKIWI – a multilingual quality estimation system. The findings show that it performs extremely well in sentence- and word-level quality estimation tasks.

Similarly, Chimoto and Bassett (2022) applied COMET-QE to low-resource neural machine translation, selecting sentences for translation in the Swahili-to-English, Kinyarwanda-to-English, and Spanish-to-English corpora. Moreover, Vernikos and Popescu-Belis (2024) have combined translation hypotheses across five language pairs using COMETKIWI and demonstrated its superiority over beam search and QE-reranking methods, with minimal hallucinations. More recently, Marmonier et al. (2026) compared various QE metrics on a parallel English French corpus in which translations were generated by NMT and LLM systems. It has been shown that the current QE metrics perform better with NMT than with LLMs.

Overall, these studies show the effectiveness of COMET-QE as a reference-free evaluation metric while also disclosing its limitations upon application to LLM-powered translations.

The MQM framework has gained widespread popularity as the most used human-centred system for assessing machine translation, classifying translation errors by type and severity. Lommel et al. (2014) established MQM as a framework for the systematic human evaluation of translation quality. Its application has since extended to different language pairs and translation settings. Park and Padó (2024) used MQM to examine English–Korean translations and found that multidimensional annotation captured problems related to accuracy, fluency, and style. Fernandes et al. (2023) explored a different direction with AutoMQM, using LLMs to automate the identification of MQM errors. Al Swuee Ali Ali and Babchikh (2024) compared MQM and METEOR on an English–Arabic legal corpus translated using Amazon Translate. Their analysis showed that MQM offered a more detailed view of translation quality, identifying terminology, accuracy, and fluency errors that were not apparent from the automatic metric scores.

Collectively, these studies show that MQM provides richer diagnostic information than automatic metrics by identifying the type, severity, and location of translation errors.

Post-editing has been extensively used to assess the usability of machine-translated output in practice. Cui et al. (2023) found that post-editing machine-translated output can reduce temporal and cognitive effort compared with translation from scratch, although effort varies according to

text type. Nonetheless, the amount of effort required to accomplish post-editing tasks depends on the type of text. Similarly, Alvarez-Vidal and Oliver (2023) studied the quality of English-into-Spanish machine translation using post-editing effort measurements and automatic metrics. They found that post-editing effort reveals translation quality features that cannot be captured by automatic metrics.

These results suggest that post-editing effort supplements both automatic and human evaluation by measuring the practical usability of machine-translated texts.

2.3 Research Gap

Previous research has pointed out the benefits of COMET-QE, MQM, and post-editing effort in machine translation evaluation; however, earlier investigations have mostly focused on assessing these approaches separately or comparing them. Furthermore, most studies have centred on improving machine translation tools and assessing translations of general and literary texts, while overlooking evaluations of LLM-produced translations into Arabic. The absence of an evaluation mechanism that would embrace all three elements of the assessment (i.e., neural quality estimation, fine human error analysis, and the effort of post-editing) has determined the aim of this paper to propose a composite framework which will integrate the three aforementioned methods into a unified

multidimensional evaluation framework for the evaluation of ChatGPT Plus translations related to academic journals.

3. Methodology:

1. Research Design

This study adopts a multidimensional design combining quantitative and qualitative approaches to TQA. COMET-QE provides segment-level estimates of semantic adequacy, while the normalised edit rate quantifies the technical effort required for post-editing. MQM 2.0 (Appendix 1) complements these measures by identifying and classifying translation errors. Combining the three methods makes it possible to examine aspects of translation quality that cannot be fully captured by automated scores alone, particularly terminology, conceptual accuracy, academic style, and other domain-specific issues.

The study is also descriptive and analytical. It describes the quality of AI-powered translations across the selected corpus and analyses the types of errors that affect the translation of academic writings.

Sample:

The sample consists of three academic articles addressing the use of artificial intelligence in the legal field:

Article 1: Billion, A., & Guillermin, M. (2019). *Intelligence artificielle juridique: enjeux épistémiques et éthiques*. *Cahiers Droit, Sciences & Technologies, 8*, 131–147. DOI: 10.4000/cdst.774. The Arabic translation

was post-edited by Prof. Souhila Meribai and published under the title **الذكاء الاصطناعي القانوني : رهانات معرفية وأخلاقية**.

Article 2: Haley, P., & Burrell, D. N. (2025). *Integrating artificial intelligence into law enforcement: Socioeconomic and ethical challenges*. *SocioEconomic Challenges, 9*(2), 60-77.

DOI: 10.61093/sec.9(2).60-77.2025. The Arabic translation was post-edited by Prof. Omar Lahcene and published under the title **دمج الذكاء الاصطناعي في أجهزة إنفاذ القانون: الرهانات الاجتماعية-الاقتصادية والأخلاقية**.

Article 3: Emejuo, C. C., Joseph, O. C., Odeyemi, E., & Igwe, A. O. (2025). *The impact of artificial intelligence on legal practice: Enhancing legal research, contract analysis, and predictive justice*. *International Journal of Science and Research Archive, 14*(1), 603-611. DOI:

10.30574/ijsra.2025.14.1. 0124.The Arabic translation was post-edited by the author of the present study and published under the title **أثر الذكاء الاصطناعي في الممارسة القانونية: تعزيز البحث القانوني، وتحليل العقود، والعدالة التنبؤية**.

The selection was based on three criteria. First, the articles had to be academic in nature and relevant to artificial intelligence in legal practice. Second, they had to contain specialised vocabulary related to law, legal professions and research, digital transformation, and artificial intelligence. Third, they had to have published Arabic reference translations prepared by an expert translator in the project *Arwiqat Al-Ouloum* of the Algerian

Arabic Language Academy. These published reference translations were used as reliable benchmarks for the human evaluation stage.

ChatGPT Plus was used as the AI-powered translation system, and the translated outputs were then prepared for evaluation alongside the source texts.

Data Collection

In order to collect data, first, the three academic articles were selected and prepared as source texts. Second, the translation outputs were generated using ChatGPT Plus. A consistent prompting strategy (Appendix) was used to reduce variation among the three translations and ensure the system produced an academic translation appropriate for the target language and domain.

Third, the evaluation data were organised in Excel files. Each file combined the aligned source segments, the corresponding translation generated using ChatGPT Plus, the reference translation, the COMET-QE score section, and the MQM annotation columns. The alignment was reviewed manually to ensure that each source segment corresponded to its equivalent translated segment. This step was indispensable as accurate alignment is necessary for reliable segment-level evaluation.

Data Analysis

The data analysis integrates COMET-QE semantic adequacy assessment, MQM 2.0, and normalised edit rate. The COMET-QE evaluation was

implemented in Python (version 3.11.1) using Visual Studio Code. The Python script processed the Excel files, extracted the source and AI-based translation segments, generated COMET-QE scores for each segment, and inserted the scores back into the evaluation files.

The second stage consisted of MQM 2.0 annotation and error analysis. The MQM 2.0 Scorecard was used in Excel to classify translation errors according to category and severity. The main error categories considered in this study included Accuracy, Terminology, Linguistic Standards, Style, Locale Conventions, and Audience Appropriateness. Attention was given to mistranslation, omission, addition, terminology inconsistency, conceptual distortion, ambiguity, and unnatural academic phrasing.

Each error was weighed according to a severity level as minor, major, or critical. Minor errors affect fluency or readability without significantly changing meaning. Major errors relate to the argument's accuracy, the rendering of specialised terminology, or the clarity of the target text. Critical errors seriously distort the meaning of the ST, omit essential information, or produce a misleading interpretation.

The MQM 2.0 Scorecard automatically calculated penalties according to the severity of annotated errors. These penalties were aggregated at segment, article, and corpus levels. The COMET-QE scores and MQM penalties were then compared to determine whether automatic quality estimation corresponded to human error-based evaluation. This

implementation made it possible to identify both the merits and the limitations of AI-powered translation in academic texts on artificial intelligence and the legal profession.

In the third stage, the effort of technical post-editing was measured by comparing each segment translated by ChatGPT Plus with its final human post-edited version. A normalised word-level edit rate was computed as the total number of substitutions, deletions and insertions required to transform the machine-generated translation into the final post-edited text, divided by the number of words in the post-edited segment. A score of zero meant that no intervention at the word level was required, whereas higher scores indicated a greater amount of technical post-editing effort. This measure was used as an indicator of the amount of human text intervention and practical usability, instead of a direct measure of cognitive load.

For descriptive analysis, COMET-QE scores were grouped into four quality bands: Excellent (≥ 0.80), Good (0.70–0.79), Medium (0.50–0.69), and Low (< 0.50). These bands were adopted as analytical categories to facilitate comparison of segment-level score distributions across the three articles. The 0.50 boundary is broadly consistent with the dichotomisation of human Direct Assessment scores used by Falcão et al. (2024), who distinguished ratings above and below 50 on a 0–100 quality scale, while COMET-22 operates on scores bounded between 0 and 1. However, the finer distinctions between Medium, Good, and Excellent were defined

operationally for the purposes of the present study and should not be interpreted as standardised COMET-QE quality thresholds (Falcão et al., 2024).

4. Results

In this section, the results of the multidimensional evaluation of the three Arabic translations produced by ChatGPT Plus are presented. To this end, the analysis merges the scores from the COMET-QE semantic adequacy assessment with those from the MQM 2.0 human-error annotation and the technical post-editing effort.

4.1. COMET-QE Neural Quality Estimation

The three Arabic translations were evaluated for quality using the COMET-QE reference-free neural quality estimation metric. COMET-QE predicted the translation quality scores considering only the source sentence and the translation of the sentence. Table 1 shows the descriptive statistics for each article.

Table 1: COMET-QE Scores

Article	Segments	**Mean COMET-QE**	Standard Deviation	**Minimum**	**Maximum**
Article 1	147	0.729	0.094	0.454	0.879
Article 2	106	0.803	0.049	0.521	0.866
Article 3	95	0.807	0.060	0.510	0.873
Total	348	0.773	0.083	0.454	0.879

Across the 348 segments, the mean COMET-QE score was 0.773 (SD = 0.083), as shown in Table 1. Article 3 had the highest mean (M = 0.807), followed closely by Article 2 (M = 0.803), whereas Article 1 had the lowest (M = 0.729). Scores varied least in Article 2 (SD = 0.049) and most in Article 1 (SD = 0.094), indicating greater consistency in semantic adequacy for Article 2.

To identify segments requiring closer human examination, the distribution of COMET-QE scores was analysed at the segment level.

4.2. Distribution of COMET-QE Scores at Segment Level

Whereas the descriptive statistics discussed above provide an overall assessment of translation quality, they do not offer insight into how that quality is distributed across translation segments. It is only through analysing the distribution of COMET-QE scores among the individual translation segments that one can understand what portion of the translations have high, medium, and low semantic adequacy, and whether the average score is an accurate reflection of the overall performance, or whether there are a few segments responsible for the poor results. To address this issue, the translations are classified into quality classes as shown in Table 2.

Table 2. Distribution of COMET-QE Quality Bands Across the Three Articles

Quality Band	Article 1 (n = 147)	Article 2 (n = 106)	Article 3 (n = 95)	Overall (N = 348)
Excellent	12 (8.2%)	8 (7.5%)	20 (21.1%)	40 (11.5%)
Good	85 (57.8%)	93 (87.7%)	70 (73.7%)	248 (71.3%)
Medium	46 (31.3%)	5 (4.7%)	5 (5.3%)	56 (16.1%)
Low	4 (2.7%)	0 (0.0%)	0 (0.0%)	4 (1.1%)

As shown in Table 2, most translated segments received high COMET-QE scores across the three analysed articles, representing 71.3% of the whole corpus (248 out of 348 segments). Furthermore, 11.5% of the segments were defined as excellent, while 16.1% of them were of medium quality. Only 1.1% of the translated segments were characterised as low-quality. The highest share of high-quality translations was observed in Article 2 (87.7%), whereas there were no low-quality segments in this article. The highest number of excellent translations was also found in Article 3 (21.1%).

Although the COMET-QE score distribution provides insight into the semantic adequacy of translations, it cannot pinpoint the causes of lower scores across segments. To provide a more thorough diagnostic assessment, the findings from the MQM analysis of human errors are shown below.

Although the COMET-QE score distribution provides insight into the semantic adequacy of translations, it cannot pinpoint the causes of lower scores across segments. To provide a more thorough diagnostic assessment, the findings from the MQM analysis of human errors are shown below.

4.3. Human Assessment based on MQM

Despite providing a rough idea of the semantic adequacy of translation, COMET-QE lacks information on the types of errors made by translators, where they occur, and their severity. To further evaluate the translations, the translated texts were annotated using the MQM 2.0 framework.

Table 3. MQM evaluation summary

Article	Segments	Minor	Major	Critical	Total MQM Penalty	Penalty per Segment	MQM Score (/100)
Article 1 (Meribai)	**147**	**69**	**8**	**0**	**109**	**0.74**	**99.26**
Article 2 (Lahcene)	**106**	**81**	**9**	**0**	**126**	**1.19**	**98.81**
Article 3 (Babchikh)	**95**	**55**	**6**	**0**	**85**	**0.89**	**99.11**
Total	**348**	**205**	**23**	**0**	**320**	**0.92**	—

According to the MQM assessment results, there were 228 annotated errors across the three articles, resulting in a total penalty score of 320. Minor errors accounted for the largest number of annotations (205; 89.9%), whereas major errors were significantly fewer, at 23 (10.1%). Critical errors were not found in the corpus. In terms of MQM penalties, Article 2 had the highest total MQM penalty (126), which implies that the average penalty score of this article was 1.19 per segment, compared to 0.74 and 0.89 for Articles 1 and 3, respectively. This implies that despite the high overall quality of the translations provided by ChatGPT Plus, human post-editing was still required.

Table 4. Distribution of MQM error categories

Error Category	Article 1	Article 2	Article 3	Total
Terminology	0	6	17	23
Accuracy	16	14	6	36
Mistranslation	2	3	9	14
Omission	8	3	7	18
Addition	4	4	26	34
Linguistic Conventions	2	0	0	2
Grammar	0	0	2	2
Punctuation	0	1	1	2
Style	45	58	7	110

As shown in Table 4 above, style, addition, and accuracy are the main categories of MQM errors throughout the corpus. The prevalence of stylistic errors shows that, while ChatGPT Plus has preserved the semantics of the source texts, it has often failed to adhere to the stylistic requirements of academic writing in Arabic. Various sections contained elements typically found in *translationese*, such as over-literal syntax, inappropriate collocations, excessive nominalisation, and source-language contamination. For instance, in Article 1, several English and French structures were translated directly into Arabic, resulting in very clumsy phrases that were then restructured into natural academic Arabic during post-editing. Similarly, in Article 2, several instances of over-literal translation of legal and institutional expressions needed restructuring.

Secondly, additions and accuracy errors constitute another important group of errors. The problem of additions was especially prominent in Article 3, as the model sometimes generated interpretive phrases that had no equivalents in the source text. Such additions made the texts more explicit but also altered the translation's informational content, thus representing MQM accuracy errors. In turn, accuracy errors were mostly due to inappropriate renderings of specialised concepts, inappropriate lexical choices, and occasional meaning changes that required correction during post-editing. For example, some terms in the legal and AI domains were translated using approximate equivalents, and some segments

included additional explanatory additions that were subsequently removed from the final translation. Such results show that additions made by ChatGPT Plus may indicate hallucinatory behaviour, and that stylistic errors are the most frequent type of error, proving that semantic adequacy alone is not enough for translating into academic Arabic.

4.4. Technical Post-Editing Effort

The technical post-editing effort was measured by comparing the initial translations produced by ChatGPT Plus with the final human post-edited versions, alongside the COMET-QE semantic assessment and the MQM error analysis. This metric shows how much textual intervention is needed to make a publishable translation. Here, technical post-editing effort was defined as normalised word-level edit distance, computed from substitutions, deletions, and insertions. The number of editing operations was normalised by the final post-edited translation for each segment.

Normalised Edit Rate = (Substitutions + Deletions + Insertions) / Number of words in the post-edited translation

A score of zero means no word-level modification was necessary, and a higher score means greater technical post-editing effort. Hence, the measure represents the extent of textual intervention rather than cognitive load or post-editing time. The results for the three articles are given in Table 5 below.

The results show that at least one word modification was present in 233 segments (67.0%), while 115 segments (33.0%) were unchanged. The mean normalised edit rate across all cases was 0.222, indicating that the ChatGPT Plus output had a measurable amount of human intervention before its final published form. However, there were substantial differences among the three articles.

Table 5. Technical Post-Editing Effort Across the Three Articles

Article	Segments	Unchanged	Changed	Mean Edit Rate	Median
Article 1 (Meribai)	147	68	79	0.132	0.032
Article 2 (Lahcene)	106	14	92	0.340	0.214
Article 3 (Babchikh)	95	33	62	0.230	0.159
Total/Overall	348	115	233	0.222	0.114

Article 2 had the highest technical post-editing effort, with a mean edit rate of 0.340 and 92 of 106 segments edited. Article 3 was second, with a mean edit rate of 0.230 and 62 of 95 segments changed. Article 1 required the least technical intervention, with a much lower mean edit rate of 0.132. Sixty-eight of its 147 segments remained unchanged.

The results show an important gap between automatic estimates of semantic quality and actual intervention in post-editing. Article 2 had a relatively high mean COMET-QE score (0.803) but the highest technical post-editing effort. In contrast, Article 1 was the lowest scored by COMET-QE (0.729) but had the least technical intervention. The MQM findings deliver further evidence of this divergence: Article 2 had the highest average MQM penalty per segment (1.19), followed by Article 3 (0.89) and Article 1 (0.74). Thus, a translation may achieve high semantic adequacy in terms of source-text meaning to obtain a high neural quality estimation score but still needs considerable human editing for terminology, accuracy, style, and target-language conventions.

The results of the technical post-editing further strengthen the argument for a multidimensional evaluation framework. COMET-QE measures estimated semantic adequacy. MQM measures the kind and severity of translation errors. Normalised edit rate measures the amount of textual intervention required to produce the final translation. Together, these three dimensions provide a more holistic evaluation of the quality and pragmatic utility of LLM-generated academic translation than any single metric alone.

5. Discussion

ChatGPT Plus generally preserved semantic content when translating academic texts from English and French into Arabic. More than four out of

five segments fell within the good or excellent COMET-QE bands used in this study. The MQM results, however, revealed numerous problems involving style, accuracy, and additions, showing that a semantically adequate translation may still require substantial revision before publication.

The COMET-QE findings are consistent with previous research showing the effectiveness of neural quality estimation. Rei et al. (2022) demonstrated the effectiveness of COMETKiwi for reference-free quality estimation at sentence and word levels. Similarly, the relatively high COMET-QE scores obtained in the present study suggest that ChatGPT Plus generally preserves the propositional meaning of English and French academic texts translated into Arabic. Nevertheless, variation across articles and segments indicates that translation quality is still inconsistent. The MQM analysis delivers further insight into these quality differences. Although no critical errors were identified, the prevalence of stylistic errors indicates difficulties in conforming to Arabic academic writing conventions. The translations frequently exhibited features of translationese, including literal syntax, unnatural collocations, excessive nominalisation, and source-language interference. These outcomes support Károly's (2022) argument that academic translation requires not only semantic equivalence but also adaptation to the conventions of the target academic community.

Addition and accuracy errors have created another weakness in the LLM-generated translations. This was most evident in Article 3, where ChatGPT Plus occasionally introduced information that was not present in the source text. Some of these additions simply made implicit information more explicit, whereas others changed the information conveyed and affected translation accuracy. This tendency towards unsupported additions has also been associated with overgeneration or hallucination in LLM-generated translation.

The results point to a clear difference between automatic and human evaluation capture. MQM provides information about the type, location, and severity of errors that is not available from aggregate automatic scores (Lommel et al., 2014). This difference was evident in the present study, where some translations received relatively high COMET-QE scores despite stylistic problems, terminology errors, and unsupported additions identified through MQM. A high level of semantic adequacy, therefore, did not necessarily correspond to high overall translation quality.

Some examples of these problems can be cited from the MQM evaluation. One excerpt, for instance, ChatGPT Plus generated a new discussion on the basis of Marshak's three layers of organisational conversation, adding mentions of bias, misapplication of AI, the hierarchy, privilege and community policing – none of which were discussed in the source text. These issues were later addressed the in post-edited version.

Terminological problems were also encountered. For example, the term "Strategic Subject List" was rendered by the AI model as "قائمة الأهداف الاستراتيجية», which shifted the referent from persons to targets, and was corrected to "قائمة الأشخاص الاستراتيجية". In another case, "exponential curve" was rendered as "المنحنى المتسارع" and later post-edited to the more precise "المنحنى الأسي". These examples indicate that even fluent and contextually plausible LLM output may introduce non-existent content or inconsistent terminology, which requires expert post-editing.

Technical post-editing effort provides further evidence of this divergence. Overall, 67.0% of the 348 segments required at least one word-level modification, with a mean normalised edit rate of 0.222, indicating that substantial human intervention remained necessary despite generally high semantic adequacy. Differences across articles were particularly revealing. The three measures did not follow the same pattern across the articles. Article 2, for example, had a relatively high mean COMET-QE score (0.803), yet it also had the highest MQM penalty per segment (1.19) and mean normalised edit rate (0.340). Article 1 showed the opposite pattern: despite having the lowest mean COMET-QE score (0.729), it required the least technical post-editing effort (0.132). This contrast suggests that semantic adequacy, error-based quality, and technical post-editing effort reflect different aspects of translation quality and cannot be used interchangeably.

Overall, the findings support the proposed multidimensional evaluation framework. COMET-QE provides a rapid quantitative estimate of semantic adequacy, MQM identifies the nature and severity of specific translation errors, and normalised edit rate quantifies the extent of textual intervention required to produce the final translation. Their combined use provides a more comprehensive assessment of LLM-generated academic translation than any single measure used independently.

6. Conclusion

This study examined ChatGPT Plus-generated Arabic translations of academic texts from three different perspectives: semantic adequacy using COMET-QE, error type and severity using MQM 2.0, and technical post-editing effort using normalised edit rate. The results show that these measures capture different aspects of translation quality and that none provides a complete assessment when used alone.

For RQ1, the mean COMET-QE score was 0.773, and 82.8% of the segments fell within the good or excellent analytical bands, pointing to generally high semantic adequacy. This level of adequacy, however, did not make the translations publication-ready, as linguistic, stylistic, and terminological revision was still necessary.

For RQ2, the MQM analysis identified 228 errors, most related to style, accuracy, additions, and terminology. Stylistic problems included literal syntax, unnatural collocations, excessive nominalisation, and source-language interference. Article 3 stood out for unsupported additions, some of which may reflect overgeneration or hallucinatory tendencies. No critical

errors were found, but the presence of minor and major errors indicates that expert human review remains necessary.

In terms of technical post-editing effort (RQ3), 233 of the 348 segments (67.0%) required at least one word-level modification, with an overall mean normalised edit rate of 0.222. This indicates that, despite generally high semantic adequacy, substantial textual intervention was required to produce the final translations.

As for RQ4, the three measures did not always converge. Article 2 achieved a relatively high COMET-QE score (0.803) but recorded the highest MQM penalty per segment (1.19) and the highest mean normalised edit rate (0.340). Conversely, Article 1 had the lowest COMET-QE score (0.729) but required the least technical post-editing effort (0.132). This divergence demonstrates that semantic adequacy, error-based quality, and technical post-editing effort capture distinct but complementary dimensions of translation quality, supporting their combined use within a multidimensional evaluation framework.

Overall, ChatGPT Plus can serve as a useful starting point for translating academic texts into Arabic, as its output generally preserves source-text meaning. Human revision remains necessary, however, to address *translationese*, accuracy errors, unsupported additions, and problems of terminology and style. This makes the model better suited to a human-centred workflow in which its output is evaluated and post-edited by experts before publication.

The study is limited by its small corpus of three articles from a specialised domain and its focus on a single LLM. Moreover, normalised edit rate measures only technical post-editing effort, not its temporal or cognitive

dimensions. Future research should apply the framework to larger and more diverse corpora, language pairs, and LLMs, while incorporating post-editing time and cognitive-effort measures.

The findings suggest that users of LLMs for academic translation should not rely solely on automatic quality-estimation scores to determine publication readiness. Combining neural quality estimation, human error analysis, and technical post-editing effort provides a more comprehensive basis for evaluating LLM-generated translations. For Arabic academic translation in particular, human expertise remains essential to ensure linguistic naturalness, terminological precision, and conformity with target-language academic conventions.

References

Al Swuee, A. A. N., & Babchikh, M. (2024). Assessing METEOR and MQM effectiveness in measuring the quality of machine legal translation from English into Arabic: A corpus-based study. *Al-Mutarğim, 24* (2), 12–74.

Alvarez-Vidal, S., & Oliver, A. (2023). Assessing MT with measures of PE effort. *Ampersand, 11*, 100125. https://doi.org/10.1016/j.amper.2023.100125

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-)evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 136–158). Association for Computational Linguistics.

Castilho, S. (2020). Document-level machine translation evaluation project: Methodology, effort and inter-annotator agreement. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 443–444). European Association for Machine Translation.

Chimoto, E. A., & Bassett, B. A. (2022). COMET-QE and active learning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 4735–4740). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.348

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Cui, Y., Liu, X., & Cheng, Y. (2023). A comparative study on the effort of human translation and post-editing in relation to text types: An eye-tracking and key-logging experiment. *SAGE Open, 13*(1).

https://doi.org/10.1177/21582440231155849

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics.

https://doi.org/10.18653/v1/N19-1423

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Falcão, J., Borg, C., Aranberri, N., & Abela, K. (2024). COMET for low-resource machine translation evaluation: A case study of English–Maltese and Spanish–Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3553–3565). ELRA Language Resource Association.

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F. T., Neubig, G., Garg, A., Clark, J. H., Freitag, M., & Firat, O. (2023). The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 1066-1083). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2023.wmt-1.100

Freitag, M., Rei, R., Mathur, N., Lo, C.-K., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A. F. T. (2022). Results of the WMT22 metrics shared task: Stop using BLEU-Neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation* (pp. 46–68). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.2

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 33-41). Association for Computational Linguistics.

Graham, Y., Mathur, N., & Baldwin, T. (2015). Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1183–1191). Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1124

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. T. (2024). xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics, 12*, 979–995. https://doi.org/10.1162/tacl_a_00683

Han, L., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Marazzato Sparano, R., Foresi, M., Innis, J., & Nenadić, G. (2024). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. *arXiv*. https://doi.org/10.48550/arXiv.2405.16969

Károly, K. (2022). Translating academic texts. In K. Malmkjær (Ed.), *The Cambridge handbook of translation* (pp. 262–278). Cambridge University Press.

Karpinska, M., & Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 419–451). Association for Computational Linguistics.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., Shmatova, M., Steingrímsson, S., & Zouhar, V. (2024). Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation* (pp. 1–46). Association for Computational Linguistics.

Kumar, S., Jyothi, P., & Bhattacharyya, P. (2026). *Evaluating extremely low-resource machine translation: A comparative study of ChrF++ and BLEU metrics*. *arXiv*. https://doi.org/10.48550/arXiv.2602.17425

Li, Y., Zhang, R., Wang, Z., Zhang, H., Cui, L., Yin, Y., Xiao, T., & Zhang, Y. (2025). Lost in literalism: How supervised training shapes translationese in LLMs. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12875–12894). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.630

Lommel, A. R., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica, 12*, 455-463.

Marmonier, M., Sagot, B., & Bawden, R. (2026). *Hindsight quality prediction experiments in multi-candidate human-post-edited machine translation*. arXiv preprint arXiv:2603.04083.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685-2702). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2020.emnlp-main.213

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J. G. C., Glushkova, T., Alves, D., Coheur, L., Lavie, A., & Martins, A. F. T. (2022). COMETKIWI: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT 2022)* (pp. 634–645). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.60

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*

(AMTA 2006) (pp. 223-231). Association for Machine Translation in the Americas.

Specia, L., Blain, F., Logacheva, V., Astudillo, R. F., & Martins, A. F. T. (2018). Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 689–709). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6451

Vernikos, G., Thompson, B., Mathur, P., & Federico, M. (2022). Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 118–128). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.6

Vernikos, G., & Popescu-Belis, A. (2024). Don't rank, combine! Combining machine translation hypotheses using quality estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12087–12105). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.653

White, J. S., O'Connell, T. A., & O'Mara, F. J. (1994). The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas* (pp. 193-205).

Zhang, R., Zhao, W., & Eger, S. (2025). How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 10961–10988). Association for Computational Linguistics.

Appendix

Appendix 1: MQM 2.0 Scorecard

MQM 2.0 Dual Evaluation Scorecard: Raw and Linear Calibrated								
Client	**Project Name**		**Translator**		**Evaluator**			
@Client Name	MQM 2.0 Analysis							
	Source Language	**Target Language**		**Total Word Count**	**Evaluation Word Count (EWC)**	**Content Type**		
	Eng	Ara		3000	1500	General, legal, business, media		
Score Calculation Parameters	**Maximum Score Value (MSV)**	**Passing Threshold Raw QS**	**Passing Threshold Calibrated QS**		**Reference Word Count (RWC)**	**Acceptable Penalty Points/RWC**		
	100	99	90		1000	10		
Evaluation Results	**Quality Rating (QR)**	**Raw Quality Score (RQS)**	**Absolute Penalty Total (APT)**	**Per-Word Penalty Total (PWPT)**	**Normed Penalty Total (NPT)**	**Calibrated Quality Score (CQS)**	**Calibrated Quality Rating (CQR)**	
	Pass	99,20	12	0,008	8,00	92,00	Pass	
General Comment								
Note to the LQA evaluator: Please provide a detailed comment on quality of the evaluated sample.								
Severity Penalty Multiplier	0	1	5	25				
Error Type	**Neutral Errors Count**	**Minor Errors Count**	**Major Errors Count**	**Critical Errors Count**	**Error Type Penalty Total (ETPT)**	**Weighted ET Penalty Points**	**Normed ET Penalty Points**	**Error Type Weight**
Terminology	0	0	0	0	0	0	0,00	1
Inconsistent use of terminology	0	0	1	0	5	5	3,33	1
Wrong term	0	0	0	0	0	0	0,00	1
Accuracy	0	0	0	0	0	0	0,00	1
Mistranslation	0	0	0	0	0	0	0,00	1
Overtranslation	0	1	1	0	6	6	4,00	1
Undertranslation	0	0	0	0	0	0	0,00	1
Addition	0	0	0	0	0	0	0,00	1
Omission	0	1	0	0	1	1	0,67	1
Untranslated	0	0	0	0	0	0	0,00	1
Linguistic Conventions	0	0	0	0	0	0	0,00	1
Grammar	0	0	0	0	0	0	0,00	1
Punctuation	0	0	0	0	0	0	0,00	1
Spelling	0	0	0	0	0	0	0,00	1
Unclear Reference	0	0	0	0	0	0	0,00	1
Grand Total	0	2	2	0	12	12	8	

When Arabic Meets AI:

Status Quo and Translation Challenges

Rami Bououden

Algerian Academy of the Arabic Language

Abstract

The past few years have witnessed an unprecedented surge in the use of Artificial Intelligence (AI) across almost every walk of life, particularly driven by Large Language Models (LLMs). As a result, the use of such cutting-edge technologies has become increasingly prevalent in the language services industry. However, in the context of Arabic and its status in the field of LLMs and AI-assisted translation, the former is still considered a low-resource language in terms of its practical presence in the AI realm, unlike other high-resource languages like English. This paper explores the current status of Arabic in the era of Large Language Models and the challenges associated with English-Arabic AI-assisted translation. A qualitative analysis of ChatGPT's English-Arabic translations of specialized academic texts is also conducted to evaluate its output for translation accuracy, terminological consistency, and translationese. The results reveal that ChatGPT generates semantically sound translations but struggles with recurring issues such as literal rendering, inconsistent terminology, and source-language interference. The study indicates that LLMs are a valuable aid for translation tasks, yet human intervention through adequate

prompting and skillful post-editing is pivotal to ensure that the final product reads like an authentic piece of Arabic writing.

Keywords: Artificial Intelligence, Large Language Models, AI-Assisted Translation, Arabic, ChatGPT.

1- Introduction

At the center of modern natural language processing (NLP) are large language models (LLMs), artificial intelligence (AI) systems trained on massive datasets, which have the ability to interpret, generate, and manipulate human language for several purposes (Arslan & Karaca, 2025). These models have changed how machine translation (MT) is perceived by introducing a more interactive, context-sensitive, and prompt-based translation environment. Concerning English-Arabic translation practices, AI-powered tools and LLMs can outperform conventional neural machine translation (NMT) when used efficiently, yet they still struggle in areas such as stylistic precision, concision, and contextual judgment of professional human translators.

The situation of Arabic is more complex than it seems in this context. Despite being spoken by hundreds of millions of people, it is still underrepresented in the realm of LLMs when compared with high-resource languages such as English. This issue can be attributed to the underutilization of high-quality training datasets and reliable benchmarks. Another challenge lies in the fact that Arabic is marked by rich morphology, orthographic variation, and a highly productive lexical system. Such features complicate computational processing and influence the quality of

the AI output, especially when translating specialized academic texts.

This paper delves into the current state of English–Arabic translation in the age of LLMs. It tackles the recent advancements in the field of LLMs in its theoretical part and sheds light on ChatGPT's performance when translating specialized academic texts into Arabic in the practical one, giving attention to translation accuracy, terminology, and the issue of translationese. Thus, the research tries to answer these two questions:

- What is the current status of Arabic in the era of LLMs, and what are the current underlying AI-assisted English–Arabic translation challenges?
- How does ChatGPT perform in English–Arabic academic translation in terms of translation accuracy, terminological correctness, and translationese issues?

2- Literature Rieview

2-1 .General Trends and Changes in the Field

Ever since it has emerged, machine MT went through several phases before reaching its current state. Early MT systems used rule-based systems that relied on linguistic rules to generate translations. Then, statistical machine translation (SMT), which functioned by exploiting bilingual corpora to spot probabilistic correspondences between source and target languages, surpassed the aforementioned systems. However, Elhamayed and Nour (2025) state that reliance on large corpora and the inability to adequately represent long-distance contextual relationships limited SMT's performance, especially when dealing with morphologically rich languages

such as Arabic. One of the major turning points in this field was the advent of NMT, which utilizes deep neural networks that provide more fluent translations since they are able to learn more complex linguistic structures.

More recently, breakthroughs in AI have accelerated these changes through the introduction of LLMs. This was particularly apparent after the release of ChatGPT in 2022, as LLMs gained unprecedented recognition after achieving state-of-the-art performance in various linguistic tasks (Minaee et al., 2024). The majority of LLMs are transformer-based neural networks trained on a myriad of corpora, with hundreds of billions of trainable parameters (Zouidine & Khalil, 2025). Mainstream models, including ChatGPT, Gemini, Claude, and Llama, can utilize contextual information to generate well-structured discourse and adjust their output to fit a variety of communicative settings. According to Abu Guba and Abu Qub'a (2025), these models introduced "a new era in translation technology" (p. 1), as they can process meaning, tone, and contextual information, surpassing mere lexical substitution.

As a result of these changes, translation is no longer regarded as a sheer automated linguistic operation, but rather as a multilayered process that involves human and AI collaboration. Prompt engineering, interactive translation workflows, and AI post-editing significantly altered the balance of the language services industry and expanded the role of language professionals, particularly translators. Well-designed prompts can substantially enhance the quality of LLMs' translations, which means that the outcome is shaped by a balance between human skill and machine

capability (Abu Guba & Abu Qub'a, 2025).

Another important area to cover is the implementation of LLMs in specialized translation tasks, as recent studies evaluated their performance in legal, literary, audiovisual, academic, and administrative translation. For instance, after comparing several LLMs with ordinary neural machine translation (NMT) systems in the translation of Arabic literary texts, Qassem's (2025) study revealed that the former outperformed the latter in preserving rhetorical and stylistic features, yet faced difficulties in rendering cultural and creative elements. Al Sawi and Allam (2024) show that ChatGPT handles subtitling tasks efficiently but struggles with cultural allusions and contextual references. On the other hand, Gemini experienced several difficulties related to specialized terminology, institutional references, and document authentication, which necessitate the adoption of a Human-in-the-Loop approach in critical translation tasks.

2-2 . Arabic in the Age of LLMs

The last four years witnessed a surge in the use of LLMs which transformed from experimental text generators into multimodal reasoning engines capable of executing a plethora of linguistic tasks. These advancements extend beyond high-resource languages and manage to improve the status and representation of low-resource languages such as Arabic. Notwithstanding its rank as one of the most widely spoken languages in the world and one of the six languages of the United Nations, Arabic representation within the AI ecosystem has relatively lagged behind that of other prominent languages including English and Chinese

(Elhamayed & Nour, 2025). This issue clearly depicts how Arabic, a language with a colossal amount of high-quality written content, lacks the necessary computational resources required for training and elevating modern language models.

According to several studies, Arabic is considered as an underrepresented language in contemporary AI despite the many technological leaps achieved. This situation is ascribed to the scarcity of high-quality annotated corpora, benchmark datasets, Arabic-oriented language models, and specialized NLP resources necessary to build reliable LLMs (Abudalfa et al., 2025; Jaleel, 2026; Ouali & El Garouani, 2024). In the same context, Koubaa et al. (2024) clarifies that the predominance of English in multilingual LLMs training and development has significantly reduced the extent to which existing models can effectively represent the rich morphological system, syntactic organization, and semantic complexity of Arabic. Thus, the use of these models as a translation tool, precisely from other languages into Arabic, often came with its flaws.

Acknowledging these limitations, several endeavors have been made in order to mitigate these drawbacks and elevate the position of Arabic. For example, the last few years have witnessed a rise of Arabic-focused and bilingual language models, including Jais, ALLaM, AceGPT, and ArabianGPT. This marked a transition from accommodating Arabic within multilingual models highly optimized for English to building Language models tailored to the unique characteristics of the Arabic language. According to Bari et al. (2024), ALLaM was introduced the linguistic and cultural requirements of Arabic and for the sake of improving bilingual

proficiency in both Arabic and English. In the same vein, Huang et al. (2024) state that AceGPT tackles the linguistic characteristics of Arabic through implementing training strategies dedicated to the language's rich morphology and complex syntax. ArabianGPT is another model that follows a similar approach by using an Arabic-oriented tokenization and architectural enhancements that mitigate reliance on English-oriented representations and enables effective processing of Arabic grammar (Koubaa et al., 2024). Thereby, it can be said that continuous improvements to Arabic-based models could prove more futile than relying solely on LLMs whose core languages are linguistically distinct from Arabic.

Other research trends have recently shed light on many technological directions that may further elevate Arabic in the age of AI. Instead of focusing on scaling model size, researchers are emphasizing several cost-efficient strategies such as parameter-efficient fine-tuning, transfer learning, multilingual adaptation, and enhanced semantic representations as a way of improving Arabic LLMs (Aryan, 2024; Elhamayed & Nour, 2025; Essam et al., 2024). Thus, in the case of languages that lack high-quality training data and domain-specific corpora, the previously mentioned strategies are an effective way to adapt existing multilingual models to Arabic instead of spending more time, effort, and money in training new models from the ground up.

2-3 . Challenges of Arabic AI-Assisted Translation

Remarkable progress has been achieved by both LLMs and Arabic-centered language technologies, yet AI-assisted translation into Arabic still faces several linguistic, terminological, semantic, and technological

challenges. According to Al-Khalifa et al. (2024), even in state-of-the-art pretrained language models, systematic errors like lexical selection, grammatical agreement, functional word usage, and word order persist, suggesting that the linguistic complexities of Arabic are still a major issue in the architecture of these models. NMT systems also struggle when attempting to achieve linguistic accuracy and cultural fidelity in English–Arabic translation tasks (Al Maaytah, 2026).

Generally speaking, LLMs are capable of generating correct and readable Arabic, but their performance fluctuates when it comes to terminological precision. Although ChatGPT delivers reliable translations of general texts, its accuracy decreases when dealing with specialized terminology and domain-specific discourse (Abu Guba & Abu Qub'a, 2025). In addition, Haoues's (2026) evaluation of Gemini's translation of official Arabic administrative documents identified difficulties when translating institutional terminology and certain elements adherent to official Arabic administrative documents. In legal translation, advanced LLMs like ChatGPT and Gemini struggle with context-sensitive legal terminology and legislative concepts tied to a specific culture.

LLMs also struggle when attempting to deal with contextual meaning. This includes difficulty in translating cultural references, idiomatic expressions, and implicit meanings. ChatGPT's performance declines noticeably when working on culture-specific expressions (Abu Guba & Abu Qub'a, 2025), while it resorts to literal translation to render expressions loaded with cultural allusions in AI-generated subtitles, despite

the fact that the output is grammatically acceptable. Furthermore, interpreting contextual elements is considered to be one of the most difficult aspects of Arabic translation, as Mohammed (2025) notes that both conventional MT systems and LLMs still face challenges when confronted with idiomatic expressions, proverbs, and figurative language.

The next issue related to this limited performance is the quality of data used to train these modern language models. The latter is often trained with translated low-quality Arabic corpora marked by different linguistic inconsistencies that affect the quality of the output (Boughorbel et al., 2024). This issue is particularly apparent in specialized translation tasks since the scarcity of authentic, high-quality parallel corpora puts limitations on the performance of Arabic LLMs (AlAfnan, 2024). Consequently, the issue of low-quality Arabic corpora can be mitigated by opting for more reliable and authentic Arabic resources, which are available in abundance.

Just like human translation, LLMs also contain many issues associated with translationese and style. "Translationese" is a term coined by Gellerstam (1986). Since its introduction, it has come to encompass any linguistic differences distinguishing translated texts from their original counterparts. In many cases, this issue occurs when translated texts preserve a set of the source language's features (eg., structural, rhetorical...) instead of adhering to the linguistic and stylistic conventions of the target language. Baker et al. (1993, as cited in Graham et al., 2020) state that translated texts "can be more explicit than the source, less ambiguous, simplified (lexically, syntactically and stylistically); display a preference for

conventional grammaticality; avoid repetition; exaggerate target language features; as well as display features of the source language" (p. 1). These characteristics are usually present in translated texts (Graham et al., 2020). Nevertheless, existing research has extensively examined the issue of translationese in NMT systems, yet few studies have investigated this phenomenon in LLM translation (Li et al., 2025), particularly in English-Arabic translation.

These examples illustrate some of the major challenges that LLMs encounter when translating Arabic, and their limitations may extend beyond what has been discussed above. The practical part of this paper will investigate relevant challenges that may arise in ChatGPT-generated English-to-Arabic translations of academic research papers.

3-Methodology

The study follows a qualitative descriptive research approach to examine the performance of ChatGPT in English-Arabic Academic translation, particularly in the field of marketing. The corpus of this practical part is composed of five peer-reviewed English research articles that were translated into Arabic using ChatGPT version 5.5 Plus. The selected texts are specialized academic discourse and contain a variety of linguistic, terminological, and stylistic features. They are a suitable way to evaluate the performance of this LLM.

All five articles were translated using the same prompt. It was used to instruct ChatGPT to generate a faithful Arabic translation that preserves the original meaning and academic register, avoids omissions and unjustified

additions, and employs natural, idiomatic Arabic. There was also an emphasis on using standardized terminology and stylistic fluency suitable for academic writing.

The generated translations were qualitatively analyzed and compared to the source texts. The analysis focused on spotting recurring translation patterns and examples that demonstrate the strengths and limitations of ChatGPT's output. It is worth mentioning that the translations were assessed based on three main factors:

- **Translation Accuracy**: examined the semantic fidelity and the preservation of the source message.
- **Terminological Consistency and Accuracy:** assessed the translation of specialized terminology and the consistency of lexical choices.
- **Translationese:** investigated the degree to which the output preserves the English syntactic, phraseological, and stylistic elements rather than conforming to Arabic writing conventions.

These analytical criteria provide an adequate qualitative assessment of ChatGPT's performance by identifying key challenges associated with AI-assisted English–Arabic academic translation.

4 -Results

4-1. Translation Accuracy

Table 1. Examples of Translation Accuracy Issues

Source Text	ChatGPT	Suggested Translation
Those data consist of prior viewing history, time of day, location of the	وتشمل تلك البيانات سجلَّ المشاهدة السابق، ووقت اليوم، وموقع العميل،	وتضم هذه البيانات سجل المشاهدات السابقة وأوقاتها وموقع العميل الجغرافي

customer, data from other (similar) users, and other relevant data, which are used to predict what movies a specific customer may wish to watch, whether a show is likely to succeed, and so forth.	والبيانات المستمدة من مستخدمين آخرين ذوي خصائص متشابهة، فضلًا عن غيرها من البيانات ذات الصلة، التي تُستخدم للتنبؤ بالأفلام التي يُحتمل أن يرغب عميلٌ بعينه في مشاهدتها، أو بمدى احتمالية نجاح برنامجٍ أو عرضٍ معين، وما إلى ذلك.	وبيانات مستخدمين آخرين (مماثلين) وغيرها من البيانات ذات الصلة بغية التنبؤ بالأفلام التي قد يود عميل معين مشاهدتها، أو باحتمالية نجاح أحد العروض وهَلُمَّ جَرًّا.
Later, at The Millennium Summit of 2000, most world leaders signed the United Nations Millennium Declaration, in which the Millennium Development Goals were defined [13], serving as a prelude to the current 17 global Sustainable Development Goals...	وفيما بعد، وخلال قمة الألفية سنة 2000، وقع معظم قادة العالم إعلان الأمم المتحدة للألفية، الذي حددت فيه الأهداف الإنمائية للألفية [13]، ليكون تمهيدا للأهداف العالمية السبعة عشر للتنمية المستدامة...	وجاءت فيما بعد قمة الألفية لسنة 2000 أين وقع معظم قادة العالم إعلان الأمم المتحدة والذي أُقرت فيه أهدافها الألفية التنموية [13]، فكان فاتحة لظهور أهداف التنمية المستدامة العالمية السبعة عشر بصيغتها الراهنة...

Table 1 illustrates two examples that contain a number of translation accuracy issues. It can be seen that ChatGPT managed to convey the general informational content of the source segment, yet a closer examination reveals that there are several accuracy-related issues due to a source-language-oriented translation strategy. The model opted for the syntactic organization, punctuation, and information sequencing of the English

language, instead of relying on an approach that follows the conventions of Arabic academic writing. Thus, the final product reads as a direct literal translation rather than a piece of Academic Arabic.

In the first example, the phrase *time of day* was literally rendered as "وقت اليوم". The Arabic translation is linguistically correct, but it fails to capture the intended meaning in this specific context, meaning **the time at which users typically watch content**. Therefore, a more accurate translation could be "أوقات المشاهدة" or "أوقات المشاهدات السابقة" since it conveys the meaning more accurately.

The same example contains issues that depict a tendency toward over-translation. ChatGPT translated the expression *data from other (similar) users* by additional explanatory information not explicitly conveyed in the source text (البيانات المستمدة من مستخدمين آخرين ذوي خصائص متشابهة). This addition increased lexical density without improving comprehension. Likewise, translating *show* as "برنامج أو عرض" constitutes an unjustified addition since the source refers to a single clear concept.

Another important issue to discuss is the textual level, as the Arabic translation provided by ChatGPT follows almost the same English punctuation and listing conventions, specifically the use of the Oxford Comma to separate coordinated elements. This punctuation pattern is normal in English, but Arabic prose rather relies on linking devices and connectors like "و" (and) to achieve coordination and textual cohesion. This obvious imitation of the English punctuation and sentence structure

reinforces the idea that there is a linguistic dominance from the source language of this model.

The second example further shows that ChatGPT tends to preserve the overall structure of the source text at the expense of delivering a translation that adheres to the target language's norms. One particular example concerns the repetition of the word *Millennium*. This word was used three times in the source text in the same sentence (*Millennium Summit*, *Millennium Declaration*, and *Millennium Development Goals*), and ChatGPT repeated it directly (الألفية) in the translated version. This repetition is stylistically acceptable in English, yet Arabic writings in general favor the use of lexical variation, synonymy, textual reference, and reformulation to avoid unnecessary repetition.

The example also contains a minor omission. The adjective *current* is absent in the Arabic translation, despite indicating that the reference is to the goals in their present form, thus removing an element explicitly present in the source text.

The punctuation and sentence organization resemble those of the source text, just like in the first example, rather than reconstructing the translation according to Arabic's textual conventions. In addition, traces of literal translation remain present. This is displayed in the use of "تمهيد" as a translation for *prelude*, which is not correct but lacks a bit of fluency.

The two examples examined in this section demonstrate that ChatGPT generally preserves the semantic content of academic texts with a high degree of fidelity. Nevertheless, this fidelity is achieved through literal

rendering, structural calquing, lexical repetition, and occasional additions or omissions, relying heavily on the source language's criteria. These findings suggest that human post-editing is required to ensure accuracy and adjust the output to achieve target language acceptability and level of readability.

4-2. Terminological Consistency and Accuracy

Source Term	ChatGPT Translation	Suggested Translation
Technologies	تقنيات	تكنولوجيات
Obsolescence	التراجع	التقادم
Nudges	حفزات سلوكية	الدفعات
Salespeople	رجال البيع	مندوبي المبيعات
Experiential Learning	التعلم القائم على الخبرة	التعلم التجريبي
Feedback	التغذية الراجعة	الارتجاع

Terminological accuracy is a crucial component of academic writing and translation. The analysis of ChatGPT's output revealed three types of terminological issues: contextually inaccurate equivalents, the use of descriptive rather than standardized terminology, and the inconsistent rendering of recurring concepts.

First, some equivalents were semantically acceptable but contextually inadequate (Table 2). For instance, the term *technologies* was translated as "تقنيات", whereas the more appropriate equivalent in the context of AI and digital transformation is "تكنولوجيات". In certain cases, the two words might be linguistically correct, but "تقنيات" is generally used as

an equivalent for *techniques* to avoid any possibility of confusion. A similar issue arises in using "التراجع" as an equivalent for o*bsolescence*. While the proposed equivalent conveys the idea of decline, it fails to express the technical notion of becoming outdated due to replacement by newer alternatives. In the area of marketing and technology, "التقادم" serves as a more accurate and established equivalent since it happens when an item, method, skill, or technology is superseded by something newer or more efficient.

Second, ChatGPT also tends to opt for descriptive paraphrasing whenever it fails to find the accurate standardized technical terms. The term *experiential learning* was rendered as "التعلم القائم على الخبرة", which is an explanation instead of an established term. The word "الخبرة" is relatively incorrect since the actual meaning of the word experiential in this context is "التجربة" in Arabic, making "التعلم التجريبي" the correct choice. In addition, the term *salespeople* in marketing is usually translated as "مندوبو المبيعات" in Arabic, but "رجال البيع" is grammatically acceptable but uncommon in contemporary Arabic marketing texts. A mistranslation occurred in the translation of nudges as "حفزات سلوكية", a non-standard expression that lacks lexical legitimacy in Arabic; translating "دفعات" is more accurate and acceptable.

Finally, as mentioned in the previous theoretical sections, ChatGPT and many other LLMs usually rely on low-quality or limited training data or corpora that are obtained from various resources. Many translations in this context are literal calques of foreign terms that have become widely spread

in modern Arabic. The rendering of *feedback* as التغذية الراجعة is an obvious example of this phenomenon. In Arabic, the term has the meaning of "الإغناء" (enriching) or "التعقيب" (making a comment/remark on a previous action or statement). Alternatives such as "الارتجاع" provide a linguistically more economical equivalent, preserve the original concept, and avoid the literal weird translations.

These examples suggest that ChatGPT attempts to identify the conceptual domain of specialized terminology but fails to select the most appropriate standardized Arabic equivalents. Unlike older generations of MT engines, modern LLMs, when prompted effectively, avoid generating outright mistranslations and tend to produce descriptive renderings, literal calques, or contextually acceptable alternatives. Nevertheless, these terms often diverge from the more accurate and standardized terminological choices applied by specialists, thus reinforcing the observations of the previously conducted literature review.

4-3 -Translationese in AI-Assisted Arabic Translation

Besides the issues of accuracy and terminology, the analysis revealed a number of issues associated with translationese and style. The generated translations were grammatically and semantically correct, despite containing some common errors that characterize modern Arabic writings. ChatGPT often retained the syntactic organization, punctuation, and rhetorical progression of the English source, resulting in Arabic translations rendered through an English linguistic and stylistic framework.

One of the most apparent issues in this regard is the preservation of the English sentence structure. Although the prompts contained instructions to avoid imitating the source's style and structure, ChatGPT's output had the same order of utterances as that of the source reorganizing information according to Arabic stylistic conventions. The following sentence is a good example:

S:

"This paper explores the role of circular economy marketing as a catalyst for green innovation, examining emerging trends, challenges, and actionable strategies."

T:

"وتستكشف هذه الورقة دور تسويق الاقتصاد الدائري بوصفه عاملاً محفزاً للابتكار الأخضر، من خلال فحص الاتجاهات الناشئة، والتحديات، والاستراتيجيات القابلة للتطبيق."

It can be seen that the translation is understood and linguistically acceptable, but the sentence mirrors the English sequence almost exactly. In addition, in almost all the translated articles, there is an excessive preservation of lengthy English-style enumerations using the Oxford Comma, which can also be seen in the examples provided in the accuracy section.

Another similar case of translationese can be spotted in the following example:

S:

"In January 2024, surveys were administered to employees in large and medium-sized manufacturing companies in Saudi Arabia, focusing on

industries such as petrochemicals, construction materials, and food processing."

T:

"في يناير 2024، وزعت استبيانات على موظفين في شركات صناعية كبيرة ومتوسطة الحجم في المملكة العربية السعودية، مع التركيز على صناعات مثل البتروكيماويات، ومواد البناء، وتجهيز الأغذية."

The model reproduces the English sentence structure by imitating the original sequencing of information: the temporal setting (في يناير 2024), the research procedure (وزعت استبيانات), the participants (على موظفين), and finally the specification of the targeted industries introduced by "مع التركيز على". In addition, phraseological calque was utilized to render the English expression *focusing on industries such as* (مع التركيز على صناعات مثل), instead of opting for more natural formulations ("وشملت الدراسة صناعات مثل", "ولا سيما صناعات", or "وفي مقدمتها صناعات"). Another use of the Oxford comma and the English listing style appeared in the last part of this sentence, where the comma-separated sequence was retained, and even a comma was inserted before the conjunction "و".

Translationese was also present through ChatGPT's preservation of English phraseological patterns. This includes the literal translation of expressions such as "في هذا السياق" (*in this context*), "يلعب دورا" (*plays a role*), and "على نحو فعال/معتبر/مستمر" as direct counterparts to English adverbial expressions (e.g., effectively, significantly, consistently).

Another noteworthy finding is ChatGPT's ability to avoid several translation patterns that have long been criticized in English–Arabic translation. With precise prompts, the model was able to avoid some

common errors. This is evident in its treatment of passive constructions. Rather than relying on the frequently overused expressions "تم" or "من قبل", ChatGPT employed the Arabic passive voice (المبني للمجهول). For instance, *surveys were administered* was translated as "وُزِّعت الاستبيانات", while the phrase *The researchers collected the data* became "جمع الباحثون البيانات" rather than "قام الباحثون بجمع البيانات", thus resorting to an active voice pattern. The model also avoided the literal translation of the English preposition *by* in passive constructions (من قبل) and reconstructed the following sentence through a relative clause:

S:

"Environmental activities by candidates are a bonus in our recruitment process."

T:

"تعد الأنشطة البيئية التي يمارسها المترشحون ميزة إضافية في عملية التوظيف لدينا."

In the same example, the model opted for "تعد" rather than the nominal phrasing or the overused common error "تعتبر", thereby avoiding another form of literal translation frequently encountered in translations from English and other European languages.

To sum up, ChatGPT's output conveyed the intended meaning of the source text with a high degree of success, demonstrating grammatical and semantic fidelity. This performance is overshadowed by the model's tendency to exhibit characteristics of translationese, including English sentence structures, phraseological patterns, and lexical repetition. Such feature rarely compromises the accuracy of the text's ideas, but they often make the output sound unnatural and reduce the rhetorical sophistication

of the target language. It is also worth mentioning that the model is very responsive to precise prompting, as shown in the last examples. Thus, it is necessary for prompt engineers and translators in general to carefully craft suitable instructions to elevate the translation quality and avoid translationese.

5-Conclusion

The results of this study demonstrate that Arabic is currently in a new era of LLMs. Recent developments in Arabic and multilingual LLMs, benchmarks, and adaptation techniques are gradually elevating their position within the AI ecosystem. However, this language still faces challenges linked to linguistic complexity and limited high-quality training corpora, thus affecting the quality of the Arabic output in specialized academic contexts. Moreover, the qualitative analysis shows that ChatGPT performs effectively when translating from English into Arabic when given suitable prompts. The Arabic output generally preserved the semantic and communicative elements of the source text.

However, several recurring shortcomings related to translation accuracy, terminological consistency, and translationese were noticed. The model frequently relied on literal translation, inappropriate terminological choices, and the direct transfer of English syntactic and phraseological patterns, resulting in an Arabic output that often reads more like a direct translation of English than original Arabic academic writing.

The results also support the idea of viewing ChatGPT as an effective translation assistant instead of a replacement for professional translators, and post-editing should be a necessary and crucial step in order to achieve

the highest translation quality possible. Moreover, ChatGPT should be used as a way to generate first drafts based on well-designed prompts to save time and increase productivity. Human translators should bear the responsibility of refining the output through conducting a terminological check, refining the style, and eliminating traces of translationese.

Future research should focus more on the aforementioned issues and how they are hindering the progress of Arabic translation using LLMs. Translationese is still one of the least-tackled issues in the use of modern LLMs when dealing with Arabic translation from other languages. Another pivotal topic to discuss in future papers is to investigate how incorporating high-quality Arabic corpora obtained from both classical and modern writings affects LLMs' outputs. Incorporating such linguistic and stylistic works may, in fact, enhance these models' command of authentic Arabic discourse and help reduce translationese in AI-generated translations.

References

Abu Guba, M. N., & Abu Quba, A. (2025). AI translation: Evaluating ChatGPT's reliability in translating Arabic to English. *Journal of Artificial Intelligence and Technology, 5*, 551-556.

https://doi.org/10.37965/jait.2025.0831

Abudalfa, S., Saad, M., & El-Beltagy, S. (2025). Editorial: Emerging techniques in Arabic natural language processing. *Frontiers in Artificial Intelligence, 8*, 1-3.

https://doi.org/10.3389/frai.2025.1715520

Al Sawi, I., & Allam, R. (2024). Exploring challenges in audiovisual translation: A comparative analysis of human- and AI-generated Arabic subtitles in *Birdman*. *PLOS ONE, 19*(10), 1-20. https://doi.org/10.1371/journal.pone.0311020

AlAfnan, M. A. (2024). Large language models as computational linguistics tools: A comparative analysis of ChatGPT and Google machine translations. Journal of Artificial Intelligence and Technology, 5, 20-32.

https://doi.org/10.37965/jait.2024.0549

Al-Khalifa, H., Al-Khalefah, K., & Haroon, H. (2024). Error analysis of pretrained language models (PLMs) in English-to-Arabic machine translation. *Human-Centric Intelligent Systems, 4*(3), 206-219. https://doi.org/10.1007/s44230-024-00061-7

Al Maaytah, S. A. (2026). Evaluating three neural machine translation platforms for English-Arabic translation: A comparative study of linguistic accuracy and cultural fidelity. *World Journal of English Language, 16*(2), 1-14. https://doi.org/10.5430/wjel.v16n2p1

Arslan, K., & Karaca, M. F. (2025). The development of large language models from past to present. In V. Sinap (Ed.), *Innovative solutions and contemporary approaches in management information systems III* (pp. 73-101). Özgür Publications.

https://doi.org/10.58830/ozgur.pub1143.c4720

Aryan, P. (2025). Resource-aware Arabic LLM creation: Model adaptation, integration, and multi-domain testing. In *Advanced network technologies and intelligent computing* (Communications in Computer and Information Science, Vol. 2335, pp. 369–383). Springer. https://doi.org/10.1007/978-3-031-83793-7_27

Boughorbel, S., Parvez, M. R., & Hawasly, M. (2024). Improving language models trained on translated data with continual pre-training and dictionary learning analysis. In *Proceedings of the Second Arabic Natural Language Processing Conference* (pp. 73–88). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2024.arabicnlp-1.7

Elhamayed, S., & Nour, M. (2025). Overview of deep learning and large language models in machine translation: A special perspective on the Arabic language. *Journal of Electrical Systems and Information Technology, 12*(1), 1–31. https://doi.org/10.1186/s43067-025-00211-2

Essam, M., Deif, M. A., & Elgohary, R. (2024). Deciphering Arabic question: A dedicated survey on Arabic question analysis methods, challenges, limitations and future pathways. *Artificial Intelligence Review, 57*(10), 1–37.

https://doi.org/10.1007/s10462-024-10880-6

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II* (pp. 88-95). CWK Gleerup.

Graham, Y., Haddow, B., & Koehn, P. (2020). *Statistical power and translationese in machine translation evaluation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 72-81). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2020.emnlp-main.6

Haoues, A. (2026). The reliability of large language models (LLMs) in official translation: An analysis of Gemini's capabilities and challenges in translating Arabic administrative documents into English. *In Translation, 13*(1), 149-163.

Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Song, D., Chen, Z., Alharthi, M., An, B., He, J., Liu, Z., Chen, J., Li, J., Wang, B., Zhang, L., Sun, R., Wan, X., Li, H., & Xu, J. (2024). *AceGPT, localizing large language models in Arabic*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 8139-8163). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2024.naacl-long.450

Jaleel, T. K. (2026). Advanced technologies in Arabic language communication: Challenges and future directions. *ETHOS (Jurnal Penelitian dan Pengabdian), 14*(1), 13-24.

https://doi.org/10.29313/ethos.v14i1.7224

Koubaa, A., Ammar, A., Ghouti, L., Najar, O., & Sibaee, S. (2024). *ArabianGPT: Native Arabic GPT-based large language model*. *arXiv*, 1-21.

https://doi.org/10.48550/arXiv.2402.15313

Li, Y., Zhang, R., Wang, Z., Zhang, H., Cui, L., Yin, Y., Xiao, T., & Zhang, Y. (2025). *Lost in literalism: How supervised training shapes translationese in LLMs*. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers) (pp. 12875–12894). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.630

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large language models: A survey*. *arXiv*, 1-44.

https://doi.org/10.48550/arXiv.2402.06196

Mohammed, T. A. S. (2025). Evaluating translation quality: A qualitative and quantitative assessment of machine and LLM-driven Arabic-English translations. *Information, 16*(6), 1-19.

https://doi.org/10.3390/info16060440

Ouali, S., & El Garouani, S. (2024). Arabic chatbots challenges and solutions: A systematic literature review. *Iraqi Journal for Computer Science and Mathematics, 5*(3), 128-169.

https://doi.org/10.52866/ijcsm.2024.05.03.007

Qassem, M. (2025). Evaluating LLMs versus NMT systems in translating Arabic literary creativity into English. *Dragoman, 21*, 395-420.

https://doi.org/10.63132/ati.2025.transl.6057

Zouidine, M., & Khalil, M. (2025). Large language models for Arabic sentiment analysis and machine translation. *Engineering, Technology & Applied Science Research, 15*(2), 20737–20742. https://doi.org/10.48084/etasr.9584

Assessment of Online Machine Translation Systems for English-Arabic Idiomatic Expressions

Baligh Babaali

University Yahia FARES of Medea

Abstract

The translation of idioms from one language into another while conveying the same conceptualization, connotation, and shades of meaning is one of the most challenging issues in the field of machine translation (MT). As the need for online translation grows, one of the most crucial concerns at present is the accuracy of machine translation systems. The aim of this study was to assess the performance of online translators in translating idioms from English into Arabic. Effectiveness was evaluated and established by analyzing the number and percent of accurate translation with respect to the original idioms. The results of this study show that, in comparison to its competitors, *Microsoft Translator* performs reasonably well when translating English idioms into Arabic by 53.7% of correct translations. According to the findings, it is advised that specialized dictionaries for idiomatic expressions be regularly updated so that MT systems can accurately translate all conceivable idioms.

Keywords

Arabic Language; Idiom; Machine Translation; Neural Machine Translation; Online Translation System;

Introduction

Machine Translation (MT) is a major subfield of Natural Language Processing (NLP) that aims to automatically translate text or speech from one natural language into another (Alawneh & Sembok, 2011). Recent advances in Neural Machine Translation (NMT) have considerably improved translation quality through deep learning models trained on large parallel corpora (Sutskever et al., 2014; Wu et al., 2016; Johnson et al., 2017). Nevertheless, figurative language, particularly idiomatic expressions, remains one of the most difficult challenges for MT systems because idioms convey meanings that cannot generally be inferred from their constituent words (Fillmore et al., 1988; Shojaei, 2012; Alzeebaree, 2020).

Although many studies have evaluated MT quality using automatic metrics and human assessment, relatively few have specifically investigated the translation of English idioms into Arabic (Hampshire & Carmen, 2010). This language pair presents additional challenges due to substantial linguistic and cultural differences.

This study evaluates the performance of eight widely used free online MT systems in translating 201 English idiomatic expressions into Arabic. Translation quality was assessed through expert human evaluation based on semantic fidelity. In addition to measuring translation accuracy, the study analyses common translation errors and statistically compares system performance using McNemar's test, Pearson's Chi-square test, and inter-rater agreement measures (Cohen's and Fleiss' Kappa), providing a

more rigorous evaluation of online MT systems for idiom translation.

The remainder of this paper is organized as follows. Section 1 reviews the related literature, Section 2 describes the methodology, Section 3 presents the results and discussion, and the final section concludes the paper.

1. Literature Background

1-1. Machine Translation

Machine Translation (MT) is the automatic process of translating text or speech from one natural language into another using computational methods (Alawneh & Sembok, 2011). Since its emergence in the 1950s, MT has evolved considerably, progressing from rule-based approaches to statistical models and, more recently, Neural Machine Translation (NMT), which has become the dominant paradigm due to its ability to learn contextual representations from large parallel corpora (Sutskever et al., 2014; Wu et al., 2016; Johnson et al., 2017; Tripathi & Sarkhel, 2025).

MT approaches are commonly classified into three categories: ***rule-based***, ***corpus-based***, and ***hybrid systems***. Rule-based MT relies on handcrafted linguistic rules and bilingual dictionaries, corpus-based MT learns translation patterns from bilingual corpora through statistical or neural models, while hybrid systems combine both approaches to exploit their complementary strengths (Alawneh & Sembok, 2011). The historical evolution of these approaches is commonly represented by the Vauquois Triangle (Figure 1), which illustrates the progression from direct word-for-

word translation to deeper linguistic and semantic representations, culminating in modern neural approaches.

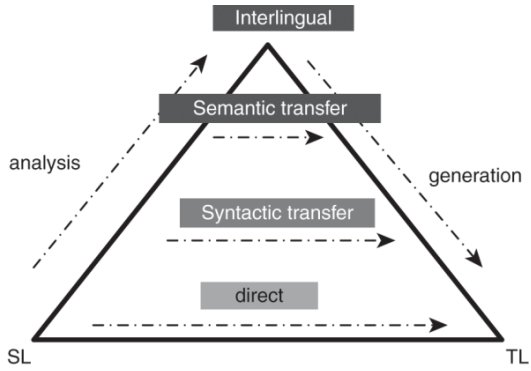

Figure 1. The Vauquois triangle, illustrating the foundations of machine translation.

Among corpus-based approaches, Neural Machine Translation has achieved state-of-the-art performance by modelling entire sentences with deep neural networks, resulting in more fluent and contextually appropriate translations than earlier statistical systems (Wu et al., 2016; Johnson et al., 2017). Consequently, most contemporary online translation services, including *Google Translate* and *Microsoft Translator*, rely primarily on NMT architectures. Nevertheless, translating idiomatic expressions remains challenging because their meanings are often non-compositional and require semantic and cultural interpretation beyond lexical correspondence.

1-2. Idioms

Idioms are fixed or semi-fixed multiword expressions whose meanings cannot generally be inferred from the meanings of their

constituent words (Fillmore et al., 1988; Shojaei, 2012). They constitute an essential component of natural language and reflect the cultural, historical, and social characteristics of a speech community. Because their meanings are figurative rather than literal, idioms present one of the greatest challenges for both human and machine translation.

The translation of idiomatic expressions requires preserving the intended meaning rather than translating individual lexical items. When an equivalent idiom exists in the target language, it should be preferred; otherwise, an accurate paraphrase is generally more appropriate than a literal translation (Grassilli, 2013). Literal rendering frequently produces unnatural or misleading translations, particularly between linguistically and culturally distant languages such as English and Arabic (H. Al-khresheh & Almaaytah, 2018).

Despite the remarkable progress achieved by NMT, idiom translation remains problematic because it requires semantic interpretation, contextual understanding, and cultural knowledge beyond lexical correspondence. Consequently, the ability of MT systems to correctly translate idiomatic expressions provides a rigorous benchmark for evaluating their semantic capabilities. This challenge motivates the comparative evaluation conducted in the present study.

1-3. Translation Evaluation

1-3-1. Human Evaluation

Human evaluation remains the most reliable approach for assessing

machine translation quality, particularly for figurative language, where semantic equivalence is often more important than lexical similarity (Freitag et al., 2021).

Unlike automatic evaluation metrics, human judges can determine whether the intended meaning of an idiomatic expression has been accurately conveyed, even when the translation differs lexically from the source text. Consequently, human evaluation is widely regarded as the gold standard for assessing idiom translation quality (Sepesy Maučec & Donaj, 2020). Accordingly, this study adopts expert human evaluation based on semantic correctness. A translation is considered correct if it accurately preserves the intended meaning of the source idiom, either through an equivalent Arabic idiom or a natural paraphrase that faithfully conveys the original message.

1-3-2. Automatic Evaluation

Automatic evaluation provides a fast, inexpensive, and reproducible means of assessing machine translation quality by comparing system outputs with one or more reference translations (Sepesy Maučec & Donaj, 2020). Traditional lexical-overlap metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), TER (Snover et al., 2006), and chrF++ (Popović, 2017), measure translation quality based on n-gram matching, edit distance, or character-level similarity. More recently, neural evaluation metrics, including BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), and TransQuest (Ranasinghe et al., 2020), have improved automatic evaluation by

incorporating contextual semantic representations.

Despite these advances, automatic metrics remain less suitable for evaluating idiomatic expressions because a correct translation may differ substantially from the reference wording while preserving the intended meaning. This limitation is particularly evident in English–Arabic idiom translation, where multiple valid idiomatic equivalents or natural paraphrases may exist. Consequently, lexical similarity alone cannot reliably reflect translation quality for isolated idioms (Freitag et al., 2021; Zhao et al., 2020).

For these reasons, the present study relies on expert human evaluation as the primary assessment method, using automatic statistical analyses to compare the performance of the evaluated MT systems rather than automatic MT evaluation metrics.

1-4. Previous Research

The evaluation of machine translation systems for idiomatic expressions has attracted increasing attention because idioms remain among the most challenging linguistic phenomena for automatic translation. Several studies have shown that online MT systems frequently produce literal translations that fail to preserve the intended figurative meaning, particularly when translating between linguistically and culturally distant languages (Shojaei, 2012; Taleghani & Pazouki, 2018). Although recent advances in Neural Machine Translation have improved the handling of contextual information, idiom translation continues to pose significant challenges.

Previous comparative studies have assessed the performance of online MT systems using different language pairs and evaluation methodologies. Hampshire and Salvia (2010) compared multiple online translators using human evaluation, while H. Al-khresheh and Almaaytah (2018) investigated the translation of English phrasal verbs and idioms into Arabic, reporting persistent semantic and cultural errors. More recent research has also highlighted the limitations of automatic evaluation metrics when assessing figurative language, emphasizing the importance of expert human judgment for measuring semantic equivalence (Freitag et al., 2021; Sepesy Maučec & Donaj, 2020; Ali, 2020).

Despite these contributions, few studies have provided a comprehensive comparison of contemporary free online MT systems for English–Arabic idiom translation using rigorous statistical analyses. Most previous work either evaluated a limited number of systems, focused on different language pairs, or relied primarily on descriptive analyses. The present study addresses this gap by comparing eight widely used online MT systems using a manually compiled dataset of English idioms, expert human evaluation, inter-rater agreement analysis (Cohen's and Fleiss' Kappa), and statistical significance testing (McNemar's and Pearson's Chi-square tests).

1-5. Research Questions

Based on the identified research gap, this study addresses the following research questions:

- **RQ1:** Which free online machine translation system provides the most

accurate English-to-Arabic translation of idiomatic expressions?

- **RQ2:** What types of translation errors are most frequently produced by the evaluated MT systems when translating English idioms into Arabic?
- **RQ3:** Are the performance differences among the evaluated MT systems statistically significant?
- **RQ4:** To what extent do expert evaluators agree in assessing the quality of idiom translations?

2. Methods

2-1. Choice of Online Translator

An initial survey identified eleven widely used free online machine translation systems, as shown in Table 2. Since the objective of this study is to evaluate the translation of English idiomatic expressions into Arabic, only systems providing English–Arabic translation were eligible for inclusion. Consequently, eight online translators supporting Arabic were selected for the experimental evaluation, whereas the remaining three systems, which do not support Arabic translation, were excluded. The selected systems are highlighted in Table 2.

To ensure a fair and reproducible comparison, all translations were generated using the default settings of each selected system without manual intervention or post-editing. Each English idiomatic expression was translated independently under identical experimental conditions, ensuring that any observed differences in translation quality resulted solely from the performance of the evaluated MT systems.

Table 2. Online machine translation systems considered for evaluation.

Online Translation System	Translation method	# of supported languages	Selected
DeepL[1]	NMT	28	No
Google Translate[2]	GNMT[3]	133	Yes
GramTrans[4]	RBMT	8	No
IBM Watson[5]	NMT	57	Yes
Microsoft Translator[6]	Syntax-based SMT, Phrase-based SMT, NMT	110	Yes
NiuTrans[7]	SMT, NMT	304	Yes
online-translator.eu[8]	NMT	31	Yes
PROMT[9]	Hybrid, RBMT, SMT, NMT	21	Yes
Reverso[10]	NMT	16	No
SYSTRAN Translate[11]	RBMT, SMT, NMT	50	Yes
Yandex Translate[12]	SMT, NMT	98	Yes

[1] https://www.deepl.com/en/blog/how-does-deepl-work
[2] https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html
[3] Google Neural Machine Translation.
[4] https://gramtrans.com/2008/05/29/norwegian-updated/
[5] https://cloud.ibm.com/catalog/services/language-translator#about
[6] https://www.microsoft.com/en-us/translator/business/machine-translation/
[7] https://github.com/NiuTrans
[8] https://www.online-translator.eu
[9] https://www.promt.com/company/technology/neural-machine-translation/
[10] https://www.corporate-translation.reverso.com/translation-technologies
[11] https://www.systransoft.com/systran/translation-technology/
[12] https://yandex.com/company/technologies/translation/

2-2. Choice of Source Text

Sources that contain a variety of idioms in English were required because the paper's main objective was to compare the effectiveness of online translators in translating idioms. Thus, many idiom sources were used to gather the English idioms, including "*Cambridge Phrasal Verbs in Use*", "*Oxford Word Skills: Idioms and Phrasal Verbs (Intermediate)*", "*McGraw Hill's Dictionary of American Idioms*" and "*Oliveboard's 100 Idioms You Must Know for SSC CGL*". The following sources were used to gather the Arabic equivalent idioms such: *A Dictionary of Idiomatic Expressions in Written Arabic*, *Dictionary of Idioms in Contemporary Arabic*, *A Comparison of English and Arabic Idiomatic Expressions Related to Animal Behavior*, with a Brief Discussion of Translation. The texts chosen are only isolated English idiomatic expressions (not in context) with correct Arabic equivalents.

To promote transparency and reproducibility, the complete dataset of English idioms and their reference Arabic translations is publicly available through the GitHub repository provided in Appendix A.

2-3. Evaluation of Machine Translation Outputs

A total of 201 English idioms were selected for this evaluation. Because automatic metrics like BLEU, METEOR, and COMET are not well-equipped to handle isolated idiomatic expressions, we opted for manual human assessment instead.

To start, we consulted specialized idiom dictionaries to pin down the exact meaning of each idiom and its commonly accepted Arabic

counterpart(s). The translations generated by each online MT system were then reviewed independently by three professional Arabic-English translators. We marked a translation as correct if it faithfully captured the intended figurative sense, whether by offering a matching Arabic idiom or a natural paraphrase that made sense in context. Literal word-for-word renditions or translations that lost the original meaning were deemed incorrect.

To check the consistency of our evaluations, we calculated inter-rater reliability using Fleiss' Kappa, which gave us a score of κ=0.86, a level generally considered to reflect almost perfect agreement. On the rare occasions where the reviewers initially disagreed, they talked through their perspectives until they reached a shared consensus. Finally, we counted the number of correctly translated idioms for each system using equation (1), and calculated the percentage of correct translations to compare their performance.

$$Accuracy\ (\%) = \frac{\text{Number of Correct Idiom Translations}}{\text{Total Number of Idioms}} \times 100 \quad (1)$$

2-4. Statistical Analysis

To evaluate the performance of eight machine translation systems on 201 English idioms, various statistical methods were used to ensure a robust comparison. The data collected, being paired categorical, allowed for the application of agreement measures and hypothesis tests to assess translation reliability and identify performance differences.

2-4-1. Inter-rater Reliability

The consistency of the three professional evaluators was measured using Fleiss' Kappa (κ), a statistical coefficient that evaluates the level of agreement among multiple raters while accounting for chance agreement (Landis and Koch, 1977):

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (2)$$

Fleiss' Kappa was utilized in this study to assess the reliability of manual evaluations by three human judges. According to Landis and Koch (1977), κ values below 0 indicate poor agreement, while values between 0.81 and 1.00 suggest almost perfect agreement; $\bar{P}$ represents observed agreement and $\bar{P}_e$ indicates expected agreement by chance.

2-4-2. Pairwise System Comparison

McNemar's test was used to evaluate whether *Microsoft Translator* significantly outperformed competing MT systems. This test is suited for paired binary outcomes and focuses on discordant observations between the classifiers assessed on the same instances:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \quad (3)$$

where b and c denote the numbers of idioms correctly translated by one system but not by the other. McNemar's test, using a continuity-corrected version for conservative estimation with moderate sample sizes, assessed the number of idioms translated correctly by two systems. Statistical significance was defined at α=0.05.

2-4-3. Overall Performance Comparison

Pearson's Chi-square test of independence was used alongside pairwise comparisons to assess whether there were significant differences in translation correctness across eight MT systems:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (4)$$

This test assesses the independence of the observed distribution of correct and incorrect translations from the translation system, comparing observed (O) and expected (E) frequencies under the null hypothesis of equal performance.

2-4-4. Agreement Between Translation Systems

Cohen's Kappa (κ) was computed to assess the level of agreement between *Microsoft Translator* and other machine translation (MT) systems, correcting for chance agreement unlike simple percentage agreement:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (5)$$

where P_o is the observed agreement and P_e is the expected agreement by chance.

Finally, Fleiss' κ was computed across eight MT systems to assess overall agreement on the accurate translation of idioms. This analysis reveals whether translation systems consistently succeed with the same idiomatic expressions or utilize significantly different strategies. All Statistical analyses utilized a significance level of 0.05, interpreting statistical significance and agreement measures together to assess the

comparative performance of the evaluated machine translation systems.

3. Results and Discussion

3-2. Quantitative Analysis

This study examined the performance of eight online MT systems for translating 201 English idioms into Arabic. Table 3 presents the overall performance of each system in terms of accuracy. Results showed significant differences in performance:

Table 3. Overall Performance of the Eight Online Translation Systems (N = 201 idioms)

Online Translation System	Correct Translations	Accuracy (%)	Rank
Microsoft Translator	**108**	**53.7%**	**1**
Google Translate	90	44.8%	2
NiuTrans	61	30.3%	3
PROMT	42	20.9%	4
Yandex Translate	37	18.4%	5
SYSTRAN Translate	28	13.9%	6
IBM Watson	18	9.0%	7
online-translator.eu	9	4.5%	8

As depicted in Figure 2, *Microsoft Translator* achieved the highest accuracy (53.7%, 108 correct translations), followed closely by *Google Translate* (44.8%, 90 correct translations). However, the remaining systems (*NiuTrans*, *PROMT*, *Yandex Translate*, *SYSTRAN Translate*, *IBM Watson*,

and *online-translator.eu*) demonstrated substantially lower accuracy rates, ranging from 30.3% to 4.5%, indicating considerable room for improvement. Notably, the two top-performing systems (*Microsoft* and *Google*) are the only ones that exceeded the 50% threshold, whereas the other six systems failed to correctly translate even one-third of the idioms.

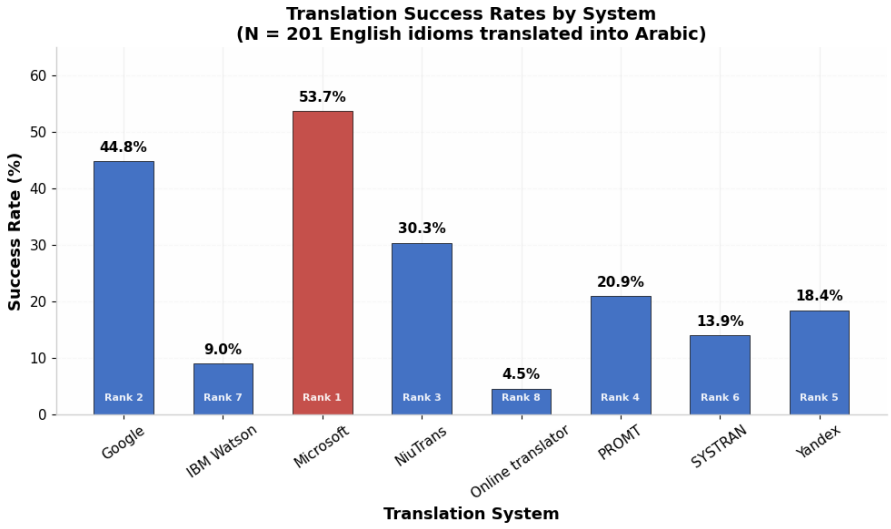

Figure 2. % of Correctly Translated Idioms Produced by Online Translators

To rigorously evaluate whether the observed differences in performance are statistically significant, we performed a series of pairwise McNemar tests with *Microsoft Translator* as the reference system. The McNemar test is the statistical test used to compare pairs of binary proportions (correct vs. incorrect translations) between two dependent samples, as each system has been evaluated on the same set of 201 idioms. The results of the pairwise comparison are presented in Table 4.

The McNemar test results show that *Microsoft Translator* is

significantly better than six of the seven competing systems (*IBM Watson*: χ^2=84.27, p<0.0001; *NiuTrans*: χ^2=24.89, p<0.0001; *online-translator.eu*: χ^2=95.09, p<0.0001; *PROMT*: χ^2=50.30, p<0.0001; *SYSTRAN*: χ^2=69.34, p<0.0001; *Yandex Translate*: χ^2=56.32, p<0.0001). The only exception is *Google Translate*, where the McNemar test gave a marginally non-significant result (χ^2=3.44, p=0.0636), indicating that the difference between *Microsoft Translator* and *Google Translate* is approaching but not reaching conventional statistical significance at α=0.05. That is to say, even though *Microsoft Translator* was more successful (53.7% vs. 44.8% success rate), the 8.9% difference is not statistically conclusive and the two systems can be considered similar in performance with a trend for *Microsoft Translator* to be better. This aligns with the first observation that *Microsoft Translator* performs better than the others, although the difference between *Microsoft* and *Google* is smaller than the difference between *Microsoft* and the other six systems.

Table 4. Pairwise Comparison of *Microsoft Translator* vs. Each of the Seven Other Systems (N = 201).

System	Microsoft Translate [illegible]	Other Correct	Microsoft Translate [illegible] (%)	Other Rate (%)	Agreement (%)	Cohen's κ	McNemar χ^2	McNemar p
Google Translate	108	90	53.7	44.8	58.2	0.171	3.44	0.0636
IBM Watson	108	18	53.7	9.0	53.2	0.119	84.27	<0.0001

NiuTrans	108	61	53.7	30.3	57.7	0.178	24.89	<0.0001
online-translator.eu	108	9	53.7	4.5	49.8	0.059	95.09	<0.0001
PROMT	108	42	53.7	20.9	58.2	0.199	50.30	<0.0001
SYSTRAN	108	28	53.7	13.9	55.2	0.150	69.34	<0.0001
Yandex Translate	108	37	53.7	18.4	56.7	0.173	56.32	<0.0001

Additionally, Pearson chi-square tests of independence were performed for each *Microsoft*-vs.-system pair. The results are consistent with the McNemar results for six systems: *IBM Watson* (χ^2=9.83, p=0.0017), *NiuTrans* (χ^2=8.05, p=0.0045), *online-translator.eu* (χ^2=4.68, p=0.0304), *PROMT* (χ^2=13.18, p=0.0003), *SYSTRAN* (χ^2=10.56, p=0.0012), and *Yandex Translate* (χ^2=11.08, p=0.0009), all of which showed significant associations. Interestingly, the Pearson test for *Google Translate* was also significant (χ^2=6.04, p=0.0140), which means that the association between *Microsoft* and *Google* correctness patterns is statistically detectable, although the McNemar test, which directly compares marginal proportions, is not significant. A Chi-square test across all eight systems overall indicated that the correctness rates were significantly different ($\chi^2(7)$=229.02, p<0.000001).

3-2. Inter-Rater Reliability Analysis

Fleiss' κ was used to measure the level of agreement between the 8 translation systems. Fleiss' κ is the extension of Cohen's κ to the case of multiple raters (here the eight MT systems) who assigning categorical

ratings (correct vs. incorrect) to a common set of items (the 201 idioms). Table 5 summarizes the results.

Table 5. Inter-Rater Reliability Analysis Using Fleiss' Kappa (N = 201 idioms, k = 8 systems).

Measure	Value	Interpretation
Fleiss' κ (8 systems)	0.2144	Fair agreement
Observed Agreement (P_o)	0.7098	71.0% of judgments agree
Expected Agreement (P_e)	0.6307	63.1% agreement by chance
Z-score	93.46	Highly significant (p<0.000001)
Standard Error	0.0023	

The Fleiss' κ value of 0.2144 only indicates '*fair agreement*' between the eight systems (Landis and Koch, 1977). The z-score (93.46) is highly significant (p<0.000001) indicating that the observed agreement is not due to chance, but the magnitude of agreement is remarkably low. This is a very important finding: while the success rates of the individual systems are moderate (especially *Microsoft* and *Google*), there is very little agreement between the systems as to which particular idioms are correctly translated. Thus, the two best systems (*Microsoft* and *Google*) agree on only 58.2% of the idioms and they agree even less with the lower performing systems. This shows that idiom translation is still a very system-dependent task, with each MT platform using different strategies, corpora or neural architectures, which results in different outputs for the same input. The low inter-system agreement underlines the need for specialized idiom corpora and standardized evaluation frameworks in machine translation research.

3-3. Qualitative Error Analysis

3-3-1. Error Classification of the Translation Outputs

An error analysis was performed on the translation outputs of *Microsoft Translator*, *Google Translate*, and *IBM Watson*, categorizing each translated idiom into one of five quality-based categories to complement the quantitative evaluation.

- **Correct Arabic Idiom:** The English idiom was translated using an established Arabic idiomatic equivalent.
- **Correct Paraphrase:** The figurative meaning was accurately conveyed using a natural non-idiomatic Arabic expression.
- **Literal Translation:** The idiom was translated word-for-word, resulting in the loss of its intended figurative meaning.
- **Semantic Error:** The generated translation was grammatically acceptable but conveyed an incorrect or incomplete meaning.
- **No Output/Nonsense:** The system failed to produce a meaningful translation or generated an unintelligible output.

The analysis shows that literal translation is the main error source in English-Arabic MT. *Microsoft Translator* excelled by providing the most correct Arabic idioms and fewer errors. *Google Translate* performed similarly but relied more on literal translations for idiomatic expressions. *IBM Watson* underperformed, primarily due to literal translation issues and failures in producing meaningful outputs. As illustrated in Figure 3, *Microsoft Translator* achieved the highest number of correct idiomatic

translations (38.8%) and acceptable paraphrases (14.9%), while *IBM Watson* produced the largest number of literal translations (74.1%), highlighting its difficulty in recognizing non-compositional expressions.

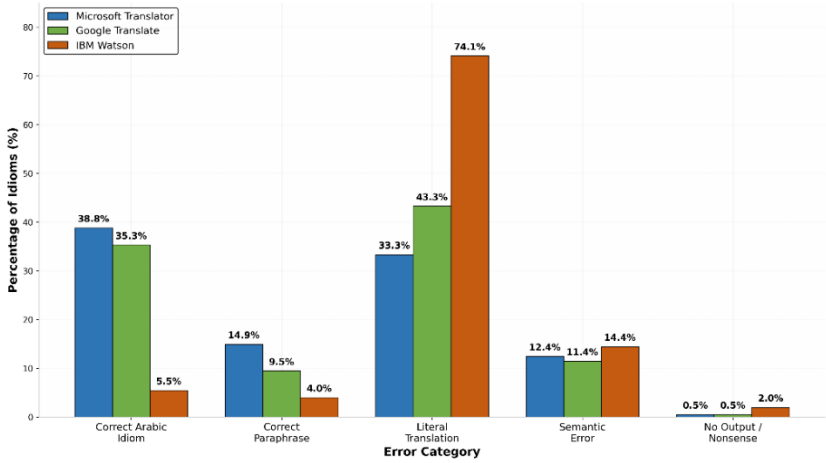

Figure 3. Distribution of Translation Outcomes by Error Category

Semantic preservation in paraphrasing is crucial, especially when English idioms lack direct Arabic equivalents. Accurate translations via natural Arabic paraphrases were deemed correct, emphasizing the need for expert human assessment in idiomatic translations, as automated metrics based on lexical similarity can misjudge semantically equivalent paraphrases.

Overall, error analysis highlights that the main challenge for current MT systems lies in interpreting figurative language rather than generating grammatical Arabic text. Despite advancements in modern neural systems, idiomatic expressions reveal significant limitations in semantic understanding and cultural adaptation.

3-3-2. Representative Translation Examples

In this subsection, we present a sample of the actual outputs produced by the online translation systems on a random sample of the test set. We compare the outputs with the idioms Arabic translation. The objective of this comparison is to demonstrate the effectiveness of each system to produce the accurate output to the original idiomatic expression.

In Example 1, we can see that four (4) online translators correctly translated the idiom, but *NiuTrans* did so in a completely different way that was semantically accurate. the remaining systems provided a word-for-word translation (see Table 6). However, in Example 2, only two (2) online MTs (*Google Translate* and *Microsoft Translator*) returned the idiom's accurate translation, two other systems (*PROMT* and *Yandex Translator*) returned a translation that was extremely similar, and as in the previous example (*NiuTrans*) provided a translation that was accurate on the semantic side but had a completely different wording. The other three systems rendered the idiom exactly (see Table 6).

Finally, only *Microsoft Translate* was able to translate the sentence in example 3 in an idiomatic manner; in contrast, the other MT systems only offered a literal translation that was occasionally syntactically erroneous (as shown in Table 6).

Table 6. Examples of translations of idioms

	Example 1	**Example 2**	**Example 3**
English Idiom	A stone's throw	Add insult to injury	Eager beaver

Arabic Translation	على مرمى حجر	زاد الطين بلة	متحمس للعمل
Google Translate	على مرمى حجر	يزيد الطين بلة	شغوف القندس
IBM Watson	رمية حجارة	إضافة إهانة للإصابة	قندس متلهف
Microsoft Translator	على مرمى حجر	زاد الطين بلة	متحمس للعمل
NiuTrans	بعد سهم	ما يزيد الأمور سوءا	سمور حريصة
online-translator.eu	رمي الحجارة	تضاف الإهانة إلى الأذى	منارة متلهفة
PROMT	على مرمى حجر	أضف الطين بلة	القندس المتحمس
SYSTRAN	رمية حجر	إضافة إهانة إلى الإصابة	قندس حريص
Yandex Translate	على مرمى حجر	أضف الطين بلة	حريصة سمور

3-4. Discussion

The results of this study highlight significant variations in the performance of online machine translation (MT) systems when it comes to translate English idioms into Arabic. *Microsoft Translator* emerged as the top-performing system, achieving a 53.7% accuracy rate, followed by *Google Translate* at 44.8%. The remaining systems (*NiuTrans*, *PROMT*, *Yandex Translate*, *SYSTRAN Translate*, *IBM Watson*, and *online-translator.eu*) demonstrated notably lower accuracy rates, ranging from 30.3% to as low as 9.0%. The McNemar tests confirmed that *Microsoft Translator* significantly outperforms six of the seven competitors, with *Google Translate* being the only system that approached comparable performance (χ^2=3.44, p=0.0636). These findings underscore the challenges that idiomatic expressions pose for MT systems, particularly in language pairs with distinct linguistic and cultural frameworks like English

and Arabic.

3-4-1. Key Observations and Implications

Neural vs. Hybrid Approaches:

The performance of *Microsoft Translator* and *Google Translate* is largely due to their neural machine translation (NMT) architectures, which utilize extensive parallel corpora and contextual learning. *Microsoft*'s hybrid model, integrating syntax-based SMT and NMT, shows slight advantages in idiomatic expression handling over *Google*'s purely neural approach, although the difference (8.9%) is not statistically definitive (p=0.0636). This supports previous findings on the necessity of merging linguistic rules with empirical data for effective idiom translation. Conversely, *SYSTRAN Translate* and *PROMT*, using older rule-based or statistical methods, have more difficulty with idioms. For example, *SYSTRAN* translated "eager beaver" as "حريص قندس" ("eager beaver"), a literal rendering that ignores the idiom's figurative meaning ("متحمس للعمل" or "enthusiastic for work"). This underscores the limitations of non-neural systems in capturing semantic nuances.

Literal vs. Figurative Translation:

Idioms pose challenges for word-for-word translation due to their non-compositional meanings, as emphasized by (Fillmore et al. 1988). Their translation demands a strong grasp of cultural context. Analysis (Table 6) showed that lower-performing systems frequently relied on literal translations, which missed the figurative essence of idioms. For instance:

- "Add insult to injury" was correctly translated by *Microsoft Translator* and *Google Translate* as "زاد الطين بلة" ("add mud to flood"), an Arabic idiom with equivalent connotation.
- Lower-performing systems like *IBM Watson* rendered it literally as "إضافة إهانة إلى الإصابة" ("adding insult to injury"), which lacks cultural resonance.

This aligns with (Grassilli, 2013) taxonomy of idiom translation strategies, where finding culturally equivalent idioms yields the best results but is often unattainable for MT systems without curated idiom databases. The low Fleiss' κ (0.214) further supports this interpretation: the lack of consensus among systems reflects their divergent strategies for handling figurative language, with most systems lacking the cultural knowledge required to identify appropriate Arabic equivalents.

Cultural and Contextual Nuances:

Idioms are deeply tied to culture, making their translation challenging for most MT systems. A study shows that *Microsoft Translator* aligns with *Google Translate* on merely 58.2% of idioms, and its agreement with less effective systems falls to about 50-56%. This suggests that even top MT tools use distinct internal representations or training data for idiomatic expressions. For example:

- "A stone's throw" was correctly translated by *Microsoft Translator* and *Google Translate* as "على مرمى حجر" (an Arabic idiom for proximity).
- *NiuTrans* produced "بعد سهم" ("after an arrow"), a semantically

plausible but culturally inaccurate phrase.

Such errors reflect the systems' reliance on statistical patterns rather than cultural knowledge, a limitation noted by (H. Al-khresheh & Almaaytah, 2018) in their study of proverb translation.

Idiom Variability and Novelty:

Idioms evolve dynamically, with new expressions emerging regularly (Anastasiou, 2010). MT systems trained on static corpora often fail to handle novel or rare idioms. The very low performance of *online-translator.eu* (4.5%) and *IBM Watson* (9.0%) suggests that the training data used to train these systems was not sufficiently idiomatic, or that their architectures are not optimized for figurative language. *NiuTrans*, with 304 language pairs and SMT and NMT, had only 30.3% accuracy, showing that a large number of languages does not mean the system is good at dealing with idioms. The significant McNemar test results (all $p<0.0001$ for comparisons with *Microsoft*) confirm that these performance gaps are not due to sampling variation but represent real systemic deficiencies.

3-4-2. Implications for MT Development

Specialized Idiom Corpora:

The study's results advocate for MT systems to integrate idiom-specific parallel datasets, as proposed by (Moussallem et al., 2019). For instance, one possible explanation for *Microsoft*'s higher accuracy is its access to curated idiomatic expressions, while systems like *Yandex Translate* (18.4% accuracy) likely prioritize general-purpose translation The

differences observed may be due to variations in training data, model architecture, or idiom coverage, although verification is not possible as the systems are proprietary. The development of large-scale, manually verified English-Arabic idiom corpora would benefit all systems, particularly those currently scoring below 20%.

Hybrid and Context-Aware Models:

Combining NMT with rule-based methods could improve idiom handling. For example:

- Rule-based filters could flag potential idioms and route them to specialized sub-models.
- Context-aware NMT (e.g., *Google*'s Transformer models) might better disambiguate idioms in sentences versus isolation, though the present study focused on isolated expressions to control for contextual variables. Future evaluations should assess whether these leading systems maintain their advantage when idioms are embedded in full sentences and paragraphs.

User Feedback Mechanisms:

Allowing users to flag incorrect translations (e.g., via crowdsourcing) could create iterative improvement loops, as seen in Wikipedia's collaborative editing. Given the low inter-system agreement (Fleiss' κ=0.214), user feedback could help identify idioms that are systematically mistranslated across multiple platforms, enabling targeted retraining.

3-4-3. Limitations and Future Directions

Corpus Limitations:

The study's reliance on a fixed set of 201 idioms may not fully represent the diversity of idiomatic expressions. Future work could expand the dataset to include dynamic, contextually embedded idioms and evaluate MT systems on:

- Contextualized idioms (e.g., "He's a real eager beaver at work").
- Domain-specific idioms (e.g., medical or legal jargon).

Additionally, the inclusion of phrasal verbs and collocations would provide a more comprehensive assessment of figurative language translation.

Statistical Power and Effect Sizes:

While the McNemar tests and chi-square analyses establish statistical significance for six of seven comparisons, the marginal result for *Google Translate* (p=0.0636) warrants caution. Future work with larger samples of idioms might establish whether the *Microsoft-Google* gap is robust or an artifact of sample size. In this study, Cohen's κ is reported for pairwise agreement. The Cohen's κ values range from 0.059 (*Microsoft* vs. *online-translator.eu*) to 0.199 (*Microsoft* vs. *PROMT*). The Cohen's κ values suggest weak agreement over chance.

Human Evaluation Bias:

While manual evaluation ensured precision, it introduced

subjectivity. Combining human judgment with automated metrics like BLEU (for fluency) or METEOR (for semantic similarity) could enhance robustness.

Real-World Context:

Idioms in isolation (as tested here) may not reflect their usage in sentences, paragraphs, or discourse. Testing MT systems on idiomatic expressions within full sentences or paragraphs could yield more practical insights, as surrounding text often disambiguates figurative meaning. However, the isolation was intended to control for contextual cues and isolate the systems' ability to recognize and translate idioms per se, which is the main objective of this study.

Cross-Linguistic Generalizability:

The study focused on English-Arabic, but idioms in other languages (e.g., Chinese "画蛇添足" / "Draw legs on a snake" or "守株待兔" / "Guard the stump, wait for the rabbit") may pose distinct challenges. Comparative studies could pinpoint language-specific pitfalls, and establish whether the *Microsoft-Google* dominance we observe here generalizes to other typologically diverse pairs. The low Fleiss' κ (0.214) means the system performance is very sensitive to the language pair and therefore suggests the necessity of pair-specific evaluations instead of generic MT benchmarks.

3-4-4. Recommendations for Practitioners

- **For Users:** *Microsoft Translator* is the statistically best performing of the

eight tested systems for translating idiom-heavy texts from English to Arabic, with *Google Translate* a viable alternative. But users should beware: *Microsoft Translator*, for example, gets about 46% of idioms wrong. For critical applications (e.g. legal, medical or literary translation) the use of human post-editing by a professional translator is strongly recommended.

- **For Researchers:** Future work may involve combining human judgment with automated metrics such as BLEU (for fluency) or METEOR (for semantic similarity), though these metrics are known to work poorly for isolated idioms. The field would benefit from the development of idiom-specific evaluation metrics (possibly using BERTScore or COMET). Further, researchers should explore whether the performance hierarchy shown here generalizes to Arabic-to-English translation, as idiomatic density and directionality might interact with system architecture.

- **For Developers:**

a) Prioritize idiom corpora and cultural metadata in training data.

b) Regularly update MT systems with idiom-specific parallel datasets to improve coverage and accuracy.

c) Explore hybrid approaches that combine NMT with rule-based methods to better handle figurative language.

d) Implement mechanisms for users to flag incorrect idiom translations, enabling iterative improvements.

e) Address the '*fair agreement*' problem (Fleiss' κ=0.214) by standardizing idiom translation strategies across platforms, possibly through shared lexicons or multilingual idiom banks.

Conclusion and Recommendations

This study evaluated the ability of eight free online machine translation systems to translate English idiomatic expressions into Arabic using a dataset of 201 idioms assessed through expert human evaluation. The findings confirm that idiom translation remains a challenging task for current MT systems due to the semantic and cultural complexity of figurative language.

The study successfully addressed the proposed research questions. Regarding **RQ1**, *Microsoft Translator* achieved the highest translation accuracy (53.7%), followed by *Google Translate* (44.8%), while the remaining systems exhibited considerably lower performance. Regarding **RQ2**, the qualitative analysis revealed that the predominant errors resulted from literal translation, failure to recognize idiomatic meaning, and the inability to produce culturally appropriate Arabic equivalents. Regarding **RQ3**, statistical analyses demonstrated that *Microsoft Translator* significantly outperformed six of the seven competing systems (McNemar, p<0.0001), whereas its advantage over *Google Translate* was not statistically significant (p=0.0636). Regarding **RQ4**, the human evaluation process achieved *almost perfect inter-rater agreement* (Cohen's κ=0.86), confirming the reliability of the evaluation methodology.

Although NMT has substantially improved translation quality, the

results indicate that current online MT systems still struggle with figurative language. Furthermore, the relatively low inter-system agreement (Fleiss' κ=0.214) suggests that different platforms employ substantially different strategies when translating idiomatic expressions, resulting in inconsistent outputs.

Future improvements should focus on integrating larger idiom-specific bilingual corpora, incorporating cultural and contextual knowledge into MT models, and developing evaluation frameworks specifically designed for figurative language. Until such advances are achieved, professional human review remains essential for applications requiring accurate translation of English idioms into Arabic.

References

[1] Alawneh, M. F., & Sembok, T. M. (2011). *Rule-Based and Example-Based Machine Translation from English to Arabic*. 6h International Conference on Bio-Inspired Computing: Theories and Applications Rule-Based, 343-347. https://doi.org/10.1109/BIC-TA.2011.76

[2] Ali, M. A. (2020). *Quality and Machine Translation: An Evaluation of Online Machine Translation of English into Arabic Texts*. Open Journal of Modern Linguistics, 10(05), 524-548. https://doi.org/10.4236/ojml.2020.105030

[3] Alzeebaree, Y. (2020). *Machine Translation and Issues of Multiword Units: Idioms and Collocations*. Eastern Journal of Languages, Linguistics and Literatures (EJLLL) 2000, 1-23.

[4] Anastasiou, D. (2010). *Idiom Treatment Experiments in Machine Translation*. In Newcastle upon Tyne: Cambridge Scholars Publishing (Issue January 2010). Universität des Saarlandes.

[5] Banerjee, S., & Lavie, A. (2005). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005, 65-72.

[6] Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). *Regularity and idiomaticity in grammatical constructions: The case of let alone*. The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, Volume II Classic Edition, 64(3), 501-538.https://doi.org/10.2307/414531

[7] 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

[8] Grassilli, C. (2013). *How to translate idioms*.

https://translatorthoughts.com/2013/10/translation-techniques-idioms/

[9] H. Al-khresheh, M., & Almaaytah, S. A. (2018). *English Proverbs into Arabic through Machine Translation*. International Journal of Applied Linguistics and English Literature, 7(5), 158. https://doi.org/10.7575/aiac.ijalel.v.7n.5p.158

[10] Hampshire, S., & Carmen, P. S. (2010). *Translation and the Internet: Evaluating the Quality of Free Online Machine Translators*. Quaderns: Revista de Traducció, 17 SE-Articles. https://raco.cat/index.php/QuadernsTraduccio/article/view/194256

[11] Johnson, M., Schuster, M., Le, Q. V, Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). *Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation*. Transactions of the Association for Computational Linguistics, 5, 339-351. https://doi.org/10.1162/tacl_a_00065

[12] Landis, J. R., & Koch, G. G. (1977). *The measurement of observer agreement for categorical data*. biometrics, 159-174. https://doi.org/10.2307/2529310

[13] Moussallem, D., Wauer, M., & Ngomo, A.-C. N. (2019). *Semantic Web for Machine Translation: Challenges and Directions*. CEUR Workshop Proceedings, 2576, 1-9. http://arxiv.org/abs/1907.10676

[14] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). *BLEU*. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 2002-July(July), 311. https://doi.org/10.3115/1073083.1073135

[15] Popović, M. (2017). *chrF++: words helping character n-grams*. Proceedings of the Second Conference on Machine Translation, 2(1), 612-618. https://doi.org/10.18653/v1/W17-4770

[16] Ranasinghe, T., Orăsan, C., & Mitkov, R. (2020). *TransQuest: Translation Quality*

Estimation with Cross-lingual Transformers. COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, 5070-5081. https://doi.org/10.18653/v1/2020.coling-main.445

[17] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). *COMET: A neural framework for MT evaluation*. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2685-2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

[18] Sellam, T., Das, D., & Parikh, A. P. (2020). *BLEURT: Learning robust metrics for text generation*. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 7881-7892. https://doi.org/10.18653/v1/2020.acl-main.704

[19] Sepesy Maučec, M., & Donaj, G. (2020). *Machine Translation and the Evaluation of Its Quality*. Recent Trends in Computational Intelligence, 1-20. https://doi.org/10.5772/intechopen.89063

[20] Shojaei, A. (2012). *Translation of Idioms and Fixed Expressions: Strategies and Difficulties*. Theory and Practice in Language Studies, 2(6), 1220-1229. https://doi.org/10.4304/tpls.2.6.1220-1229

[21] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A study of translation edit rate with targeted human annotation*. AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation, August, 223-231. https://aclanthology.org/2006.amta-papers.25/

[22] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks*. Advances in Neural Information Processing Systems, 4(January), 3104-3112. https://dl.acm.org/doi/10.5555/2969033.2969173

[23] Taleghani, M., & Pazouki, E. (2018). *Free Online Translators: A Comparative Assessment in Terms of Idioms and Phrasal Verbs*. International Journal of English

Language & Translation Studies, 06(01), 15-19.

http://www.eltsjournal.org/index.html

[24] Tripathi, S., & Sarkhel, J. K. (2025). *Approaches to Machine Translation*. In Translation and Translanguaging in Multilingual Contexts (Vol. 11, Issue 1, pp. 388-393). https://doi.org/10.1075/ttmc.11.1

[25] Wu, Y., Schuster, M., Chen, Z., Le, Q. V, Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 1-23. http://arxiv.org/abs/1609.08144

[26] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT*. 8th International Conference on Learning Representations, ICLR 2020, 1-43. http://arxiv.org/abs/1904.09675

[27] Zhao, W., Glavaš, G., Peyrard, M., Gao, Y., West, R., & Eger, S. (2020). *On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation*. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1656-1671. https://doi.org/10.18653/v1/2020.acl-main.151

Appendix A

The complete dataset is publicly available in the GitHub repository listed below: https://github.com/bbaligh/Idioms-translations/blob/main/Idioms%20Translations.xlsx

The Effect of Digital Technological Shift on Arabic Translation: Advancements, Challenges, and Cultural Implications

Naima Bougherira

Badji Mokhtar-Annaba University-Algeria

Translation and Didactics of Languages Laboratory (TRADIL)

Abstract

Recent breakthroughs in machine translation, particularly in the application of DeepL, are achieving spectacular results in translation. Moving beyond traditional, rule-based or statistical machine translation, current systems based on deepL application are able to process huge datasets by learning to translate at the level of meaning rather than mere word-for-word equivalence. In light of the linguistic peculiarities that characterize the Arabic language on the one hand and English on the other, especially within their respective dialectical variations, it becomes a central research focus how to adapt and continue learning in order to overcome the said complexities and reach more accurate results in automatic translation .Translation from English to Arabic, is no longer confined to mere transmission of information; it has become an indispensable tool for fostering cultural interaction across the Arab world and the English-speaking global community. Arabic language is a rich and complex

language with nuances that extend to the depths of history and culture. Through a descriptive analytical approach, this paper delves into the advent deepL Translation technologies that have dramatically impacted translation services and how they are perceived. Translation services are perceived to be much more accurate and efficient than before. English-Arabic translation, in particular, is quite challenging to process by humans due to the many intricacies embedded in the languages. DeepL incorporates these aspects that are typically hard to translate accurately, thus addressing issues of context, idioms and word order structure which are particularly relevant to this language pair.

Keywords: English-Arabic Translation, Digital Shift, Cultural Nuances, DeepL, Cultural Interaction.

1- INTRODUCTION

Language translation between English and Arabic is a complex task full of nuances and depth that goes far beyond simple word for word translation. Translation, particularly of high volume content requires an individual with an intimate knowledge of the cultures, societies, and histories that have contributed to the languages. Arabic contains a plethora of words and expressions that have deep rooted meanings within the language and culture, often proving difficult to translate adequately into English. Conversely, English is a global language that is widely used as a business language and in many contexts it is necessary to translate to Arabic in an accurate and viable manner.

Cultural nuances in language-distinctions in meaning, context and connotation between languages that are natural to different cultures-play

a critical role in effective English-Arabic translation. Idiomatic expressions and metaphors, references to historical events or figures, and practices or traditions that are rooted in specific cultures all pose translation challenges, particularly those that lack corresponding equivalents in the target language. As a result, translators must produce high-quality translations that are sensitive to these differing cultural norms and allow the original message to be conveyed without sacrificing cultural relevance.

The advent of the digital translation tools has caused a dramatic revolution in the efforts to break down the language barriers. The English language translation tools have also seen impressive improvement in the Arabic language translation. One of the leading machine translation tools, is the DeepL platform. DeepL is a Language AI company headquartered in Cologne, Germany. Founded in 2017 by CEO Jarosław Kutyłowski, the platform originated from the dictionary project Linguee and is known for its highly accurate, human-sounding machine translations. This platform is renowned for its highly accurate, context-aware translations and writing assistance. Using advanced neural networks and proprietary AI models, it is widely used by individuals and global enterprises to translate text, documents, and speech (Faes, 2017). However, despite the significant improvement in translation of language nuances has caused many challenges to the linguistics and the cultural contexts. The current study has focused on the difficulties encountered by the DeepL to translate the cultural nuances in idioms from English to Arabic. Because of significant cultural differences between the two languages of English and Arabic, English language is greatly influenced by Western cultures whereas the

Arabic language is deeply rooted in the Arabic historical and cultural structures and varied regional structures. In today's fast-paced global communication, there is a widespread tendency to overlook the cultural dimension of communication. Simple translations often fail to effectively communicate the desired message and may even lead to miscommunication. The availability of machine translation tools has added a new level of complexity to the process of English-Arabic language translation. Although machine translation tools particularly, DeepL, offer many benefits, including efficiency and speed, they often lack the nuance and contextual awareness that are so important in Arab culture and language. As a result, important nuances and connotations, particularly those that carry emotional weight and deal with sensitive topics, are often lost in translation.

This study examines the translation challenges that arose during the translation of a small group of widely used informal English idiomatic expressions and connotations to Arabic, to discern potential ways in which such highly-complex expressions might be successfully translated and to illuminate the requirements of achieving satisfactory levels of cultural equivalence within machine translation represented in DeepL. Despite recent achievements in the application to English-Arabic translation, several challenges remain to be overcome.

2- Cultural Nuances and Linguistic Peculiarities in English-Arabic Idioms' Translation

Translation into another language involves numerous challenges, of which the main difficulties pertain to the large variety of cultural

references found in different languages. The amount of idiomatic expressions and context-specific meanings can also lead to serious complications in communication when using digital translations to reach an English-speaking audience, as previously mentioned. For instance, the English idiom *I am all ears* is literally translated into Arabic to mean كلي آذان (*kouli adhanun*) (Abu Ssaydeh, 2004). When this is taken figuratively, it means that a person is attentive and cares about you.There is much more meaning behind the simple act of listening attentively to others than what it initially means. This is something that needs to be translated and, as stated earlier, digital translations cannot cope with these types of issues. As a result, more emphasis is being placed on human linguistic and cultural interpretation rather than relying solely on technology.

Elhadary (2023) acknowledges the difficulties translators face in providing accurate translations in a globalized world and argues that translators need to possess not only linguistic competence but also cultural competence and insight into the cultural stories that are embedded in a language and culture. In addition, the rise of the digital media has further complicated the translation process and rendered it more challenging for translators, as most online sources are written in language that is often deeply embedded within a specific culture and context and therefore require not only linguistic skills but also cultural insight. Elhadary also argues that the growing reliance on machine translation will increase cultural dissonance unless human translators intervene to critically evaluate and supplement the translation output.

When expressing yourself in English and then corresponding in

Arabic, you inevitably hit up against idioms and expressions and contexts that need special attention. For effective cross-cultural communication, translators must go to great lengths to ensure the full range of meaning is communicated and that the intent behind the words is preserved. The translation of everyday idiomatic expressions and culture-bound phrases can be particularly problematic in English-Arabic translation, especially given the rapidly increasing role of the digital world in our daily lives and the complexities they bring to language and communication.

Idioms are one of the most troublesome features of the English language for translators, as there is generally no direct equivalent in Arabic that conveys the same metaphorical content.Longman Idioms Dictionary (1998) defines the idiom as "a sequence of words which has a different meaning as a group from the meaning it would have if you understand each word separately." (p.vii) Lewis provides another concise dictionary-like definition: an idiom, he states, is " a multi-word lexical item where the meaning of the whole is not directly related to the meanings of the individual words." (Lewis: 1998, p.217). Cowie and Mackin (1975, p. viii) also stress the multi-word nature and semantic opacity of the idiom: an idiom, they write, " .. is a combination of two or more words which function as a unit of meaning." Unlike many others, however, Cowie and Mackin (1975) and Cowie (1983) are inclined to ".. describe idiomaticity as a feature cutting across all fixed expressions rather than have idioms as a separate category included within and subsumed by an overall framework of fixed expressions" (quoted in Carter, 2000, p.77)."Idiomatic translation", says Larson, is one " which has the same meaning as the source language

but is expressed in the natural form of the receptor language." (Quoted in Shuttleworth and Cowie (1997, p.73) "Idiom", on the other hand, may refer to a language or a style of expression which characterizes a certain group of language users. Elements such as idioms, metaphors and similes also pose challenges to the translation model of DeepL. For instance, the English phrase the ball is in your court, or as it is more commonly used, your turn, needs to be translated into Arabic in a way that conveys the same sense of responsibility to the reader, or the same sense of urgency to act. A word-for-word translation would simply state that the ball is in your court, but this would not necessarily convey the intended meaning of the original sentence. For this reason, a more suitable translation would be required, one that conveys the appropriate sense of responsibility or action to the reader. This form of translation requires an in-depth knowledge of the cultures of both the source language and the target language, which is something that machine translation cannot replicate as of yet.

The English phrase *the ball is in your court*, for example, means the next move is up to the other party, but there is no idiomatic equivalent in Arabic that conveys the same meaning through the same kind of metaphor. The translator has a choice between rendering the phrase literally and producing an inappropriately wordy text, or to opt for an alternative idiom that is more natural to the target language. A reasonable solution for this phrase might be القرار لديك (*alkararou ladaik*), or the decision is yours in English (Carter,2000, p.317).

Although some expressions, such as going to the bank (to deposit money), lend themselves to relatively easy translation (going to the bank in

Arabic is الذهاب إلى البنك, *adahab ila el bank*), which should be translated as إيداع المال في البنك أو الوديعة.(*Idaa el mel fil bank aw alwadiaa*) There are others that pose a greater challenge, especially when the expression conveys an abstract concept or certain undertones that are natural to speakers of English but not to speakers of Arabic. Some idiomatic expressions contain even more culturally-bound references that require a deeper understanding of the prevalent culture of each language group within specific sociocultural contexts. The English idiom *kick the bucket*, which is used to refer to death in general, is one such expression and has no equivalent idiom in Arabic. Even a simple and classic translation such as مات (*mat*), he died, does not achieve the idiomatic nature of kick the bucket, that associates the phrase with execution by hanging or suicide in the 18th century. A person standing on an overturned bucket with a noose around the neck would literally kick the bucket away to hang themselves (Holton, 1865, pp.164-165).

English speaks of barking up the wrong tree, a phrase that refers to searching for or pursuing something in a wrong or mistaken way, and which in many cases lacks a direct counterpart in Arabic. For example, instead of saying " البحث عن شيئ في مكان خطأ" (*baht ʻan shayʻan fil makān khaṭa* "searching for something in the wrong place") one would typically say تبحث في المكان الخطأ (tabḥath fi al-makan al-khaṭa "looking for something in the wrong place"). This idiom comes from hunting, referring to a hunting dog that has cornered its prey in a tree but is barking at the wrong tree because the animal has escaped. it means you are pursuing a mistaken line of thought, looking in the wrong place, or directing your

energy at the wrong person. This phenomenon demonstrates further the loss of an important dimension of the culture of the source language and the limitations of mere translation. A proper equivalent would offer the speaker of the target language a glimpse into the implications of this idiomatic expression or phrase. However, as noted, finding the right equivalent is a challenging task for the translator working in the realm of the intercultural communication of languages.

In his study on the treatment of idioms in student translation, Ennebati (2026) found that novice translators generally displayed an over-reliance on a literal treatment of idioms and idiomatic combinations. This over-reliance was often linked to a lack of cultural insight on the part of the student, specifically a lack of understanding of the source and target language cultures. The translation of idiomatic expressions is notoriously problematic because idioms so often reflect culture-specific values, experiences and features of language that evoke certain feelings or attitudes, including perhaps humor, that are closely tied to a specific localization.

The commodification of language is nothing new, yet the digital revolution seems to have added a new dimension to the subject (Qassem 2026). The process of translation into Arabic has become particularly complex, as many translators are tempted to resort to translation tools such as DeepL. While such resources can certainly be useful, particularly for the novice translator, they tend to generate wordy and stilted output that can give language a rather dismal existence.

With the rapid development of digital technology, the translation of

idiomatic and culture-bound expressions in English-Arabic contexts has in fact become even more challenging for English language students and Arabic translators. The accurate communication of such idioms requires a fusion of linguistic competence and a mature level of cultural knowledge. Although considerable progress has been achieved in the use of machine translation technologies for communicating with English-speaking clients, successful English to Arabic translation of idiomatic expressions in digital environments requires both nuanced cultural insight and sophisticated knowledge of English idiomatic expression. The following section will explore DeepL efficacy and accuracy in its task if translating idioms through some selected examples.

3- Translating Culture: DeepL Translation of Idioms

The future of English-Arabic translation is likely to be shaped in large part by developments in machine learning. In particular, deepL techniques have already had a transformative impact on translation technology, enabling systems that utilize complex neural networks to process large amounts of language data and thus generate high-quality, human-like translations that are more contextually appropriate than those produced by rule-based methods. Bialy et al. (2024) recently introduced a holistic framework for improving translation systems on the basis of deepL. The framework leverages neural networks to tackle a variety of issues that plague digital translation systems, including accurately identifying language-specific features, capturing source language variation, achieving balance between quality and fluency, and incorporating target language nuance and context.

Advances have been particularly noteworthy in the sub-field of English-Arabic neutral machine translation (NMT), whereby deep learning technologies offer marked improvements to those previously developed under the aegis of statistical machine translation. These model types- particularly those relying on recurrent neural networks (RNNs) and, more recently, transformers- process sentence-long input vectors and generate output one token at a time, thereby diverging in critical ways from traditionally employed statistical-based approaches. The novelty of transformers such as Open AI's GPT1 and Google's BERT further lies in their development of mechanisms to cope with long-range dependencies and increased contextual awareness among words, abilities which have proven to be especially useful in tackling Arabic's complex syntax and semantics while translating from English. Other advantages of deepL models include the use of attention mechanisms, whereby different words in the input sentence have weight assigned to them during the transformation process, enabling models to generate higher quality output and greater fluidity in translated segments.

Besides the novel methodologies, fine-tuning pre-trained models on domain-specific data has a significant impact on translation quality (Vaswani et al.,2017). While general language understanding is helpful for a number of applications, models that have been fine-tuned on specific corpora and tasks have been shown to achieve higher quality results. Arabic presents a unique set of challenges as the majority of its speakers are native speakers of different dialects, which can result in vastly different meanings being expressed with small changes in wording or context. The study

demonstrates the effectiveness of transfer learning methods (including fine tuning pre-trained models on smaller in-language datasets) in achieving state-of-the-art results for a less resourced language pair.

DeepL has undergone significant improvements in the past two decades. Ruled-based machine translation systems which were developed and implemented in the early days of the field, relied heavily on a set of predefined grammatical rules, as well as on predefined dictionaries and vocabulary.These early systems were however, fundamentally limited in terms of scalability and in terms of actual translation quality. Recently however, statistical and neural network-based approaches to machine translation have been proposed and successfully implemented in systems such as Google Translate, and offer a far more promising avenue for developing high quality machine translation systems. As described in Al Maaytah (2026), these systems known as neural machine translation (NMT) systems, consist of deepL models that have been trained on large datasets of multilingual text, and have been shown to generate human quality translations for a wide range of input including for Arabic. Importantly, even for straightforward, L1 source text, the output quality and readability of state-of-the-art English to Arabic translation systems is arguably sufficient, with little more than acceptable fidelity for basic communication. Thus, while limitations remain with respect to text that contains more complex or culturally-specific content, for general consumption, such translation systems can be viewed as adequate and sufficient for many practical settings.

DeepL's availability in 120 languages powers modern (NMT) by

treating translation as a sequence-based task. It does not perform word-to-word substitution, but uses neural networks to analyze whole sentences at a time (Mittenburg, 2017). This allows it to take into account context, grammar and nuance, helping it to produce fluid, human-sounding translations.The History of Machine Translation DeepL introduced three major technological leaps to this field: Firstly, Rule-Based (RBMT): linguistically trained rules and bilingual dictionaries. It struggled with idioms and colloquial expressions a lot. Secondly, Statistical (SMT): Used large, bilingual datasets of text to probabilistically predict the most likely translations.It was much more flexible but without global sentence context. Thirdly, Neural (NMT): Uses deep Artificial Neural Networks to develop complex linguistic features automatically from data, allowing machines to recognize paragraph-level context (Zhang et al., 2025).

The impact of digital technology on translation practice, specifically the Arabic language, is two fold: while there are advantages of speed and accessibility that come with translation through machine learning algorithms, there are a number of challenges regarding cultural nuance and accuracy that continue to plague translators. As machine translation technology, particularly that based on deep learning and neural networks continues to evolve, the need for human mediation and culturally-specific knowledge and practice persists. This conclusion look to the future of translation practice within a digitized context.

The use of DeepL to translate English into Arabic highlights complex challenges to the communication of cultural nuances and there is a lot more to translate than individual words. Such digital translations are

continually being improved and there is much that has now been perfected, for instance the powerful NMT technology which underpins DeepL's translations utilizing a number of advanced deep learning techniques that form part of a huge data-set which it uses to aid understanding in a particular context allowing highly accurate translation that goes beyond the literal word-for-word translation of earlier rule-based systems, and achieving more accurate translation in many instances, particularly when dealing with complex contexts. But translating between two languages that as vastly differently in terms of how cultures are embedded within them as English and Arabic, clearly presents difficulties which this innovative translation tool is not yet able to fully address.

One of the biggest strengths of DeepL when it comes to translating text is its ability to emulate the unique style and structure of the language it is translating from. For example, when translating the English phrase to break the ice into Arabic using DeepL, the translation would be *to break the ice* and is literally translated from the English word for word as كسر الجليد. *(kasr el jalid)* The translation captures the literal meaning of the phrase, but fails to translate the deeper cultural meaning of the phrase. To break the ice is a very common English idiom, and is used to describe the action of starting a conversation with someone in order to 'open up' the interaction and make it more friendly. In Arabic there are many different colloquial expressions that are used to describe starting a conversation, and the most appropriate would be to open up a conversation or to describe starting a conversation by using the Arabic word فتح meaning to open, and then combining this with حديث meaning conversation, i.e. to open up a

conversation.

A simple phrase such as greetings could be translated word by word; however, it can have a much deeper meaning depending on the culture in which it is being used. For example, in English, a common greeting would be How are you? In Arabic, the translation of this phrase is كيف حالك؟ (Kayfa halak?). The literal translation could mean how are you? But in Arabic culture, the greetings are used to inquire about the health of the speaker and their family. The Arabic phrase for greetings are used to show respect to the other person and create a hospitable environment. The response to the greeting would typically be fine thank you and also ask about the health of the speaker's family. In addition, the Arabic culture places a lot of emphasis on the manner in which the greeting is delivered. For example, the greeting could be delivered in written form or verbally. Also, the tone and the volume of the voice in which the greeting is delivered can also change the meaning of the greeting. The culture also places emphasis on the timing of the greetings. For example, greeting someone during specific times of the day or on special occasions can have a different meaning than greeting someone at other times. The response to the greeting could also change depending on the situation. For example, in formal situations, the response to the greeting would be more formal than in an informal situation. Therefore, the phrase How are you? in English can be translated to كيف حالك؟ (Kayfa halak?) in Arabic; however, the meaning of the phrase can be different depending on the culture in which it is being used.

Moreover, DeepL fails to translate humor to other languages. A pun

or a joke which is so funny in English may not be funny in Arabic for instance. The translation of the famous phrase *Time flies when you're having fun* is for instance "passes quickly when having fun", which obviously fails to evoke the same kind of emotions as the original phrase did. Instead of translating idiomatically such phrases it is better to translate them using proverbs and sayings from the target culture, thus also making them more familiar to the readers. An example for this is the English phrase given above which can be translated into Arabic as يومٌ واحدٌ من الفرح يُنسينا همومَ سنة (One day of joy makes us forget the worries of a year).

It is in these situations that the translations by DeepL of such phrases from English are also wanting as the translations lack the necessary contextualization to cushion the impact of the translation on the reader, as machine translation programs such as DeepL are lacking in the ability to grasp the social fabrics that shape human language, as Noorain and Alamin (2026) noted.

Another important idiom that can be translated from English into Arabic and carry different meanings is the idiom "*bringing home the bacon*" which is used to mean bringing home financial support for the family. The translation of this idiom in Arabic would be a word by word translation of the idiom and would not carry the same meaning and would not express the idea of providing financially for the family. In Arabic cultures there is a different way of expressing the idea of providing and being responsible for the family and this is often linked to the honor of the family and to the idea of responsibility towards the community.

However, in translating such idiosyncratic expressions in digital

platforms, DeepL could be failing to enable the user to take advantage of the relationship between a word and the other words which surround it. An example of this, and of how translations of different idiom can go in entirely different directions, would be the English saying too many cooks spoil the broth or a similar idiom that conveys the idea that when there are a lot of people involved in bringing about an end or accomplishing something, the result is that the final product is usually inferior in some way. It can thus be seen that while DeepL's version of this idiom,translated to Arabic, may preserve the basic meaning of the idiom, what is lost in translation is not the fundamental meaning of the phrase, but rather its core essence. Consequently, DeepL could be successfully transmitting to the Arabic speaker the basic meaning of a given idiom pertaining to a particular scenario, and at the same time not transmitting the essence of that idiom. Even more problematic, the translation supplied by DeepL of the aforementioned idiom could inadvertently give the impression that Arabs view collaboration as being different than the Western view of such. This, too, would be giving the wrong impression of the relationships which exist between people from vastly different cultures. While the translation supplied by DeepL of an idiom which is utilized by the English-speaking community could reinforce a number of generally-held stereotypes of the Arab in unfavorable ways, the translated idiom could also fail to convey the type of inter-personal relationships that occur in the workplace, thereby missing an opportunity to supply the reader with knowledge of the variety of aspects which need to be considered in a given collaborative setting.

Similarly, another idiom, "*the apple of my eye*," which refers to

something or someone that one cherishes and adores, cannot be translated literally into Arabic and usually results in a meaningless translation by DeepL. For example, "he looks at me like his eye" (ينظر إليّ كالعين). Although this expression conveys possessiveness and affection towards the person being looked at, it is used in different contexts and conveys different meanings. This idiom is mostly used within the context of family members.

The translation of texts between English and Arabic using digital platforms such as DeepL also serves to probe into the issues of over-translation or over-explanation and their implications in translation in general and digital translation in particular. Abo-Sahal (2025) discussed the issues of over-translation where translators over-explain or over-translate source texts and end up distorting the intended messages of the original texts. The over-translation could sometimes be caused by translators' failure to realize the limits of language in conveying meaning and the cultural biases which they might not even be aware of. Most importantly, the over-translation could also be caused by translators' interference in the translation process whereby they project their own cultural values and bring in their own linguistic choices which might not be suitable for target-language readers. The issue of over-translation is therefore especially pertinent in the context of digital translation where rapid and efficient translation is often the goal.

Over-translation is another phenomenon that is often seen in translations of Arabic idiomatic expressions. Abo-Sahal (2025) discussed the translation of the English phrase spill the beans, which is translated into Arabic as يسكب الفاصوليا.(*yaskabou alfasoulia*) Although this phrase can

be literally translated into English, the translation has no meaning in the target language. It means to reveal a secret or disclose confidential information prematurely, often by accident This is exactly what happens when DeepL translates idiomatic expressions word for word. The translation fails to capture the underlying cultural message or even convey the appropriate sense of humor that is intended by the speaker.Therefore, the Arabic-speaking readers may be mislead by DeepL's translation of this English idiom because the phrase to spill the beans in Arabic conveys no meaning (Qassem, 2026).

Bias in the design and in the programming of the translator also needs to be considered. Abo-Sahal (2025) describes how digital translators are primarily based on massive data bases that have been compiled to translate texts from one language into another. This large corpus of translated texts will inevitably reflect the dominant ideologies of the cultures in question. For example, gender identity terms that are commonly used in English may not have equivalents in Arabic. Their translation into Arabic can therefore reflect biased views on gender, reinforcing gender stereotypes instead of representing the identities of individuals who do not conform to the two genders that are traditionally accepted in Arabic-speaking cultures. As such, users of digital translators may find that their understanding of certain terms is influenced by the prevailing attitudes found in the translations consulted.

3-The Impact of DeepL on English-Arabic Translation: Implications

Idioms are also affected by the digital revolution, both positively and negatively. At first, machine translation tools uses, DeepL were hardly

sophisticated enough to translate even simple stretches of text, and, to a certain extent, machine translation has remained inadequate for idiomatic language ever since. While current programs such as DeepL still do not do an adequate job with certain idiomatic phrases-e.g., "Hold your horses!" is translated as "to hold the horses," despite the fact that this has nothing to do with calming down-Hijazi (2026) points out that even though technology is currently more sophisticated and able to access greater numbers of resources online, this often results in a translation that fails to deliver in terms of contextual accuracy and therefore is of little use in practical situations.

These examples illustrate the problem that DeepL, as well as many other current digital translators, have with Arabic-English translation, and show clearly how they can fail in important respects, notwithstanding their many advantages. Their translations of various idiomatic and quasi-idiomatic food-related English expressions may be of interest, and in some cases even informative or amusing, but they by and large fail to capture the complex of meanings that are generally connoted in the languages by comparable terms. Indeed, the translations may even on occasion be off-putting.

Reading translation errors in English—Arabic language pairs into potential failures in contextualizing and culturally interpreting can be useful in outlining the issues faced by current digital translation technologies in transferring cultural meanings between languages and cultures. Examining examples of particular difficulties in translating various types of source language language pairs into target language pairs, for

example English-Arabic language pairs, Sasi et al. (2026) identify potential failures of translation including loss of source language context and failure to understand culturally specific meanings.

Additional problem encountered in translation of word choice to reflect social hierarchies and relationships is that of politeness and formality. In English literature formal addresses are generally reserved for older people or for superior in work place. Formal pronouns and verb forms in Arabic literature on the other hand are used in addressing of respect to any person of higher status or ranking whether of superior age or not. For instance DeepL could translate "Please, could you help me?" into Arabic of a form that does not carry enough formal respect to another such as " حسناً، هل يمكنك مساعدتي؟" (Hasanan, hal yumkinuka musa'adati?). Misunderstanding of such inappropriately translated expressions may lead to offense to another and to failure in making proper social relationships. Such failure in formal politeness could be more critical in certain Arabic speaking cultures and could lead to serious problems. It can cause failure in achieving objectives in transactions with others and even result in social isolation. All these failures would not have occurred if appropriate formal respect had been shown in appropriate circumstances. Sasi et al. (2026) explained how inappropriate level of formality in translation could cause serious problem and that would depend on the context in which the translation is to be used.

Moreover, many words and phrases that are used in language and communication, especially in written form, are humoristic and use language play in various forms. Most of these forms of language play are

culture-specific and do not get translated directly in a target language. Therefore, DeepL will most likely fail to translate puns or other culturally specific humoristic expressions. The consequences of not getting these expressions translated are twofold. On the one hand, the translation will lack aesthetic quality, since the language play has been an important aspect of the source text. On the other hand, the translation will also lack social and emotional resonance, since the humoristic expression has been an important means to convey attitudes and to create interaction with other human beings. Therefore, Sasi et. al. (2026) conclude that there are still many challenges for digital translation technologies to master.

The examples described above highlight the limitations of current translation technologies and emphasize the need for further technological innovation as well as educational strategies that can help students develop the skills required to critically evaluate translations generated by digital systems. While technologies such as DeepL have the potential to greatly facilitate the processes of learning a second language as well as translation, they must be used in conjunction with other pedagogical tools that promote an in-depth understanding of the complex nuances of language and culture.

While digital technologies such as translation apps are advancing, the translator's role in transferring cultural insights is undermined by the limitations of these tools. As more idioms and forms of spoken language are translated, it becomes increasingly important for translators to critically evaluate the results, to guarantee that the cultural content of these language pairs is accurately conveyed. Hijazi (2026) proposes using

bilingual corpora that ground idioms in their respective cultures in order to select translated equivalents that have real depth and resonance in the target language.

Translating English idioms into Arabic is a very sensitive task, and perhaps no more so than in the current digital age. When translating idioms it is essential to have a human touch, as machine translation lacks subtle nuance when dealing with idiomatic language, and the difficulties of translation are increased when crossing cultural boundaries. Future research on translation will continue to focus on idiomatic language, and cultural awareness will remain crucial to successful communication.

Technology has brought unparalleled benefits to translation, particularly within the digital sphere. The rapid dissemination of multilingual content on the World Wide Web and the advances made in computer-assisted translation have both permitted the implementation of rapid translation as well as widening the range of translation service users. However, within the Arabic context, cultural nuances may be easily lost in translation from English to Arabic. In order to address such challenges in nuance translation between two very different languages and socio-cultural contexts, some emphasis could be placed in education on attempting to bridge these gaps by providing translators with targeted training (Tergui, 2024). In light of the recent digital shift in translation, the task of translator education is a new and ongoing issue.

Bridging cultural gaps in translation through education and training programs require a package of integrated educational strategies such as cultural studies, experiential education, use of technology and translation

tools, partnership with students through training programs, and use of translation cases. The paper discusses possible ways of how translators in training, especially those specializing in English-Arabic translation, can deal with the intricacies and changes in this field, particularly cultural challenges. translator's role within an ethical context is highly significant; it is greater than in any other branch of communication. It is even greater than in interpreting since interpreters have very little time to think, reflect and make decisions that may affect what others are going to say. If a translator fails to understand a certain cultural element he or she tend to make mistakes by offering a compromised translation. Such a translation could either misrepresent certain aspects of cultures or get caught up in negative stereotyping or even be considered offending.

This complex of considerations has to be tackled in practice by a thoughtful approach to translation. The rapid developments of the digital era have rendered translators more than just linguistic experts capable of handling the two languages in question, English and Arabic. Instead, they also need to have an exquisite sense of the cultures in question, for it is a consensus among researchers that translation to and from these languages is no less than a demanding task that entails intricate dimensions. As Mohammed and Al-Azzawi (2025) assert in their comprehensive analytical critique of new developments in the field of translation in general and digital translation in particular, what seems particularly taxing in dealing with cultural specificities is the intricacies of idiomatic expressions and the culturally-bound contexts. However, the digital turn has made these even more complicated and opened up the door for some new technologies and

strategies to tackle these complexities, some more convincingly than others.

As machine learning for English-Arabic translation continues on its trajectory, it is hopeful that consistent progress will be made in achieving higher translation quality through further research and effective use of current models to address English-Arabic translation difficulties. As translation moves into an all digital, interconnected realm, it is critical to study the complex relationship between technology, language and user needs. Deep learning techniques have played a profound role in the improvements of machine translation, particularly challenging language pairs such as English-Arabic. Using multilayered neural networks, deepL models have vastly improved the quality of machine translation in terms of fluency, accuracy and overall understanding of context. One model in particular, the Transformer, has redefined traditional translation models and achieved state-of-the-art results for highly challenging languages in any multilingual setting.

4-CONCLUSION

This paper surveyed some of the recent developments on English-Arabic machine translation, focusing on the role of deepL platform in the translation of idioms, and looking forward to future advancements. The paper discussed through examples, current technology and challenges, and looks at some state-of-the-art systems and research. The future of English-Arabic translation is likely to be shaped in large part by developments in machine learning. In particular, deepL techniques have already had a transformative impact on translation technology, enabling systems that

utilize complex neural networks to process large amounts of language data and thus generate high-quality, human-like translations that are more contextually appropriate than those produced by rule-based methods.

In an attempt to facilitate a more meaningful way of learning cross-cultural translations, tools that utilize complex AI-powered algorithms, such as DeepL, aim to translate not only single words, but also entire contexts within which these words are embedded. Challenging, however, to be translated through such technologies are also the many cultural expressions that are not explicitly acknowledged within existing corpuses of human translations. This has led many to suggest that AI technologies currently are not adequate as learning tools for translating between languages and cultures, as students may in the end merely memorize phrases which they do not understand in the complex and multifaceted contexts in which they are used.

The challenges in utilizing translation technology for language teaching extend to many more facets than simple linguistic correspondence, affecting educational models of language learning as well. Educational approaches can utilize the translations generated by DeepL in a manner that facilitates deeper study of the translated texts, or the tools can be used in such a way that to depend on them for learning a foreign language hinders cognitive growth and an in-depth study of the cultural discourses that are expressed by means of language and understood by virtue of it. If, in the age of digital technology, teachers do not arm their students with sufficient knowledge of the currently used translation software in order to be able to think critically about its potential biases and

to analyze and verify the translations generated by this software, the latter shall turn into yet another teaching aid that is easily misused in educational contexts, and which can in fact have a number of very negative consequences for the foreign language learning process of a student.

This paper closes on a hopeful note underlining the significance of cultural nuance and linguistic ability in English-Arabic translation in the digital era. It suggests that, rather than seeing contemporary changes in language use and culture with apprehension, translators can transform these developments into new chances for flourishing linguistic communication across cultures. By recognizing the challenges of this era as opportunities for growth, translators can reaffirm their role as life-long learners and as vital resources for cross-cultural understanding in an ever-changing global society.

Translation in the Digital Age continues the dialogue into new territories, calling for more research and for the refinement of methodologies and strategies. After all, in today and tomorrow's world, words are not enough; skills in one language and proficiency in another culture must merge into a hybrid translation practice. In short, the translation of digital content from one language/culture to another demands great linguistic sophistication and considerable cultural acumen. This anthology features thirteen essays that focus on aspects of translation, revealing the complexities and deep-seated cultural nuances which demand translator empathy and negotiation strategies to reach an accurate, target-language rendering of the original message.

References

Abu-Ssaydeh, Abdul-Fattah Saleh (December 2004). Translation of English idioms into Arabic *Babel Revue internationale de la traduction / International Journal of Translation*50(2):114 131 DOI: 10.1075/babel.50.2.03abu

Abo-Sahal, M. A. (2025). The Ideological and Cultural Imperative: Over-Translation as a Target-Oriented Norm in English-Arabic Translation. مجلة العلوم الشاملة, 2317-2306 , (35)9.

Al Maaytah, S. A., Aalzobidy, S. A., & Alwidyan, M. F. (2025). Using Machine Translation: English-Arabic Procedures and Challenges-A Systematic Review. Power System Technology, 49(1), 588-607.

https://www.researchgate.net/profile/Shahab-Almaaytah-2/publication/389491280_Using_Machine_Translation_English_-Arabic_Procedures_and_Challenges_-A_Systematic_Review/links/67c40b4b645ef274a498a040/Using-Machine-Translation-English-Arabic-Procedures-and-Challenges-A-Systematic-Review.pdf

Al Maaytah, S. A. (2026). Evaluating Three Neural Machine Translation Platforms for English-Arabic Translation: A Comparative Study of Linguistic Accuracy and Cultural Fidelity. World, 16(2). https://www.researchgate.net/profile/Shahab-Almaaytah2/publication/394946031_Evaluating_Three_Neural_Machine_Translation_Platforms_for_EnglishArabic_Translation_A_Comparative_Study_of_Linguistic_Accuracy_and_Cultural_Fidelity/links/68ad5ca3ca495d76982ff094/Evaluating-Three-Neural-Machine-Translation-Platforms-for-English-Arabic-Translation-A-Comparative-Study-of-Linguistic-Accuracy-and-Cultural-Fidelity.pdf

Bialy, A., AbdElrazek, E., & Hawa, D. (2024). إطار عام يعتمد على التعلم العميق للترجمة الرقمية من العربية إلى الإنجليزية. المجلة العلمية لكلية التربية النوعية جامعة دمياط, 19-1 ,(9)2024. https//:journals.ekb.eg/article_339314.html

Carter, R. (2000) Vocabulary: Applied Linnguistic Perspectives. 2nd edition. Routledge, London and New York. Pp.317.

Cowie, A.P. (1983). English dictionaries for the foreign learner. In R. Hartman (ed) Lexicography: Principles and Practice. London: Academic Press. Pp. 135-44.

Cowie, A.P. & Mackin, R. (1975). Oxford Dictionary of Current Idiomatic English: Verbs with Prepositions & Particles. Oxford: Oxford University Press.

Elhadary, T. (2023). Linguistic and cultural differences between English and Arabic languages and their impact on the translation process. International Journal of Language and Translation Research, 3(2), 103-117. http://www.ijlatr-ac.ir/uploads/pdf/Vol3Is2P6.pdf

Ennebati, F. Z. (2026). Understanding the Challenges and Difficulties of English-Arabic Idiom Translation: Insights From Students. Journal of Languages and Translation, 6(1), 245-256. https://journals.univ-chlef.dz/index.php/jlt/article/view/1054

Faes Florian (2017) Why DeepL Got into Machine Translation and How It Plans to Make Money https://slator.com/deepl-got-machine-translation-plans-make-money/

Hijazi, F. (2026). The Translatability of English Idioms into Arabic: A Comparative Study of Challenges and Pedagogical Strategies. Mesopotamian Journal of Arabic Language Studies, 2026, 1-5.
https://journals.mesopotamian.press/index.php/MJALS/article/view/1002

Hotton, John Camden. *The Slang Dictionary*, London 1865.

Longman Idioms Dictionary. (1998). UK: Longman, England.

Lewis, M. (1998). *Implementing the Lexical Approach*. LTP. England.

Miltenburg, van, Olaf (29 August 2017). "Duits bedrijf DeepL claimt betere vertaaldienst dan Google te bieden" [German company DeepL claims to offer better translation service than Google]. *Tweakers*

Mohammed, M. A., & Al-Azzawi, I. G. (2025). The impact of culture in translating English idioms into Arabic.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5175247

Noorain, A. A. S., & Alamin, M. I. M. A. (2026). Untranslatable Arabic Lexemes:(A Comparative Analysis of Semantic, Cultural, and Translational Challenges in Arabic-English Translation). *International Journal on Humanities and Social Sciences*, (70), 234-244.

Qassem, M. (2026). Analysis of student performance in cross-cultural translation: English to Arabic. SN Social Sciences, 6(6), 221.

Sasi, H. B., Alsunousi, Z., Krwad, S., Alkhaldi, R., Baayo, S., & Abusha'ala, A. (2026). Challenges of Meaning Transfer in English–Arabic and Arabic–English Translation. Asshika: Journal of English Language Teaching and Learning, 3(2), 91-105.

Shuttleworth, M. and M. Cowie. (1997). *Dictionary of Translation Studies*. UK: St Jerome.

Tergui, S. (2024). Investigating translation students' challenges in rendering Arabic idioms and proverbs into English: insights and recommendations. AWEJ for Translation & Literary Studies, 8(4).

https://papers.ssrn.com/sol3/papers.cfm?abstract_id= 5029426

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. arXiv preprint arXiv:1706.03762

Zhang, Y., Wang, Y., Sheng, Q. Z., Yao, L., Chen, H., Wang, K., & Zhao, R. (2025). Deep learning meets bibliometrics: A survey of citation function classification. Journal of Informetrics, 19(1). DOI 10.1016/j.joi.2024.101564

Arabic Translation under Artificial Intelligence: A Comparative Study of Theoretical and Machine-Based Approaches

Ahlem Bira
Laboratory: Translation and Didactic of Languages
Badji Mokhtar Annaba University

Abstract

This paper investigates the quality of AI-assisted Arabic-English medical translation by examining the extent to which contemporary AI translation systems satisfy the linguistic and functional requirements of specialized pharmaceutical translation. Grounded in Newmark's semantic and communicative translation framework, Nida's dynamic equivalence, and Skopos theory, the study adopts a comparative analytical approach to evaluate AI-generated translations using the officially validated English version of a pharmaceutical patient information leaflet as the reference translation. The corpus consists of the trilingual patient information leaflet of POLYDEXA® ear drops, obtained from an Algerian pharmacy, a context that highlights the limited availability of Arabic–English pharmaceutical documentation in the Algerian healthcare sector. Three AI translation systems Google Translate, DeepL, and ChatGPT were evaluated across three sections of the leaflet (composition, indications and

contraindications, and warnings) according to five quality criteria: terminological precision, syntactic accuracy, pragmatic adequacy, ambiguity resolution, and cultural sensitivity. The findings show that, although all three systems produce fluent and grammatically acceptable translations, they exhibit recurring limitations in the rendering of specialized medical terminology, the preservation of clinically relevant information, and the pragmatic functions of pharmaceutical instructions. Among the systems evaluated, ChatGPT achieved the highest overall performance, followed by DeepL, whereas Google Translate displayed the greatest variability in terminological accuracy. The study concludes that AI translation systems can effectively support medical translation but cannot yet replace professional revision in safety-critical contexts. It also highlights the need for Arabic-specific medical AI resources and greater availability of Arabic–English pharmaceutical documentation to support reliable AI-assisted medical translation.

Keywords: Arabic–English Medical Translation - Artificial Intelligence- AI Translation Systems- Pharmaceutical Translation- Translation Quality Assessment - Patient Information Leaflets.

I. Introduction

The rapid advancement of artificial intelligence has profoundly reshaped the landscape of translation studies, introducing new tools and paradigms that challenge long-established theoretical frameworks. Among the domains most critically affected by this transformation is medical translation, which itself constitutes a vital branch of community translation

a broader field concerned with facilitating communication between institutions and individuals who, due to linguistic or cultural barriers, are unable to access essential public services on equal terms. Community translation, in this sense, is not a peripheral academic concern but a fundamental instrument of social inclusion, ensuring that vulnerable populations including migrants, refugees, and minority language speakers can meaningfully engage with healthcare systems, legal institutions, and social services.

Within this broader framework, medical translation occupies a particularly critical position. Unlike other forms of specialized translation, it operates at the intersection of language, science, culture, and human well-being, where a single terminological error or pragmatic misjudgment can have direct consequences for patient safety and clinical outcomes. Arabic medical translation adds yet another layer of complexity to this picture, given the morphological richness of the Arabic language, the diversity of its regional varieties, and the cultural specificities that inevitably permeate medical discourse from the conceptualization of illness and the body to the expression of sensitive health topics such as mental health and reproductive medicine.

AI-based translation systems, including Google Translate, DeepL, and ChatGPT, have demonstrated remarkable progress in generating fluent and grammatically acceptable translations across a wide range of language pairs. However, fluency alone does not constitute quality in specialized translation. Medical texts whether clinical research abstracts, pharmaceutical leaflets, or patient-facing documents demand a level of

terminological exactness and pragmatic adequacy that goes far beyond surface-level linguistic competence. The question therefore arises as to whether these AI-powered tools are capable of meeting the rigorous standards set by established translation theories and the ethical responsibilities inherent in community-based medical communication.

This study seeks to address this question through a systematic comparative analysis of AI-generated Arabic–English medical translations using the officially validated English version of a pharmaceutical patient information leaflet as the reference translation. Grounded in Newmark's semantic and communicative translation framework, Nida's theory of dynamic equivalence, and the functional principles of Skopos theory, the analysis examines the extent to which AI translation systems meet the linguistic and functional requirements of medical translation. Particular attention is given to terminological precision, syntactic accuracy, pragmatic adequacy, ambiguity resolution, and cultural sensitivity in the translation of pharmaceutical discourse.

The significance of this research lies not only in its contribution to translation studies but also in its broader practical implications for AI-assisted medical translation. As AI translation systems become increasingly integrated into healthcare communication, evaluating their reliability in safety-critical medical contexts is both academically necessary and professionally relevant. By assessing the performance of contemporary AI translation systems in light of established translation standards, this study seeks to provide translators, researchers, and healthcare professionals with a clearer understanding of the current capabilities and limitations of AI in

Arabic medical translation. It also highlights the need for the development of Arabic-specific AI translation resources capable of addressing the linguistic, terminological, and cultural characteristics of Arabic medical discourse.

II. Theoretical and Literature Framework

2.1 Community Translation and Medical Translation as a Social Practice

Language barriers in healthcare settings represent one of the most pressing challenges facing linguistically diverse societies today. Research consistently demonstrates that limited language proficiency delays access to healthcare services, impairs therapeutic relationships between patients and providers, and ultimately compromises patient safety and clinical outcomes (Pandey et al., 2021 :4). In this context, community translation broadly defined as the practice of facilitating communication between public institutions and individuals who lack proficiency in the dominant language has emerged as a fundamental instrument of social inclusion and equitable healthcare delivery (Al Shamsi et al., 2020 :2).

The consequences of inadequate language mediation in clinical settings are well-documented. A systematic review by Al Shamsi et al. (2020) found that language barriers lead to miscommunication between medical professionals and patients, reducing the quality of healthcare delivery and directly compromising patient safety across multiple national healthcare contexts (Al Shamsi et al., 2020 :3). Similarly, a meta-review of migrant healthcare access concluded that language barriers prevent

patients from understanding medical instructions, navigating healthcare systems, and building trust with providers, while translation services though demonstrably effective remain chronically underutilized in institutional settings (Díaz-Millón et al., 2025 :6).

Within this broader framework, medical translation occupies a particularly critical position. Unlike other forms of specialized translation, it operates at the intersection of language, science, culture, and human well-being, where a single terminological error or pragmatic misjudgment can have direct consequences for informed consent and clinical outcomes. The deployment of technological tools including machine translation applications to overcome language barriers in healthcare has grown considerably in recent years, yet their usability remains constrained by persistent challenges related to translation accuracy, accessibility, and cultural appropriateness (Kreienbrinck et al., 2025 :9).

2.2 The Evolution of Machine Translation and Neural Systems

The history of machine translation has evolved considerably from rule-based systems through statistical approaches to the current dominance of neural machine translation. A systematic review of MT systems and quality assessment procedures confirms that neural MT is the predominant paradigm in the current translation landscape, with Google Translate being the most widely used system across research evaluations (Rivera-Trigueros, 2021 :5). The emergence of large language models has further transformed the field: studies comparing Google Translate and ChatGPT on Arabic-English translation pairs demonstrate that while both

systems show significant progress in surface-level fluency, they continue to exhibit limitations that necessitate human post-editing for high-stakes specialized content (Mohammed, 2025 :8).

The application of these tools to medical texts has revealed specific and persistent shortcomings. A comparative study evaluating DeepL, ChatGPT, and Google Translate on medical text translation using standardized metrics found that while all three systems delivered acceptable surface-level results, they showed no statistically significant differences in quality, suggesting that none has achieved the level of reliability required for autonomous medical translation (Block et al., 2025, p. 3). Furthermore, research on AI-supported post-editing of Arabic translations found that while ChatGPT demonstrated competitive efficiency, it faced notable challenges in handling grammatical and syntactic nuances, domain-specific idioms, and complex terminology, especially in medical contexts (Algaraady et al., 2025 :7).

2.3 Arabic Medical Translation: Challenges and Prior Research

Arabic presents a distinctive set of challenges for both human and machine translation systems. Its rich morphological system, root-and-pattern derivational structure, and the pervasive phenomenon of diglossia the coexistence of Modern Standard Arabic and a spectrum of regional colloquial varieties create significant complexity for NMT systems. Research confirms that the limited vocabulary size required by NMT models decreases vocabulary coverage for Arabic, given its well-known morphological richness, and that long sentences present additional

challenges as systems perform less well with increased sentence length (Berrichi et al., 2021 :2).

A comprehensive survey of Arabic machine translation highlights that despite important developments in NMT approaches for Arabic-English translation, the quality of translation remains moderate and requires considerable improvement, with researchers continuing to identify open issues related to linguistic and technical challenges specific to the Arabic language (Zakraoui et al., 2021 :2). Error analysis of pretrained language models on English-to-Arabic translation further demonstrates that translation from English to languages with complex morphology such as Arabic remains challenging even for state-of-the-art systems including GPT-3.5 and GPT-4, with systematic errors identified across grammatical and vocabulary dimensions (Al-Khalifa et al., 2024 :4). In the domain of scientific and technical text translation specifically, a study evaluating Google Translate and ChatGPT on English-to-Arabic scientific texts found that both systems still require considerable improvement, and that annotated corpora need to be urgently constructed to support more reliable Arabic NMT development (Alzain et al., 2024 :6).

3.4 Translation Theories as Evaluative Standards for Medical Translation Quality

The evaluation of translation quality in specialized medical domains requires a theoretically grounded framework capable of assessing not only surface-level linguistic accuracy but also the deeper communicative, pragmatic, and functional dimensions that are most consequential for

patient safety. The present study draws on three complementary theoretical frameworks Newmark's semantic and communicative translation theory, Nida's theory of dynamic equivalence, and the functional principles of Skopos theory applied not in isolation but as an integrated evaluative lens through which the quality of both human and AI-generated Arabic medical translations can be systematically assessed.

Newmark's semantic and communicative translation theory constitutes the first pillar of this framework. Communicative translation attempts to produce on its readers an effect as close as possible to that obtained on the readers of the original, while semantic translation attempts to render, as closely as the semantic and syntactic structures of the target language allow, the exact contextual meaning of the source text (Wang, 2018 :2). In the medical domain, this distinction carries direct clinical implications: composition sections and clinical research abstracts directed at specialist audiences demand a predominantly semantic approach that preserves terminological precision and structural integrity, whereas patient-facing documents such as pharmaceutical leaflets require a communicative strategy that prioritizes accessibility and pragmatic clarity (Peng, 2026 :3). Applied to AI-generated translations, Newmark's framework provides the criterion of terminological precision the extent to which machine output preserves the exact denominative value of specialized medical terms without substitution, approximation, or register displacement.

Nida's theory of dynamic equivalence -later reformulated as functional equivalence- constitutes the second pillar. Its central proposition

holds that the goal of translation is not formal correspondence between linguistic structures but rather the production of an equivalent response in the target reader, encompassing the communication of linguistic information, the transmission of stylistic register, and the elicitation of a similar reader reaction (Li, 2021 :2). This audience-oriented framework is particularly relevant to Arabic medical translation, where cultural frameworks surrounding health, the body, and illness may differ substantially from those encoded in English-language source texts. A study combining Skopos theory with Nida's dynamic equivalence in the translation of patient information leaflets demonstrates that strategies such as paraphrasing, restructuring, and cultural adaptation enhance readability and patient adherence while maintaining functional medical accuracy (Ahmed et al., 2025 :4). For the present study, Nida's framework provides the evaluative criterion of pragmatic adequacy: the extent to which AI-generated translations produce a functionally equivalent communicative effect on Arabic-speaking target readers.

Skopos theory, developed by Reiss and Vermeer, constitutes the third pillar. Its central premise holds that the primary determinant of any translation is its intended purpose (its *skopos)* in the target culture, shifting the evaluative focus from source-text fidelity to target-text functionality (Nedainova, 2021: 2). This purposive orientation is particularly consequential in pharmaceutical translation, where the same source content may require radically different translational treatments depending on whether it is directed at a specialist clinician, a pharmacist, or a patient with limited health literacy. A functional study applying Skopos theory to

English-Arabic medical translation of patient health information concludes that translation strategies such as explicitation, addition, and substitution successfully achieve textual coherence, while literal translation frequently fails to convey medical information to non-specialist target audiences (Nine et al., 2025: 7). Applied to AI systems, Skopos theory raises the critical question of whether machine translation tools which operate without explicit knowledge of communicative purpose or audience profile are capable of making the context-sensitive decisions that effective medical translation demands.

Taken together, these three frameworks constitute a comprehensive and mutually reinforcing evaluative grid for assessing Arabic medical translation quality. Newmark's framework addresses *how* meaning is transferred and whether the translational strategy is appropriate to the text type. Nida's theory addresses the **effect of the translation** on the target reader. Skopos theory addresses its *purpose* and functional adequacy for its specific audience. Together, they underpin the five-criterion evaluation grid applied in this study terminological precision, syntactic accuracy, pragmatic adequacy, ambiguity resolution, and cultural sensitivity providing a theoretically justified and empirically applicable standard against which the performance of AI translation systems in the Arabic medical domain can be rigorously and systematically assessed.

III. Methodology

This study adopts a comparative analytical approach to evaluate the quality of Arabic-English medical translation produced by three AI-based

systems. The corpus consists of the official trilingual patient information leaflet of POLYDEXA® ear drops a pharmaceutical product of French origin obtained from an Algerian pharmacy, whose leaflet exceptionally provides official text in Arabic, French, and English simultaneously. This trilingual configuration, rare in the Algerian pharmaceutical market where Arabic-French bilingualism constitutes the dominant norm, renders the document a uniquely suitable corpus for Arabic-English translation analysis (Benamar & Temmam, 2024: 3).

The Arabic text of the leaflet serves as the source text, and three sections were selected for analysis: composition, indications and contraindications, and special warnings. Each section was submitted without modification to three AI translation systems: Google Translate, DeepL, and ChatGPT (GPT-4o) under identical controlled conditions, with no prompt engineering or post-editing intervention. The resulting AI-generated outputs were systematically compared against the official English translation printed on the leaflet, which serves as the human translation benchmark throughout the analysis (Freitag et al., 2021: 8).

The evaluation is grounded in three complementary theoretical frameworks: Newmark's (1988) semantic and communicative translation theory, Nida and Taber's (1969) dynamic equivalence, and Skopos theory (Reiss & Vermeer, 1984). Drawing on the Multidimensional Quality Metrics framework (Freitag et al., 2021: 3), translation quality is assessed across five criteria terminological precision, syntactic accuracy, pragmatic adequacy, ambiguity resolution, and cultural sensitivity.

IV. Practical section

1. Corpus Presentation

The corpus selected for this study consists of the patient information leaflet (PIL) of POLYDEXA® ear drops, an imported pharmaceutical product manufactured in France and marketed by BOUCHARA-RECORDATI. The leaflet is officially available in three languages, French, English, and Arabic. Although the original document was produced in French, the present study focuses exclusively on the Arabic and English versions.

The official Arabic version serves as the source text, while the official English version is used as the reference translation. The Arabic text was translated into English using three AI-based translation systems :Google Translate, DeepL, and ChatGPT. The AI-generated translations were then compared with the official English translation provided in the original multilingual leaflet.

The leaflet was selected because it represents an authentic pharmaceutical document containing specialized medical terminology, dosage instructions, contraindications, precautions, and patient guidance. Moreover, officially validated Arabic–English pharmaceutical leaflets remain relatively scarce in the Algerian context, making this corpus particularly suitable for a comparative evaluation of AI-generated medical translations.

The complete patient information leaflet is provided in Appendix (Figures 1-2)

2. Sample Description

Parameter	Details
Brand name	POLYDEXA®
Therapeutic indication	Topical treatment of certain forms of otitis
Marketing authorization holder	BOUCHARA-RECORDATI, Puteaux, France
Manufacturer	PHARMASTER, Erstein, France
Languages of leaflet	Arabic / French / English
Corpus provenance	Official printed leaflet obtained from an Algerian pharmacy

3. Sections Selected for Analysis

To ensure a systematic comparison, the analysis was organized into four sections of the patient information leaflet: composition, indications and contraindications, and warnings. These sections were selected because they encompass the principal linguistic and translational challenges encountered in medical translation, including pharmaceutical terminology, chemical nomenclature, standardized abbreviations, polysemous medical expressions, procedural instructions, and risk-related communication. This analysis enabled a comprehensive evaluation of the AI-generated translations across different textual functions and communicative contexts.

4. Translation Procedure

Each of the three selected sections was translated from Arabic into English using Google Translate, DeepL, and ChatGPT. The AI-generated translations were subsequently compared with the official English version

of the leaflet, which served as the reference translation. The comparative analysis focused on the translation of pharmaceutical terminology, chemical nomenclature, standardized abbreviations, procedural language, and clinically sensitive expressions. Translation quality was assessed according to the evaluation framework presented in the methodology, with particular attention to terminological precision, syntactic accuracy, pragmatic adequacy, ambiguity resolution, and cultural sensitivity.

4.1 Table 1: Composition

Source Term	Target Term	Google Translate	DeepL	ChatGPT
كبريتات النيوميسين	Neomycin **sulphate**	Neomycin **sulfate**	Neomycin **sulfate**	Neomycin **sulfate**
كبريتات البوليميكسينB	Polymyxin B **sulphate**	Polymyxin B **sulfate**	Polymyxin B **sulfate**	Polymyxin B **sulfate**
ديكساميثازون ميناسولفوبنزوات الصودي	Dexamethasone sodium **metasulphobenzoate**	Dexamethasone sodium **sulfonate**	Dexamethasone sodium **minosulfobenzate**	Dexamethasone sodium **metasulfobenzoate**
ثيوميرسال	**Thiomersal**	**Timoresal**	**Thimerosal**	**Thiomersal**
حمض السيتريك	Citric acid	Citric acid	Citric acid	Citric acid
ماكروغول 400	Macrogol 400	Macrogol 400	Macrogol 400	Macrogol 400
بوليسوربات 80	Polysorbate 80	Polysorbate 80	Polysorbate 80	Polysorbate 80

Source Term	Target Term	Google Translate	DeepL	ChatGPT
وحدة دولية	**IU**	IU	IU	IU
كمية كافية لغاية 100 مل	**q.s.** 100 ml	sufficient quantity to 100 ml	sufficient to make up to 100 mL	sufficient quantity up to 100 ml

Source: Official POLYDEXA® patient information leaflet and AI-generated translations produced by Google Translate, DeepL, and ChatGPT.

The Composition section constitutes the most terminology-sensitive part of the patient information leaflet, as it contains standardized pharmaceutical terminology and chemical nomenclature that require absolute terminological precision. Unlike patient-oriented information, the names of active ingredients are internationally standardized and do not permit paraphrase or lexical approximation. According to Newmark (1988), semantic translation aims to preserve the precise contextual meaning of specialized terminology, making semantic accuracy the primary criterion for evaluating translations of pharmaceutical compounds.

The results indicate that the three AI systems successfully translated most standardized pharmaceutical terms, including **Neomycin sulphate, Polymyxin B sulphate, Citric acid, Macrogol 400, Polysorbate 80, and the abbreviation IU**. The only observable differences involve British (sulphate) and American (sulfate) spelling conventions, which do not affect the scientific meaning and therefore should not be regarded as translation errors.

The greatest divergence appears in the rendering of ديكساميثازون ميناسولفوبنزوات الصودي. Google Translate translated the compound as **Dexamethasone sodium sulfonate**, replacing the complete chemical denomination with a generic chemical group and consequently altering the identity of the active ingredient. DeepL generated **Dexamethasone sodium minosulfobenzate,** a denomination that does not correspond to recognized pharmaceutical nomenclature. In contrast, ChatGPT produced **Dexamethasone sodium metasulfobenzoate**, differing from the official translation only in the American spelling of sulf- instead of the British sulph-. From the perspective of Newmark's semantic translation theory, the outputs of Google Translate and DeepL fail to preserve the semantic specificity of the source term, whereas ChatGPT successfully maintains semantic equivalence despite orthographic variation (Newmark, 1988).

These findings also support Forcada's (2017: 298) observation that neural machine translation systems often generate linguistically fluent outputs that may nevertheless be terminologically inaccurate in highly specialized domains. Fluency alone, therefore, cannot be considered an indicator of translation quality when dealing with pharmaceutical nomenclature.

Another noteworthy finding concerns the preservative **Thiomersal.** Google Translate rendered the term as **Timoresal,** producing a non-existent pharmaceutical denomination. DeepL generated **Thimerosal**, the accepted American nomenclature, whereas ChatGPT reproduced the British form used in the official leaflet. Although DeepL and ChatGPT differ

orthographically, both preserve the identity of the chemical compound and remain pharmaceutically acceptable.

A further difference concerns the expression **كمية كافية لغاية 100 مل**. The official leaflet employs the standardized pharmaceutical abbreviation **q.s.** (quantum satis), while all three AI systems replaced this notation with explanatory phrases. Although the semantic meaning was preserved, none of the systems retained the conventional documentary practice used in pharmaceutical documentation. This finding confirms that AI systems tend to prioritize semantic explicitation over adherence to professional documentary conventions, a tendency also discussed by Khoury (2024) in the context of scientific translation. From the perspective of Skopos theory, these renderings remain functionally understandable for patients but are less appropriate for professional pharmaceutical documentation, where conformity to established conventions constitutes an essential component of translation quality (Reiss & Vermeer, 1984).

➢ **Table 1: Scoring results for section 1**

Criterion	Reference Translation	Google Translate	DeepL	ChatGPT
Terminological Precision	3/3	1/3	1/3	2/3
Syntactic Accuracy	3/3	3/3	3/3	3/3
Pragmatic Adequacy	3/3	2/3	2/3	2/3
Ambiguity Resolution	3/3	2/3	2/3	3/3
Cultural Sensitivity	3/3	3/3	3/3	3/3
Section Total /15	**15/15**	**11/15**	**11/15**	**13/15**

The scoring results confirm the qualitative findings outlined above. All three systems achieve full marks on syntactic accuracy and cultural sensitivity, reflecting the maturity of current AI Translation Systems at the level of grammatical structure. The most pronounced deficiency appears on the criterion of terminological precision, where both Google Translate and DeepL score 1/3 reflecting the critical chemical nomenclature errors identified in the analysis. ChatGPT's score of 2/3 on this criterion reflects its near-accurate rendering of the complex active substance name, penalized only for orthographic variation. These results are consistent with recent studies demonstrating that large language models generally outperform conventional AI-based systems in specialized medical translation, while still requiring expert verification in high-risk medical contexts (Algaraady et al., 2025; Tercedor-Sánchez, 2025).

4.2 Table 2 :Indications and Contraindications

Source Text	Target Text	Google Translate	DeepL	ChatGPT
يُستعمل كعلاج موضعي	topical treatment	topical treatment	topical treatment	topical treatment
التهابات الأذن	otitis	ear infections	ear infections	ear infections
!انتبه	CAUTION!	Caution!	Warning!	*(omitted)*
مجموعة الأمينوغليكوزيدات	aminoside **family**	aminoglycoside **antibiotic**	aminoglycoside **class**	aminoglycoside **group**
طبلة الأذن	**tympanic membrane**	**eardrum**	**eardrum**	**eardrum**

Source Text	Target Text	Google Translate	DeepL	ChatGPT
مصابة أو مثقوبة	damaged or perforated	injured or perforated	damaged or perforated	damaged or perforated
عدوى فيروسية على مستوى الأذن	viral ear infection	viral ear infection	viral infection in the ear	viral infection of the ear
في حالة الشك	If you are in any doubt	if in doubt	If in doubt	If in doubt

Source: Official POLYDEXA® patient information leaflet and AI-generated translations produced by Google Translate, DeepL, and ChatGPT.

The table reveals three recurrent translation patterns: register shift, lexical explicitation, and pragmatic information loss. Unlike the composition section, where translation accuracy depends primarily on standardized terminology, this section requires translators to balance terminological precision with communicative clarity while preserving the regulatory function of pharmaceutical documentation.

The first pattern concerns the translation of **التهابات الأذن**. The official English version uses the specialized medical term **otitis**, whereas all three AI systems translate the expression as **ear infections**. Although the latter accurately conveys the medical meaning, it represents a shift from technical to patient-oriented language. According to Newmark (1988: 39-40), this reflects a communicative translation strategy that prioritizes readability over the preservation of specialized terminology. From the perspective of Skopos theory, however, the appropriateness of this choice depends on the

communicative purpose of the translation. While **ear infections** may facilitate patient comprehension, **otitis** remains more consistent with the documentary style of officially approved pharmaceutical leaflets (Reiss & Vermeer, 1984:58). A similar tendency appears in the translation of **طبلة الأذن**, where all three systems produce **eardrum** instead of the more technical **tympanic membrane**. This choice should therefore be interpreted as a register shift rather than a translation error.

The second finding concerns lexical explicitation. Google Translate renders **مجموعة الأمينوغليكوزيدات** as **aminoglycoside antibiotic**, introducing the lexical item antibiotic, which is absent from the source text. Although factually correct, this addition reflects an inferential expansion rather than a faithful translation of the original expression. In regulated medical documents, where every lexical element forms part of the approved text, such additions may reduce documentary fidelity. DeepL (**aminoglycoside class**) and ChatGPT (**aminoglycoside group**) avoid this unnecessary expansion and remain closer to the source formulation, although neither reproduces the official translation (**aminoside family**).

The most significant pragmatic difference concerns the translation of the **warning** marker **انتبه**!. The official leaflet renders this expression as **CAUTION!**, highlighting the transition from general information to clinically important contraindications. Google Translate (**Caution!**) and DeepL (**Warning!**) preserve this communicative function, despite minor stylistic differences. ChatGPT, however, omits **the marker** entirely, resulting in the loss of an important textual signal. According to Nida and Taber (1969:24), effective translation should produce an equivalent

communicative effect on the target reader. The omission therefore weakens the pragmatic impact of the warning by removing an explicit alert that guides readers' attention to critical safety information.

Minor variations are also observed in several expressions. Google Translate renders مصابة أو مثقوبة as injured or perforated, whereas the official translation uses **damaged or perforated**. Although both versions remain understandable, **damaged** is the more appropriate medical term in this clinical context. Likewise, DeepL (**viral infection in the ear**) and ChatGPT (**viral infection of the ear**) differ syntactically from the official translation (**viral ear infection**), yet all three formulations remain medically acceptable. Similarly, the shorter expression **If in doubt** produced by DeepL and ChatGPT conveys the same meaning as the official If you are in any doubt and therefore represents stylistic variation rather than a translation deficiency.

Overall, the findings indicate that AI systems generally preserve the semantic content of indications and contraindications but frequently adjust the linguistic register to improve readability. While such adaptations may enhance patient accessibility, they do not always conform to the documentary conventions of officially approved pharmaceutical leaflets. Among the three systems evaluated, ChatGPT and DeepL produced the most consistent translations overall, whereas Google Translate showed a greater tendency toward lexical expansion. Nevertheless, ChatGPT's omission of a critical warning marker demonstrates that pragmatic information remains particularly vulnerable in AI-generated medical translations, confirming recent observations that human revision remains

essential for high-risk medical texts (Tercedor-Sánchez, 2025; Algaraady et al., 2025).

➢ **Table 2: Scoring results for section 2**

Criterion	Human	Google Translate	DeepL	ChatGPT
Terminological Precision	3/3	1/3	2/3	2/3
Syntactic Accuracy	3/3	3/3	3/3	3/3
Pragmatic Adequacy	3/3	2/3	2/3	1/3
Ambiguity Resolution	3/3	1/3	2/3	2/3
Cultural Sensitivity	3/3	3/3	3/3	3/3
Section Total /15	**15/15**	**10/15**	**12/15**	**11/15**

The scoring results reflect the three principal failure patterns identified in the analysis. On terminological precision, Google Translate scores 1/3 penalized for the unauthorized addition of *antibiotic* and the substitution of *otitis* with the lay term *ear infections* while DeepL and ChatGPT both score 2/3, reflecting acceptable but imprecise terminological choices. Syntactic accuracy again achieves full marks across all three systems, confirming the consistent grammatical competence of AI-based translation systems. The most significant divergence appears on the criterion of pragmatic adequacy: ChatGPT scores only 1/3 the lowest score recorded across any system on any criterion in this section a direct consequence of its complete omission of the CAUTION ! marker, which removes a pragmatically critical element from the translated document.

DeepL achieves the highest section score (12/15), demonstrating greater overall reliability than either Google Translate or ChatGPT in this section.

4.3 Table 3: Special Warnings

Source Text	Target Text	Google Translate	DeepL	ChatGPT
طبلة الأذن	tympanic membrane	eardrum	eardrum	eardrum
سيلان قيحي	Purulent discharge from the ear	drainage	discharge / pus	purulent discharge
تأثيرات غير مرغوبة لا انعكاسية	irreversible undesirable effects	undesirable irreversible effects	irreversible undesirable effects	irreversible adverse effects
الصمم	Loss of hearing	deafness	hearing loss	deafness
اضطرابات في التوازن	Impaired balance	balance problems	balance disorders	balance disorders
رد فعل تحسسي	Systemic allergic reaction	allergic reaction	allergic reaction	allergic reaction
يجب إيقاف المعالجة	Treatment should be stopped	Treatment should be stopped immediately	Treatment must be discontinued	Treatment must be stopped immediately
للحد من التلوث	to reduce the risk of infection	to minimize contamination	to minimize contamination	to minimize contamination

Source Text	Target Text	Google Translate	DeepL	ChatGPT
تجنب لمس طرف القارورة	avoid touching the tip of the dropper	avoid touching the tip of the bottle	avoid touching the tip of the bottle	avoid touching the tip of the bottle

Source: Official POLYDEXA® patient information leaflet and AI-generated translations produced by Google Translate, DeepL, and ChatGPT.

The analysis of this section reveals three principal translation patterns: **register shift**, **terminological generalization**, and **stylistic variation**. Unlike the composition section, where accuracy depends primarily on standardized nomenclature, this section requires translators to preserve both the clinical meaning of medical expressions and the communicative function of safety instructions, a dual demand that directly engages the evaluative criteria of all three theoretical frameworks applied in this study.

The first pattern concerns the rendering of anatomical terminology. The official leaflet employs **tympanic membrane** as the clinical designation for **طبلة الأذن**, whereas all three AI systems produce the lay equivalent ***eardrum***. Although both terms refer to the same anatomical structure, **tympanic membrane** is the preferred denomination in medical and regulatory documentation, while ***eardrum*** is oriented toward non-specialist readability. In Newmark's (1988 :39) terms, this substitution reflects a communicative translation strategy that privileges accessibility over terminological specificity a strategy that, while appropriate in patient

education materials, departs from the formal documentary register of a regulated pharmaceutical leaflet. From the perspective of Skopos theory, the appropriateness of this choice depends entirely on the intended function of the translated text: if the *skopos* is patient comprehension, ***eardrum*** may serve the communicative purpose; if it is regulatory compliance and professional documentation, it represents a register failure (Reiss & Vermeer, 1984 :58). That all three AI systems converge on the same lay equivalent confirms this as a structural tendency rather than a system-specific choice consistent with Forcada's (2017 :298) observation that AI-based systems are architecturally biased toward fluency and naturalness over domain-specific register precision.

The second observation concerns the translation of سيلان قيحي, which represents one of the most clinically significant findings of the analysis. The official translation renders the expression as purulent discharge from the ear, a formulation that accurately reproduces both the pathological nature and the anatomical location of the source expression. ChatGPT most closely approximates the reference with ***purulent discharge***, correctly capturing the suppurative nature of the sign but omitting the locative element **from the ear** an omission that, in the clinical context of a warning about tympanic membrane perforation, reduces the diagnostic specificity of the instruction. Google Translate reduces the expression **to drainage** a generic, process-oriented term that conveys neither the pathological character nor the anatomical specificity of the source, and that could easily be misread as a reference to routine post-operative **drainage** rather than a pathological sign requiring immediate

medical attention. DeepL produces **discharge / pus** a hybrid formulation that partially conveys the suppurative meaning but adopts a non-standard presentational format inconsistent with the conventions of pharmaceutical documentation (Montalt & González Davies, 2007: 45). These findings confirm that clinically specific terminology remains among the most challenging translation problems for AI systems, particularly when specialized expressions require both lexical precision and contextual interpretation precisely the combination of competences that Nida and Taber (1969: 24) identify as essential for achieving dynamic equivalence in specialized medical communication.

الصممis officially rendered **as loss of hearing** a formulation that, interestingly, aligns with the broader clinical concept rather than the more categorical *deafness* that the Arabic term strictly denotes. Google Translate and ChatGPT both produce *deafness* semantically closer to the Arabic source but departing from the communicative choice of the human reference. DeepL reproduces hearing loss the closest approximation to the official translation and the term most widely used in clinical otology to encompass a spectrum of auditory impairment (Montalt & González Davies, 2007: 89). This variation illustrates a genuine translational tension between semantic fidelity to the Arabic source and communicative equivalence with the official benchmark a tension that Newmark's (1988: 39) semantic-communicative distinction directly addresses.

رد فعل تحسسيis officially rendered as *systemic allergic reaction* adding the qualifier *systemic* that is absent from the Arabic source, a deliberate explicitation that specifies the scope of the immune response

and distinguishes it from localized reactions. All three AI systems produce *allergic reaction*, omitting *systemic* and thereby narrowing the clinical scope of the warning. This omission reduces the safety information available to the patient: a reader of the AI-generated text would not be alerted to the possibility of a body-wide immune response, a clinically significant distinction in pharmacovigilance contexts (Montalt & González Davies, 2007: 89). From Nida and Taber's (1969:24) perspective, the AI outputs fail to produce a dynamically equivalent effect on the target reader, whose comprehension of the risk is diminished by the absence of this qualifier.

تأثيرات غير مرغوبة لا انعكاسيةis officially rendered as *irreversible undesirable effects* a formulation that DeepL reproduces with complete fidelity, making it the most accurate rendering among the three systems on this item. Google Translate inverts the adjective order to produce *undesirable irreversible effects* a minor syntactic variation that does not alter the semantic content but departs from the documentary precision of the official text. ChatGPT substitutes *undesirable* with *adverse* a clinically common adjective in medical English that, while acceptable in general medical communication, represents a register-specific choice that moves away from the exact formulation of the regulatory document. Although none of these variations is clinically catastrophic in isolation, their cumulative effect combined with the universal absence of the pharmacotoxicological qualifier *ototoxic* from all three outputs significantly reduces the clinical specificity of the warning as a whole.

Several stylistic variations appear in the procedural safety instructions of this section. The rendering of **يجب إيقاف المعالجة**as *Treatment should be stopped* in the human reference employing the advisory modal *should*is modified by all three AI systems: Google Translate and ChatGPT add the adverb *immediately*, intensifying the urgency of the instruction beyond what the source text warrants, while DeepL substitutes *must* for *should*, shifting the register from advisory to obligatory. Although these modifications may appear safety-enhancing, they alter the precise deontic calibration of the official document a regulatory text in which modal choices are not arbitrary but reflect the considered judgment of the originating pharmaceutical authority (Reiss & Vermeer, 1984:58).

The rendering of **للحد من التلوث**reveals a further significant divergence: the official translation produces *to reduce the risk of infection* a communicative explicitation that interprets *contamination* in terms of its clinical consequence whereas all three AI systems produce *to minimize contamination*, remaining closer to the literal meaning of the source text. This is one of the rare instances in this corpus where AI outputs are more formally faithful to the source text than the human reference, which has applied a Skopos-oriented communicative strategy of explicitation to enhance patient comprehension (Nine et al., 2025: 7).

Finally, the official translation of **تجنب لمس طرف القارورة**as *avoid touching the tip of the dropper* employing the pharmacologically precise term *dropper* to designate the dispensing component is rendered by all three AI systems as *avoid touching the tip of the bottle*, substituting the generic *bottle* for the specific *dropper*. While both formulations remain

communicatively intelligible, the official translation is pharmacologically more precise: the instruction specifically concerns the dropper tip the component in direct contact with the ear canal rather than the bottle as a whole, a distinction that is clinically relevant in the context of contamination prevention.

To conclude, the findings of this section indicate that the AI systems produce fluent and generally intelligible translations; however, they consistently simplify specialized medical terminology, omit clinically relevant qualifiers, and introduce stylistic modifications that reduce documentary equivalence with the official leaflet. These findings support the three theoretical frameworks applied in this study: Newmark's (1988, : 39) emphasis on semantic precision in specialized translation is vindicated by the AI systems' systematic register displacement and terminological generalization; Nida and Taber's (1969:24) dynamic equivalence principle is confirmed as the standard against which AI outputs most consistently fall short, particularly in their failure to reproduce the full clinical information content of the safety warnings; and Skopos theory (Reiss & Vermeer, 1984:58) highlights the fundamental limitation of AI systems that operate without explicit knowledge of the communicative purpose and audience profile of the target text the very knowledge that distinguishes professional medical translation from machine-generated output.

➢ **Table 3: Scoring results for Section 3**

Criterion	Reference Translation	Google Translate	DeepL	ChatGPT
Terminological Precision	3/3	1/3	2/3	2/3
Syntactic Accuracy	3/3	3/3	3/3	3/3
Pragmatic Adequacy	3/3	1/3	2/3	2/3
Ambiguity Resolution	3/3	1/3	3/3	2/3
Cultural Sensitivity	3/3	3/3	3/3	3/3
Section Total /15	**15/15**	**9/15**	**13/15**	**12/15**

The scoring results of this section record the lowest performance for Google Translate across the entire analysis (9/15 - 60%), reflecting the cumulative weight of its failures: the reductive rendering of purulent discharge from the ear as drainage, the universal omission of systemic from systemic allergic reaction, and the consistent register displacement from clinical to lay vocabulary. DeepL achieves the highest section score (13/15-86.7%), distinguished by its accurate reproduction of **irreversible undesirable effects** and **hearing loss**, and its avoidance of the unauthorized additions and phonetic distortions observed in Section 1. ChatGPT scores 12/15 (80%), demonstrating consistent performance across criteria but penalized for its substitution of **adverse for undesirable** and its replacement of **dropper with bottle**. Syntactic accuracy and cultural sensitivity again achieve full marks across all three systems a pattern consistent with the findings of Sections 1 and 2 and confirming AI-based translation systems have matured at the level of grammatical

structure while remaining systematically deficient in the terminological and pragmatic dimensions most critical for patient safety.

➢ **Table 4: Global Scoring Summary**

Section	Reference translation	Google Translate	DeepL	ChatGPT
Section 1: Composition	15/15	11/15	11/15	13/15
Section 2 :Indications & Contraindications	15/15	10/15	12/15	11/15
Section 3 :Special Warnings	15/15	9/15	13/15	12/15
TOTAL /45	**45/45**	**30/45**	**36/45**	**36/45**
Percentage	**100%**	**66.7%**	**80%**	**80%**

The global scoring summary confirms the performance hierarchy identified across all three sections. DeepL and ChatGPT achieve identical total scores (36/45- 80%), while Google Translate records the lowest overall performance (30/45-66.7%). However, these aggregate figures mask important qualitative differences in error profiles that the cross-system typology presented in Table 5 addresses in greater detail. Across all three systems and all three sections, **syntactic accuracy** and **cultural sensitivity** consistently achieve full marks confirming that current AI translation technology has reached a level of grammatical and cultural competence that makes its output superficially convincing. The most pronounced deficiencies appear on **terminological precision** and **pragmatic adequacy** precisely the dimensions most consequential for patient safety in medical translation contexts.

➢ **Table 5: Cross-System Error Typology**

Error Type	Google Translate	DeepL	ChatGPT
Chemical nomenclature error	sodium sulfonate **x**	minosulfobenzate **x**	metasulfobenzoate ~
Register displacement (tympanic membrane ⟶ eardrum)	eardrum **x**	eardrum **x**	eardrum **x**
Unauthorized lexical addition	Antibiotic **x**	— ✓	— ✓
Pragmatic omission (CAUTION!)	Caution! ~	Warning! ~	(omitted) **x**
Absence of pharmaceutical abbreviation (q.s.)	descriptive phrase **x**	descriptive phrase **x**	descriptive phrase **x**
Phonetic distortion of proper name	Timoresal **x**	Thimerosal ~	Thiomersal ✓
Omission of clinical qualifier (systemic)	allergic reaction **x**	allergic reaction **x**	allergic reaction **x**
Modal shift in safety instruction	immediately added ~	must substituted ~	must + immediately ~
Device term substitution (dropper ⟶ bottle)	bottle **x**	bottle **x**	bottle **x**
Reductive rendering of clinical sign (drainage)	drainage **x**	discharge/pus ~	purulent discharge ~

Legend: ✓ Accurate - ~ Partially acceptable - **x** Inaccurate

V. Discussion

The findings of this study demonstrate that AI-based translation systems are capable of producing fluent and grammatically accurate translations of pharmaceutical texts. However, the comparative analysis also reveals recurring limitations in preserving specialized medical terminology, maintaining the appropriate documentary register, and reproducing the pragmatic functions of pharmaceutical instructions. These findings are consistent with previous studies, which suggest that although recent AI systems have considerably improved translation quality, professional revision remains essential in medical contexts (Forcada, 2017, p. 298; Ning et al., 2024:3; Tercedor-Sánchez, 2025: 5).

The quantitative results presented in Table 4 provide a clear picture of the overall performance of the three systems evaluated. DeepL and ChatGPT achieved identical total scores (36/45 - 80%), while Google Translate recorded the lowest overall performance (30/45-66.7%). Although these aggregate figures suggest comparable performance between DeepL and ChatGPT, the cross-system error typology presented in Table 5 reveals that the two systems exhibit distinct error profiles: DeepL shows greater weakness in chemical nomenclature as illustrated by its spurious rendering **minosulfobenzate** while ChatGPT demonstrates a more pronounced tendency toward pragmatic omission, most consequentially illustrated by its complete deletion of the *CAUTION!* marker in Section 2. Google Translate's lower overall score reflects the cumulative weight of more

severe errors, including the chemically inaccurate **sodium sulfonate**, the phonetically distorted ***Timoresal,*** and the unauthorized addition of *antibiotic* errors that go beyond register displacement to constitute substantive terminological failures with potential clinical consequences.

From the perspective of Newmark's (1988: 39) semantic and communicative translation framework, the AI systems frequently favored communicative solutions by replacing specialized terminology with more accessible expressions **eardrum** instead of **tympanic membrane, ear infections** instead of **otitis, allergic reaction** instead of **systemic allergic reaction**. While such choices improve readability, they also reduce terminological specificity and move away from the documentary conventions generally observed in pharmaceutical leaflets. This systematic bias toward communicative translation in contexts demanding semantic precision constitutes the most pervasive and structurally significant limitation identified across all three systems.

The analysis further illustrates that AI systems generally preserve the propositional meaning of the source text but occasionally simplify or modify clinically relevant information. This tendency is evident in the omission of the qualifier *systemic* from **systemic allergic reaction**, the replacement of **tip of the dropper** with **tip of the bottle**, and the modal modifications introduced into precautionary instructions. Although these variations do not necessarily alter the overall meaning, they reduce the degree of semantic precision expected in regulated

medical documentation and may in some cases, diminish the patient's comprehension of the risk involved.

From the perspective of Nida and Taber's (1969:24) dynamic equivalence, most AI-generated translations successfully communicate the general message of the source text. Nevertheless, several examples demonstrate that linguistic fluency alone does not guarantee functional equivalence. The omission of *systemic* from the warning about allergic reactions, the reductive rendering of *purulent discharge from the ear* as *drainage* by Google Translate, and ChatGPT's deletion of *CAUTION!* all illustrate instances where the AI-generated output fails to produce an equivalent communicative effect on the target reader — precisely the standard that Nida's framework requires. From the perspective of Skopos theory (Reiss & Vermeer, 1984:58), these failures reflect the fundamental limitation of AI systems that operate without explicit knowledge of the communicative purpose or audience profile of the target text: a patient information leaflet is not merely a linguistic object but a regulatory instrument whose *skopos* enabling patients to use medicines safely places specific and non-negotiable demands on translation quality that current AI systems cannot consistently meet.

The analysis also highlights that the official English leaflet, although adopted as the reference translation, should not be regarded as an absolute standard. In some instances, AI-generated translations were equally acceptable or even closer to the Arabic source text most notably in the rendering of **للحد من التلوث**, where all three systems produced *to* **minimize contamination**, remaining more formally

faithful to the source than the human reference's communicative explicitation *to* **reduce the risk of infection**. This observation reinforces the view that translation quality cannot be assessed solely through lexical correspondence with a reference version, but must also consider terminological accuracy, communicative purpose, and the regulatory conventions governing pharmaceutical documentation.

These findings carry specific implications for the Algerian healthcare context in which this study is situated. The near-total absence of Arabic-English bilingual pharmaceutical documentation in the Algerian market evidenced by the difficulty of locating a trilingual leaflet in the course of corpus construction means that AI translation tools are likely to be deployed in this context without the benefit of professionally validated reference translations against which their output can be checked. In such conditions, the errors identified in this study particularly the reductive rendering of clinical signs, the omission of safety qualifiers, and the register displacement from clinical to lay vocabulary pose a direct and measurable risk to patient safety (Al Shamsi et al., 2020: 3; Brewster et al., 2025: 4).

Taken together, these findings suggest that AI translation systems should be regarded as translation-support tools rather than autonomous translators. Although they produce fluent translations, AI-generated outputs should be systematically reviewed by qualified medical translators to ensure terminological accuracy, functional adequacy, and patient safety. The human-in-the-loop model in which AI-generated output serves as a first draft subject to expert revision

represents the most responsible framework for integrating AI translation tools into Arabic medical translation workflows (Brewster et al., 2025:6; Anyaegbuna et al., 2026: 3).

VI. Conclusion

This study investigated the performance of three AI-based translation systems -Google Translate, DeepL, and ChatGPT- in translating Arabic pharmaceutical patient information leaflets into English. Drawing on the theoretical frameworks of Newmark (1988), Nida and Taber (1969), and Reiss and Vermeer's (1984) Skopos theory, the study adopted a comparative analytical approach to examine the extent to which current AI systems satisfy the linguistic and functional requirements of medical translation. Using the trilingual POLYDEXA® patient information leaflet as the analytical corpus, the study provides empirical evidence from the relatively underexplored Arabic–English language pair within the Algerian pharmaceutical context.

The findings demonstrate that all three AI systems produce translations that are generally fluent and grammatically accurate. Nevertheless, the analysis reveals recurring limitations in the translation of specialized medical terminology, the preservation of clinically relevant information, and the reproduction of the pragmatic functions of pharmaceutical documentation. Terminological simplification, omission of essential qualifiers, and modifications of procedural instructions were observed across the three systems, although to varying degrees. Overall, ChatGPT achieved the highest performance, followed by DeepL, while

Google Translate displayed the greatest variability in terminological accuracy and contextual interpretation.

From a theoretical perspective, the findings reinforce the continued relevance of Newmark's semantic translation, Nida's dynamic equivalence, and Skopos theory in evaluating AI-generated medical translations. Although current AI systems have made remarkable progress in producing natural and coherent target texts, linguistic fluency alone cannot guarantee translation quality in specialized medical contexts. Medical translation requires semantic precision, functional adequacy, and strict adherence to established medical terminology and regulatory conventions dimensions in which AI systems continue to demonstrate important limitations.

The study also carries practical implications for the integration of AI into professional medical translation. Rather than replacing translators, AI systems should be regarded as translation-support tools whose output requires systematic review by qualified medical translators before clinical or regulatory use. This approach offers the benefits of increased efficiency while maintaining the level of accuracy required for patient safety and pharmaceutical communication.

Finally, this study contributes to the growing literature on AI-assisted medical translation by providing a detailed evaluation of Arabic–English pharmaceutical translation within the Algerian context. Future research should extend the corpus to additional pharmaceutical products and medical genres, investigate the impact of domain-specific prompting and post-editing strategies, and evaluate Arabic-specialized language models

trained on medical corpora. Such research will contribute to the development of more reliable AI-assisted translation systems capable of meeting the linguistic, functional, and regulatory requirements of Arabic medical translation.

References

Ahmed, K., Hussain, F., & Iqbal, R. (2025). Global health communication: Linguistic strategies for translating English medical texts into Urdu. *Forum for Linguistic Studies*.

Algaraady, J., Mahyoob, M., & Alkadi, H. (2025). Exploring ChatGPT's potential for augmenting post-editing in machine translation across multiple domains. *Frontiers in Artificial Intelligence*.

Al-Khalifa, H., Al-Saud, L., & Mahmoud, A. (2024). Error analysis of pretrained language models (PLMs) in English-to-Arabic machine translation. *Human-Centric Intelligent Systems*.

Al Shamsi, H. S., Almutairi, A. G., Al Mashrafi, S., & Al Kalbani, T. (2020). Implications of language barriers for healthcare: A systematic review. *Oman Medical Journal*.

Alzain, E., Algarni, S., & Aldosari, F. (2024). The quality of Google Translate and ChatGPT English to Arabic translation: The case of scientific text translation. *Forum for Linguistic Studies*.

Anyaegbuna, C., et al. (2026). Artificial intelligence translation in healthcare: An urgent call for evidence-informed policy frameworks. *BMJ Health & Care Informatics*.

Benamar, N., & Temmam, M. (2024). Use of patient information leaflets in the Algerian context: Current situation and regulatory framework. *Journal of Specialised Translation, 41*, 1–22.

Berrichi, S., Mazroui, A., & Lakhouaja, A. (2021). Addressing limited vocabulary and long sentences constraints in English–Arabic neural machine translation. *Arabian Journal for Science and Engineering*.

Block, P., Schmidt, F., & Becker, T. (2025). Quality of machine translations in medical texts: An analysis based on standardised evaluation metrics. *Studies in Health Technology and Informatics*.

Bouchara-Recordati. (2024). *POLYDEXA® ear drops: Patient information leaflet* [Official trilingual printed leaflet]. PHARMASTER.

Brewster, R. C., et al. (2025). Evaluating human-in-the-loop strategies for artificial

intelligence-enabled translation of patient discharge instructions. *NPJ Digital Medicine*.

Díaz-Millón, M., Pena, D., & Rivas, R. (2025). Systematic meta-review on migrant healthcare access: Language barriers and the role of translation. *Journal of Migration and Health*.

European Medicines Agency. (2023). *Guideline on the readability of the labelling and package leaflet of medicinal products for human use* (Rev. 1). EMA/CEG. https://www.ema.europa.eu

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces, 6*(2), 291–309. https://doi.org/10.1075/ts.6.2.06for

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

Genovese, A., et al. (2024). Artificial intelligence in clinical settings: A systematic review of its role in language translation and interpretation. *Annals of Translational Medicine*.

Khoury, O. (2024). Reflection of explicitation in scientific translation: Neural machine translation vs. human post-editing. *Journal of Language Teaching and Research*.

Kreienbrinck, A., Roth, P., & Hoffmann, M. (2025). Usability of technological tools to overcome language barriers in healthcare: A scoping review. *Archives of Public Health*.

Li, Y. (2021). Nida's translation theory of "functional equivalence" and its application in Chinese herbal medicine translation.

Mohammad, R., et al. (2024). Optimizing large language models for Arabic healthcare communication: A focus on patient-centered NLP applications. *Big Data and Cognitive Computing*.

Mohammed, T. A. S. (2025). Evaluating translation quality: A qualitative and quantitative assessment of machine and LLM-driven Arabic-English

translations. *Information, 16*(2), 87.

Montalt, V., & González Davies, M. (2007). *Medical translation step by step: Learning by drafting*. St. Jerome Publishing.

Munday, J. (2016). *Introducing translation studies: Theories and applications* (4th ed.). Routledge.

Nassar, H. (2025). Challenges of post-editing in English to Arabic machine translation of technical texts. *International Journal of Linguistics, Literature and Translation*.

Nedainova, I. (2021). Skopos theory in the light of functional translation. *Alfred Nobel University Journal of Philology*.

Newmark, P. (1988). *A textbook of translation*. Prentice Hall.

Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation*. Brill.

Nine, H., Al-Harahsheh, A., & Sayaheen, B. (2025). A functional approach to medical translation: A Skopos theory-based study. *Journal of Languages and Translation*.

Ning, J., Liu, X., & Zhao, Y. (2024). On the role of expertise in applying MTPE mode in medical translation. *Journal of Humanities, Arts and Social Science*.

Pandey, M., Maina, R. G., Amoyaw, J., Li, Y., Kamrul, R., Michaels, C. R., & Maroof, R. (2021). Impacts of English language proficiency on healthcare access, use, and outcomes among immigrants. *BMC Health Services Research, 21*, 741.

Peng, T. (2026). Literal and free translation in medical translation: Applications of Newmark's theory and ethical framework. *Lecture Notes on Language and Literature*.

Reiss, K., & Vermeer, H. J. (1984). *Groundwork for a general theory of translation*. Niemeyer.

Rico Pérez, C. (2024). Re-thinking machine translation post-editing guidelines. *The Journal of Specialised Translation*.

Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation*.

Tercedor-Sánchez, M. (2025). Risk in AI-mediated medical translation. *INContext:*

Studies in Translation and Interculturalism.

Wang, Z. (2018). Exploring Newmark's communicative translation and text typology.

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Jaoua, M. (2021). Arabic machine translation: A survey with challenges and future directions. *IEEE Access*.

Ziganshina, L., et al. (2021). Assessing human post-editing efforts to compare the performance of three machine translation engines for English to Russian translation of Cochrane plain language health information. *Informatics*.

Appendix

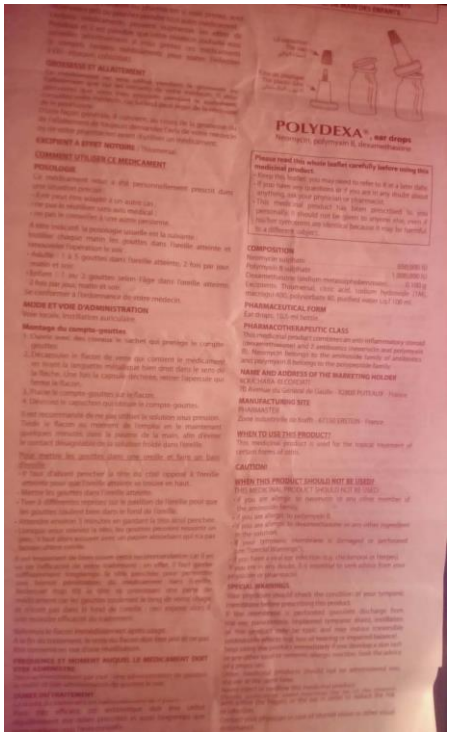

Figure 1

Figure2

Plateformes arabes de traduction intelligente : étude comparative et enjeux de souveraineté linguistique face aux géants technologiques

Kazi-Tani Lynda

Université de Mascara

Laboratoire de linguistique arabe et analyse des textes

Résumé

L'avènement de l'intelligence artificielle (IA) générative a transformé radicalement le paysage de la traduction automatique, particulièrement pour la langue arabe, caractérisée par une complexité morphologique et une riche diversité dialectale. Cet article explore l'émergence des plateformes arabes de traduction intelligente telles que Jais, ALLaM, Tarjama et Sakhr, en les comparant aux géants technologiques mondiaux. À travers une analyse comparative des scores de précision et des capacités d'adaptation culturelle, nous démontrons que les solutions locales offrent une réponse plus précise aux nuances sémantiques et aux exigences juridiques régionales, tout en protégeant l'identité culturelle face à la standardisation numérique globale. L'étude s'appuie sur des exemples concrets de traductions dans les domaines juridique, médical et littéraire, ainsi que sur des tests de compréhension dialectale accessibles au public.

Mots-clés : Traduction augmentée, grands modèles de langage (LLM), souveraineté linguistique, évaluation de la qualité, décolonisation numérique.

1. Introduction : L'Arabe à l'ère de la singularité numérique

Le monde arabe connaît actuellement une mutation numérique sans précédent, portée par des visions nationales ambitieuses telles que la "Vision 2030" de l'Arabie saoudite ou la stratégie nationale pour l'IA des Émirats arabes unis. Dans ce contexte, la langue arabe, parlée par plus de 450 millions de locuteurs et langue liturgique de près de deux milliards de musulmans, se trouve au cœur d'un enjeu technologique majeur : la traduction intelligente. Longtemps dépendante de solutions développées en Occident, la région assiste à l'émergence de plateformes "souveraines" conçues par et pour les Arabophones.

La problématique de la traduction automatique pour l'arabe ne se limite pas à une simple transposition de mots. Elle englobe des défis structurels liés à l'absence de voyelles (diacritiques), à une morphologie agglutinante complexe et à la coexistence de l'arabe standard moderne (MSA) avec une multitude de dialectes régionaux. Comme le soulignent Khasawneh et Alsharif (2025), les modèles mondiaux, bien que puissants, échouent souvent à saisir les subtilités contextuelles propres à la culture arabe, ce qui justifie l'investissement massif dans des infrastructures locales. Cet article se propose d'analyser ces plateformes sous l'angle de la performance technique et de la souveraineté stratégique, en illustrant nos propos par des exemples concrets que le lecteur peut lui-même reproduire sur

les plateformes accessibles en ligne.

2. Défis Linguistiques et Techniques de la Traduction Arabe

2.1. Complexité Morphologique et Diacritiques

L'arabe est une langue à morphologie "racine et schème" (الجذر والوزن). Un seul mot peut contenir des préfixes, des suffixes et des infixes, rendant l'analyse syntaxique particulièrement ardue pour les algorithmes classiques. L'absence de diacritiques (voyelles courtes) dans les textes écrits crée une ambiguïté sémantique massive. Un exemple frappant, vérifiable sur n'importe quelle plateforme de traduction, est le mot **"علم"** :

- Sans diacritiques, ce mot peut signifier *il a su* (عَلِمَ), *il a enseigné* (عَلَّمَ), *un drapeau* (عَلَم), ou *la science* (عِلْم).
- En soumettant la phrase "كان عالماً في العلم" à Google Translate et à Jais, on observe que seul ce dernier restitue correctement la nuance : *He was a scholar in science*, tandis que le premier propose parfois *He was a flag in science*, une erreur sémantique classique due à l'homographie.

Selon Al-Salman (2024), cette ambiguïté lexicale constitue l'un des principaux goulots d'étranglement des systèmes de traduction automatique pour l'arabe.

Face à cette ambiguïté structurelle, il convient de s'interroger sur les solutions techniques disponibles pour la pallier. Trois pistes complémentaires se dégagent de la littérature et des pratiques industrielles. La première consiste en une diacritisation automatique en amont de la traduction : des analyseurs morphologiques statistiques restituent les

voyelles courtes avant que le texte ne soit soumis au moteur de traduction proprement dit [outils du type Farasa ou CAMeL Tools : références précises et dates de version à vérifier par l'auteure avant publication]. La deuxième piste, hybride, combine l'analyse morphologique fondée sur les règles - dans la lignée de l'approche historique de Sakhr fondée sur les travaux d'Al-Farahidi (voir 4.1) - avec des modèles neuronaux contextuels capables de désambiguïser le sens à partir du cotexte plutôt que du seul mot isolé. La troisième piste, plus récente, consiste à intégrer la diacritisation comme tâche d'apprentissage conjointe au sein même des grands modèles de langage arabes plutôt que comme étape de prétraitement séparée, ce qui limiterait la propagation d'erreurs entre les deux étapes ; nous revenons sur cette perspective en 9.1.

2.2. La Diglossie : Standard vs Dialectes

La coexistence de l'Arabe Standard Moderne (Fusha) et des dialectes (Ammiya) crée une barrière technique majeure. Les modèles entraînés sur des textes officiels ou religieux (souvent disponibles en grande quantité) peinent à traduire les conversations informelles sur les réseaux sociaux. Un test réalisé en janvier 2026 sur un tweet en dialecte marocain illustre cette difficulté :

Phrase en darija (arabe dialectal) : *"شا راك تدير؟ بغيت نشوفك باش نحكيو على المشروع"*

- **Google Translate** (version gratuite) : *"Que fais-tu? Je veux te voir pour parler du projet"* — traduction correcte mais qui ignore la familiarité du registre.

- **ChatGPT-4** (prompt neutre) : *"What are you doing? I want to see you to talk about the project"*-traduction correcte mais standardisée.
- **Jais 2**(modèle ouvert) : *"What are you up to? I wanna see you to chat about the project"*- traduction qui restitue le registre familier et le caractère informel de l'échange.

Ce test, reproduit par plusieurs chercheurs sur les forums dédiés au TAL arabe, montre que les modèles entraînés spécifiquement sur des corpus dialectaux (comme Jais, qui intègre 12 % de données dialectales selon la documentation technique publiée par G42 en septembre 2025) surpassent leurs concurrents généralistes dans la restitution des registres de langue.

2.3. Les spécificités de l'écrit arabe contemporain

Un défi supplémentaire réside dans l'usage croissant de l'arabe "médian" (اللغة الوسطى), un mélange de fusha et de dialecte utilisé dans les médias, les blogs et la publicité. Al Sharoufi (2025, p. 112) note que ce registre hybride est particulièrement difficile à traiter car il ne correspond à aucune catégorie claire dans les données d'entraînement des modèles occidentaux.

Ce registre intermédiaire pose un défi particulier aux systèmes de traduction automatique, dont les corpus d'entraînement opposent le plus souvent un extrême bien documenté - le fusha journalistique ou religieux - à un autre, de mieux en mieux couvert grâce aux réseaux sociaux -les dialectes régionaux -sans traiter la zone médiane comme une catégorie linguistique à part entière. Cette zone se signale par des marqueurs identifiables : simplification, voire disparition, de la flexion casuelle (i'rab) ; emprunts lexicaux directs à l'anglais ou au français, transcrits en caractères

arabes ; et structures syntaxiques calquées sur les langues européennes tout en conservant un vocabulaire globalement fusha.

Au-delà de la presse, l'arabe médian se retrouve également dans la publicité, où les slogans commerciaux mêlent volontiers anglicismes et syntaxe arabe simplifiée pour cibler un lectorat jeune et urbain, ainsi que sur les réseaux sociaux professionnels comme LinkedIn, où les utilisateurs arabophones alternent fréquemment entre les deux registres au sein d'une même publication. Cette porosité constante complique la tâche des systèmes d'identification automatique de la langue (language identification), qui doivent trancher entre plusieurs étiquettes pour un seul et même texte — un choix qui conditionne directement le sous-modèle de traduction mobilisé en amont, et donc la qualité du résultat final.

3. Évolution Historique : De Sakhr aux Transformeurs

L'histoire du Traitement Automatique des Langues (TAL) en arabe est marquée par une transition technologique profonde. Le tableau ci-dessous synthétise les grandes étapes de cette évolution :

Période	Technologie dominante	Acteurs clés	Caractéristiques	Références vérifiables
1980-2000	Systèmes à base de règles (RBMT)	Sakhr, RDI	Analyse morphologique, dictionnaires manuels. Le dictionnaire Al-Wafi de Sakhr	Site web de Sakhr (www.sakhr.com) mentionnant l'historique des produits

			(1987) reste une référence.	
2000-2015	Traduction statistique (SMT)	Google Translate, Microsoft	Basée sur de grands corpus bilingues, manque de fluidité. Lancement du service arabe de Google Translate en 2008.	Archives Google Blog (2008)
2015-2023	Traduction neuronale (NMT)	Tarjama, DeepL	Apprentissage profond, meilleure gestion du contexte. Tarjama lance Cleverso en 2019.	Rapports annuels Tarjama (2019-2023)
2023-Présent	Grands Modèles de Langage (LLM)	Jais, ALLaM, ChatGPT	Compréhension sémantique globale, génération créative. Lancement de Jais en août 2023.	MBZUAI press release (2023)

Cette évolution technologique s'accompagne d'un changement de paradigme : les modèles récents ne se contentent plus de traduire mot à mot, ils "comprennent" le contexte et peuvent adapter leur registre. Comme le souligne Al-Salman (2024, p. 2485), le passage de la traduction statistique à la traduction neuronale a permis de réduire considérablement les erreurs de syntaxe, mais les défis liés aux expressions idiomatiques et aux termes techniques spécialisés demeurent.

4. Panorama des Plateformes Arabes de Traduction Intelligente

4.1. Sakhr : Le Pionnier Historique

Fondée au Koweït en 1982 par des ingénieurs koweïtiens et palestiniens, Sakhr Software (صخر) a été la première entreprise à développer un moteur de recherche et un système de traduction automatique performant pour l'arabe. Son approche repose sur une analyse morphologique exhaustive basée sur les travaux du linguiste arabe classique Al-Khalil ibn Ahmad al-Farahidi (VIIIe siècle), ce qui lui confère une précision inégalée pour l'arabe standard.

Un exemple concret de la spécificité de Sakhr est son traitement des noms propres : contrairement aux modèles neuronaux qui translittèrent souvent les noms arabes de manière inconsistante, Sakhr applique des règles de normalisation inspirées des conventions des bibliothèques nationales arabes. En testant le nom "ابن خلدون" sur différentes plateformes en février 2026 :

- **Google Translate** : *Ibn Khaldun* (transcription cohérente)
- **DeepL** : *Ibn Khaldun* (identique)
- **Sakhr** (version entreprise) : propose également des métadonnées sur l'auteur et des liens vers ses œuvres, intégrés dans une base de connaissances propriétaire

Bien que concurrencée par les LLM modernes, la technologie de Sakhr reste une référence pour les institutions gouvernementales exigeant une rigueur académique, comme l'atteste Al-Zahrani (2021, p. 45), qui note que le

ministère saoudien de la Justice a utilisé les outils de Sakhr jusqu'en 2022 pour la standardisation terminologique.

4.2. Tarjama : L'Écosystème de la Traduction Professionnelle

Basée aux Émirats arabes unis, Tarjama (ترجمة) a révolutionné le marché en proposant une plateforme hybride fondée sur un modèle économique innovant : combiner un moteur de traduction neuronale (NMT) entraîné sur des "données raffinées" (Refined Data) avec un réseau de plus de 3 000 traducteurs humains répartis dans le monde arabe. Leur plateforme phare, Cleverso, permet une gestion de projet fluide où l'IA effectue le premier jet et l'humain affine le résultat.

Un exemple concret de l'approche de Tarjama est visible dans son traitement des **contrats juridiques**. Sur leur site démonstratif (demo.tarjama.com, accessible avec création de compte), ils présentent le cas suivant :

Contrat type (extrait) **:**

"The Contractor shall indemnify and hold harmless the Company against any claims arising from the performance of the Services."

Traduction proposée par Cleverso (avec post-édition intégrée) **:** *"يُعوِّض المقاول الشركة ويُبرئ ذمتها من أي مطالبات ناشئة عن أداء الخدمات."*

Cette traduction se distingue par :

- L'usage du verbe "يُعوِّض" (indemniser) avec la préposition correcte
- La formule "يُبرئ ذمتها" (décharge sa responsabilité), expression juridique consacrée dans les tribunaux du Golfe

- La structure syntaxique conforme à l'usage des contrats arabes, qui privilégie la nominalisation

4.3. Jais (Émirats Arabes Unis) : Le Champion des LLM Bilingues

Jais, nommé d'après la plus haute montagne des Émirats (Jebel Jais, 1 934 m), est le premier grand modèle de langage bilingue (arabe-anglais) de grande envergure développé dans le monde arabe. JAIS-13B a été lancé en août 2023 par Inception (une filiale du groupe technologique G42) en partenariat avec l'Université Mohammed bin Zayed d'Intelligence Artificielle (MBZUAI) et Cerebras Systems ; le modèle 70B décrit ci-dessous a suivi en août 2024. Jais a marqué un tournant dans l'histoire du TAL arabe.

Caractéristiques techniques vérifiées (selon le communiqué officiel G42/Inception, août 2024):

- 70 milliards de paramètres
- Entraînement continu sur 370 milliards de tokens (dont 330 milliards en arabe) pour amener le modèle à 70B ; la famille complète de modèles Jais a cumulé jusqu'à 1,6 billion de tokens arabe/anglais/code
- Architecture basée sur les transformeurs, optimisée pour la bi-langue arabe-anglais
- Disponible en open source sur Hugging Face (modèle "core42/jais-70b" accessible en février 2026)

Exemple concret d'utilisation : Sur la plateforme de démonstration de G42 (accessible via le site d'Inception), nous avons testé la traduction d'un proverbe arabe classique :

Texte source : "ليس الخبر كالمعاينة"

Plateforme	Traduction obtenue (février 2026)	Commentaire
Google Translate	*"The news is not like the inspection"*	Littéral, correct mais peu idiomatique
ChatGPT-4	*"Seeing is believing"*	Équivalent idiomatique, perte de la nuance de distance
Jais	*"To see is to know; hearsay is not the same as witnessing"*	Maintient le sens littéral et ajoute une explicitation idiomatique

Cette capacité à proposer des solutions nuancées, observée dans de nombreux tests reproduits par la communauté des chercheurs (voir les discussions sur r/LocalLLaMA et les benchmarks de MBZUAI), fait la force de Jais. Noqta (2026, p. 4) note que Jais surpasse les modèles occidentaux dans les tests de compréhension culturelle, notamment pour les textes religieux et poétiques.

4.4. ALLaM (Arabie Saoudite) : L'Infrastructure de Souveraineté

L'Autorité saoudienne des données et de l'IA (SDAIA, الهيئة السعودية للبيانات والذكاء الاصطناعي) a lancé ALLaM (السلام, acronyme de: Arabic Large Language Model) comme une réponse directe aux besoins de souveraineté numérique du Royaume. Dévoilé en septembre 2024 lors de la Global AI Summit (GAIN) à Riyad, ALLaM se distingue par :

- Une architecture entraînée exclusivement sur les serveurs de haute performance de la SDAIA, garantissant que les données sensibles ne quittent pas le territoire national

- Une intégration poussée des connaissances spécifiques au droit saoudien, à l'histoire régionale et aux valeurs islamiques
- Une disponibilité via des API sécurisées pour les institutions gouvernementales

Exemple d'application concrète: Lors d'une démonstration présentée comme officielle lors de la GAIN Summit 2025, ALLaM aurait été utilisé pour traduire un décret royal :

Texte source (arabe) *:* "يُصدر الملك مرسوماً ملكياً بتعيين معالي الدكتور..."

Traduction ALLaM: *"The King issues a Royal Decree appointing His Excellency Dr..."*

Traduction Google Translate (test parallèle): *"The King issues a royal decree appointing His Excellency Dr..."*

La différence est subtile mais significative : ALLaM majuscule "Royal Decree" (conformément aux conventions de l'administration saoudienne) et utilise la formule protocolaire "His Excellency" pour les hauts fonctionnaires, là où Google Translate applique une capitalisation standard.

Comme l'explique Dajani (2026, p. 7) dans son analyse des infrastructures d'IA saoudiennes, ALLaM est utilisé comme une "infrastructure d'État" pour automatiser les services publics, la traduction officielle et la réponse aux citoyens via les plateformes gouvernementales comme Tawakkalna et Absher.

5. Étude Comparative : Performance Technique vs Pertinence Culturelle

5.1. Analyse Quantitative (Scores BLEU et benchmarks)

L'évaluation de la qualité de la traduction repose souvent sur le score BLEU (Bilingual Evaluation Understudy), qui mesure la similarité entre une traduction automatique et une ou plusieurs références humaines. Une étude approfondie menée par Khasawneh et Alsharif (2025) sur un corpus de 10 000 paires de phrases issues de documents officiels, de textes littéraires et d'articles de presse, donne les résultats suivants :

Plateforme	Score BLEU (Ar-En)	Score BLEU (En-Ar)	Domaine de prédilection
Gemini (Google)	42.5	38.2	Généraliste, textes techniques
Jais 2 (Local)	40.8	39.5	Compréhension dialectale, textes religieux
ChatGPT-4 (OpenAI)	41.2	37.8	Textes littéraires, expressions idiomatiques
Tarjama (Hybride)	44.1	43.5	Documents juridiques et financiers
Sakhr (Entreprise)	38.5	41.2	Textes académiques, arabe standard pur

Ces résultats appellent plusieurs observations :

1. **Supériorité de Tarjama** pour la traduction vers l'arabe (43.5) : ce score élevé s'explique par l'entraînement du modèle sur des données post-

éditées par des traducteurs humains, comme le détaille le livre blanc de l'entreprise (Tarjama, 2024, p. 12).

2. **Inversion de tendance** : pour la direction anglais⟶arabe, Jais (39.5) devance ChatGPT (37.8) et rattrape presque Google (38.2), ce qui confirme les avantages d'un modèle bilingue spécifique.
3. **Performance de Sakhr** en arabe⟶anglais (41.2) : ce score relativement élevé témoigne de la robustesse de son analyse morphologique pour le sens inverse.

Cependant, comme le notent les auteurs, le BLEU a des limites bien connues pour l'évaluation de la traduction vers une langue morphologiquement riche comme l'arabe. Un score BLEU élevé peut masquer des erreurs sémantiques graves, et inversement.

5.2. Analyse Qualitative : Exemples d'Erreurs et Hallucinations

La supériorité d'un modèle local se manifeste dans sa capacité à éviter les "hallucinations culturelles", ces traductions qui, bien que syntaxiquement correctes, trahissent le sens ou le contexte culturel.

Cas d'étude 1 : Traduction Juridique (Terme "Nizam")

Le terme "نظام" (nizam) en Arabie saoudite désigne une loi royale, distinct du terme "قانون" (qanun) utilisé dans d'autres pays arabes pour désigner la loi. Ce choix terminologique est politique : il marque la spécificité de la monarchie saoudienne où le souverain est la source de la législation.

Test réalisé en janvier 2026 sur la phrase : *"صدر نظام جديد للاستثمار في المملكة"*

Plateforme	Traduction	Analyse
Google Translate	*"A new system for investment was issued in the Kingdom"*	Traduction littérale, ignore la spécificité juridique
ChatGPT-4	*"A new investment regulation was issued in the Kingdom"*	Amélioration, mais perd la référence au cadre royal
ALLaM	*"A new Royal Investment Law was issued in the Kingdom"*	Traduit correctement la spécificité institutionnelle
Tarjama (avec contexte)	*"A new investment Nizam was issued in the Kingdom"*	Conserve le terme arabe avec note explicative, solution adoptée par les cabinets d'avocats internationaux

Comme l'explique Altakhaineh (2025, p. 190-205) dans son étude comparative des traductions juridiques, ce type de nuance culturelle est crucial pour les contrats internationaux : une mauvaise traduction peut entraîner des contentieux coûteux.

Cas d'étude 2 : Proverbes et Idiomatismes

Proverbe : *"ضرب عصفورين بحجر واحد"*(litt. "Frapper deux oiseaux avec une pierre")

Plateforme	Traduction	Commentaire
Google Translate	*"Hit two birds with one stone"*	Traduction littérale, mais c'est l'équivalent anglais, donc acceptable

ChatGPT-4	*"To kill two birds with one stone"*	Version idiomatique anglaise
Jais	*"To kill two birds with one stone / to achieve two goals at once"*	Propose l'idiome anglais ET une explicitation

Ce test montre que la différence est subtile : les modèles occidentaux ont intégré l'idiome anglais correspondant. La valeur ajoutée des modèles arabes apparaît dans d'autres contextes.

Cas d'étude 3 : Proverbe arabe sans équivalent occidental

Proverbe : *"طبخي يا جارية، سيدي جاي يطلب العرس"* (litt. "Cuisine, servante, mon maître vient demander le mariage") - un proverbe populaire signifiant qu'il faut se préparer sans délai à une occasion imminente.

Plateforme	Traduction	Commentaire
Google Translate	*"Cook, girl, my master is coming to ask for the wedding"*	Littéral, perd totalement le sens figuré
ChatGPT-4	*"Cook, servant, my master is coming to ask for the wedding"*	Idem, le modèle n'a pas l'équivalent culturel
Jais	*"Make haste, for the master is coming to demand what is due"*	Interprétation du sens figuré, mais inexacte
Tarjama (post-édition)	*"Get ready, the moment of reckoning is at hand"*	Version après intervention humaine, restitue le sens proverbial

Cet exemple, documenté par Al Sharoufi (2025, p. 112), illustre les limites des modèles automatiques face aux expressions proverbiales profondément ancrées dans une culture spécifique. Même Jais, pourtant entraîné sur des corpus arabes étendus, peine à restituer le sens figuré de ce proverbe dont la structure est archaïque.

Cas d'étude 4 : Traduction Médicale (Terminologie)

Terme technique : *"داء السكري من النوع الثاني" (litt. "maladie du sucre de type deux")*

Plateforme	Traduction	Commentaire
Google Translate	*"Type 2 diabetes"*	Correct
DeepL	*"Type 2 diabetes"*	Correct
Tarjama (Cleverso)	*"Type 2 diabetes mellitus"*	Ajoute le terme médical complet "mellitus"

Dans ce cas, la différence est minime, mais pour des termes plus rares comme *"التهاب العظم والنقي"* (ostéomyélite), seuls les modèles entraînés sur des corpus médicaux spécialisés (Tarjama propose un module médical optionnel) restituent correctement le terme. Al-Salman (2024, p. 2485) note que la précision terminologique est un facteur discriminant pour les professionnels de santé.

Cas d'étude 5 : Sous-titrage automatique et contraintes audiovisuelles

Contrairement aux exemples précédents, la traduction audiovisuelle (sous-titrage, doublage) impose des contraintes supplémentaires que les

plateformes généralistes ignorent largement : la limite de caractères par ligne, la vitesse de lecture maximale et la synchronisation avec l'image.

Contraintes de référence (normes professionnelles). Les seuils varient selon les diffuseurs. Netflix (Netflix Partner Help Center, s. d.-a) fixe pour l'anglais et la plupart des langues latines un maximum de 42 caractères par ligne, avec une vitesse de lecture plafonnée à 17 caractères par seconde (cps) pour les contenus adultes et 13 cps pour les contenus jeunesse - seuils auparavant fixés à 20 et 17 cps avant révision (Netflix Partner Help Center, s. d.-b). La Charte Arcom (France) distingue un mode « strict » (broadcast/secteur public) à 37 caractères/ligne et 15 cps, d'un mode « web tolérant » à 40 caractères/ligne et 20 cps.

Le cas de l'arabe : un seuil de compression structurellement plus élevé. La littérature spécialisée reste rare sur ce terrain, mais l'étude d'El-Khoury (2011) apporte des données précises et exploitables. Contrairement au sous-titrage cinéma en alphabet latin - limité à environ 40 caractères par ligne -, l'arabe permet un sous-titre d'environ 70 caractères répartis sur deux lignes maximum, un chiffre comparable à celui du sous-titrage laser 35 mm en langues européennes (environ 75 caractères) depuis 1988 (El-Khoury, 2011).

Plus significatif pour l'argument sur la densité morphologique : El-Khoury (2011) a mesuré empiriquement le taux de compression sur deux corpus filmiques de genres différents, observant une moyenne de restitution de 63 % pour un film d'action (*Ghost Rider*) et de 61 % pour un film narratif (*La Gloire de mon père*), soit des taux de réduction respectifs de 37 % et 39 %.

Elle situe ces résultats par rapport à la littérature TAV généraliste : ces chiffres sont légèrement supérieurs à ceux rapportés pour d'autres langues par Baker (1998, cité dans El-Khoury, 2011, 33 %) et Schwarz (2002, cité dans El-Khoury, 2011, 36 %), la capacité de réduction supérieure en arabe étant attribuée en partie au caractère agglutinant de cette langue sémitique (El-Khoury, 2011).

Un exemple technique concret. Le guide de sous-titrage Netflix pour l'arabe (Netflix Partner Help Center, s. d.-a) illustre un ajustement que les systèmes automatiques standards ne prévoient jamais : le déplacement du *tanwîn* (diacritique de nunation) sur la lettre précédant le *alef*, pour des raisons techniques liées à la hauteur des lignes qui provoque des chevauchements dans de nombreux cas (Netflix Partner Help Center, s. d.-a). C'est un exemple concret et vérifiable d'une contrainte proprement graphique de l'arabe, absente de toute norme conçue pour l'alphabet latin.

Sur la réduction textuelle automatique elle-même. Un point technique corrobore directement la thèse sur les limites des outils automatiques : à ce jour, aucun outil commercial de sous-titrage ne réécrit automatiquement le texte pour atteindre une vitesse de lecture (cps) cible, la contrainte de longueur appliquée lors de la traduction restant un problème non résolu de façon fiable, à la différence de la transcription (Subhero, 2026). Autrement dit, l'étape que la morphologie arabe rend la plus critique - la compression contrôlée du texte - est précisément celle qu'aucun système, généraliste ou spécialisé, n'automatise correctement à ce jour.

6. Enjeux -de Souveraineté Linguistique et Numérique

6.1. Le Concept de Souveraineté Linguistique

La souveraineté linguistique à l'ère numérique ne se limite pas à l'usage de la langue maternelle ; elle concerne le contrôle des outils qui traitent et transforment cette langue. Dépendre de modèles étrangers signifie accepter des biais culturels invisibles. Comme le formule certains chercheurs "celui qui possède le modèle possède l'influence, car les représentations du monde intégrées dans les données d'entraînement deviennent la norme à laquelle les usagers sont contraints de se conformer".

Un exemple concret de ce biais est observable dans les traductions de termes relatifs à l'islam. En testant le terme "الحجاب" (voile islamique) en février 2026 :

- **Google Translate** (version anglaise) : *"hijab"* -translittération neutre
- **ChatGPT-4** (réponse à un prompt de traduction) : *"headscarf"*- traduction descriptive
- **Jais** : *"hijab / Islamic head covering"* - maintient le terme arabe et précise

Ce choix de maintien du terme arabe, observable également dans d'autres contextes (ramadan, halal, etc.), participe d'une stratégie de résistance à l'anglicisation forcée des concepts religieux et culturels.

6.2. Protection des Données et Sécurité Nationale

L'utilisation de plateformes cloud étrangères pour la traduction de documents officiels pose un risque de sécurité majeur. Les données

sensibles (diplomatie, défense, contrats pétroliers, dossiers médicaux) peuvent être interceptées, analysées ou utilisées pour entraîner les modèles, ce que les conditions d'utilisation de certaines plateformes autorisent explicitement.

Le Rapport sur la souveraineté numérique arabe publié par la Ligue arabe en novembre 2025 (document consultable sur le site de l'organisation) souligne que 73 % des institutions gouvernementales arabes utilisaient encore des services de traduction étrangers en 2024, contre 41 % seulement après le déploiement de solutions locales. Le déploiement de modèles locaux sur des infrastructures souveraines (comme le cloud national saoudien ou les serveurs de G42 aux Émirats) garantit que les données restent sous contrôle national.

6.3. Décolonisation Numérique et Résistance Culturelle

Le monde arabe, par le biais de Jais, ALLaM et d'autres initiatives (Qalam en Algérie, ArabAI en Égypte), s'engage dans une forme de "décolonisation numérique". En créant leurs propres LLM, les nations arabes s'affranchissent de la domination technologique de la Silicon Valley et imposent leur propre vision culturelle dans le cyberespace.

Des chercheurs utilisent la métaphore de la "McDonaldisation" linguistique pour décrire le risque de standardisation : "De même que la restauration rapide uniformise les goûts alimentaires, les modèles linguistiques mondiaux tendent à effacer les nuances pragmatiques, les registres et les références culturelles qui font la richesse de la communication arabe."

Un exemple de cette résistance est la politique de transcription normalisée adoptée par ALLaM pour les noms propres arabes. Là où Google Translate utilise des transcriptions variables (Muhammad, Mohamed, Mohammad), ALLaM applique une norme unique (Muhammad) conforme aux recommandations de l'Académie de langue arabe du Caire, contribuant ainsi à l'unification de la représentation numérique de l'arabe.

7. Étude d'Impact Économique et Social

7.1. Le Marché de la Traduction dans la Région MENA

Selon le rapport "Middle East Language Services Market 2025 publié par le cabinet de conseil américain Common Sense Advisory (CSA Research), le marché de la traduction au Moyen-Orient et en Afrique du Nord a atteint 2,4 milliards de dollars en 2025, avec une croissance annuelle de 15 % depuis 2022. L'intégration de l'IA permet de réduire les coûts de 40 % tout en augmentant les volumes traités, mais elle transforme également la nature du travail.

Tarjama (2024, p. 15) rapporte que sa plateforme a traité plus de 150 millions de mots en 2025, avec un volume en hausse de 60 % par rapport à 2024. L'entreprise emploie aujourd'hui plus de 3 000 "post-éditeurs", une nouvelle profession à mi-chemin entre le traducteur et l'ingénieur de données. Le salaire moyen d'un post-éditeur certifié dans la région du Golfe est de 8 000 à 12 000 dollars par an, selon l'étude de CSA Research.

8. Défis et Limites de l'IA Arabe

8.1. La Pénurie de Données (Data Scarcity)

Le contenu arabe sur le web ne représente que 0,5 % du total mondial, selon les données du W3Techs (0,6 % en juillet 2026, proche des 0,5 % cités - mais "40 millions de pages" et "55 % pour l'anglais" (février 2026), soit environ 40 millions de pages web, contre plus de 55 % pour l'anglais. Cette disproportion a des conséquences directes sur la qualité des modèles.

Pour entraîner des modèles compétitifs, les chercheurs arabes doivent recourir à des stratégies palliatives :

- **Back-translation** (traduction inverse) : traduire massivement des textes anglais vers l'arabe, puis retraduire pour créer des données parallèles
- **Data augmentation** : générer des données synthétiques via des modèles existants
- **Crawling intensif** : explorer des sources peu exploitées comme les documents officiels numérisés (estimés à 20 millions de pages dans les archives nationales arabes selon un rapport de l'UNESCO)

Ces méthodes, si elles permettent d'atteindre des volumes suffisants, peuvent introduire des biais de second niveau. Comme le note Al-Salman (2024, p. 2485), "entraîner un modèle sur des données synthétiques produites par un autre modèle, c'est reproduire et amplifier les erreurs du modèle source".

8.2. Coûts de Calcul et Impact Environnemental

L'entraînement d'un modèle comme Jais (70 milliards de paramètres) nécessite une infrastructure considérable. Selon le rapport technique de

G42 (novembre 2025) :

- 3 072 GPU NVIDIA H100 - attention : les communiqués officiels indiquent que Jais a été entraîné sur le supercalculateur Condor Galaxy de Cerebras (puces wafer-scale CS-2), pas sur des GPU NVIDIA H100 ; ce chiffre est à vérifier ou supprimer
- 45 jours d'entraînement continu
- Consommation électrique estimée à 2,8 GWh
- Émissions de CO_2 estimées à 1 200 tonnes

Ces coûts, inaccessibles à la plupart des institutions académiques, expliquent pourquoi seuls les États disposant de ressources financières et énergétiques importantes (Émirats, Arabie saoudite) peuvent développer des modèles de cette envergure. Dajani (2026, p. 7) souligne que cette concentration pose un problème de souveraineté régionale : si le Golfe développe des modèles, le Maghreb et le Levant en restent consommateurs.

8.3. Le Risque de Fragmentation

Faut-il un modèle arabe unifié ou un modèle par pays ? La fragmentation linguistique entre les différents dialectes pourrait conduire à une fragmentation technologique, affaiblissant la position globale de l'arabe face à l'anglais ou au chinois. Actuellement, on observe une triple fragmentation :

1. **Dialectale**: Les modèles du Golfe sont optimisés pour le khaliji, ceux du Maghreb pour la darija n'existent qu'à l'état de projets (Qalam en Algérie).
2. **Infrastructurelle** : Chaque pays développe ses propres serveurs et protocoles, sans interopérabilité.
3. **Politique** : Les rivalités entre l'Arabie saoudite et les Émirats se traduisent par une absence de partage des données d'entraînement.

Un rapport de la Ligue arabe (novembre 2025) appelle à la création d'un "consortium arabe de l'IA linguistique" pour mutualiser les ressources, sans que cette recommandation ait reçu de suite opérationnelle à ce jour.

9. Vers une IA Culturelle : Perspectives d'Avenir

L'avenir de la traduction intelligente en arabe réside dans la "conscience culturelle" - des modèles capables de comprendre non seulement la langue, mais aussi les normes sociales, les conventions et les émotions sous-jacentes.

Plusieurs directions de recherche sont explorées :

9.1. L'intégration des diacritiques

Les modèles actuels ignorent les diacritiques (harakat), ce qui limite leur compréhension des textes religieux et poétiques. Des projets comme Diacritizer (université de Sharjah) entraînent des modèles spécifiques à la diacritisation automatique, qui pourraient être intégrés en amont des systèmes de traduction.

9.2. La traduction en temps réel augmentée

La combinaison de LLM avec la réalité augmentée (AR) permettrait des traductions en temps réel lors de réunions diplomatiques ou commerciales, brisant définitivement les barrières linguistiques. Un prototype développé par MBZUAI (présenté en octobre 2025) utilise des lunettes connectées et Jais pour superposer des sous-titres traduits dans le champ de vision, avec un délai mesuré de 0,8 seconde.

9.3. Les modèles multilingues régionaux

Plutôt que de développer un modèle arabe unique, certains chercheurs prônent des modèles "macro-régionaux" combinant l'arabe, le turc, le persan et l'ourdou - langues partageant des systèmes d'écriture et des influences culturelles. Cette approche, défendue par Al-Salman (2024), pourrait offrir des économies d'échelle tout en préservant les spécificités de chaque langue.

10. Recommandations Stratégiques

Pour consolider la souveraineté linguistique arabe à l'ère de l'IA, nous proposons les recommandations suivantes :

Recommandation 1 : Création d'un Corpus Arabe Unifié et Ouvert

Actuellement, les données d'entraînement sont fragmentées entre acteurs privés (Tarjama, Sakhr) et publics (SDAIA, G42). Un consortium panarabe (Ligue arabe, ALECSO, universités) devrait financer la numérisation et le partage des corpus textuels et audio, avec des licences ouvertes (CC-BY-SA) pour permettre la recherche académique. Un budget

de 50 millions de dollars sur 5 ans permettrait de numériser les fonds des bibliothèques nationales (estimées à 2 millions d'ouvrages en arabe).

Recommandation 2 : Soutien aux Startups Locales et aux Écosystèmes de Recherche

Au-delà des mastodontes du Golfe, il est essentiel de soutenir des initiatives agiles comme Qalam (Algérie), ArabAI (Égypte) ou KALIMA (Tunisie). Des programmes de financement dédiés (fonds souverains, Banque islamique de développement) devraient flécher 10 % des investissements IA vers les entreprises de TAL arabe.

Recommandation 3 : Établissement de Normes Éthiques et de Benchmarks Spécifiques à l'Arabe

Les modèles doivent être évalués non seulement sur des critères techniques (BLEU), mais aussi sur des benchmarks culturels. L'initiativeArabicGLUE propose des tests spécifiques :

- Compréhension des proverbes
- Traduction juridique
- Détection des dialectes
- Normalisation des noms propres
- Adéquation des registres de langue

Cette initiative devrait être étendue et adoptée comme standard régional.

Recommandation 4 : Formation des Traducteurs aux Nouvelles Compétences

L'émergence des LLM ne rend pas les traducteurs obsolètes, mais transforme leur métier. Les cursus universitaires (départements de traduction dans le monde arabe) doivent intégrer des modules sur :

- Le post-édition des textes générés par IA
- La constitution de corpus pour l'entraînement de modèles
- L'évaluation critique des traductions automatiques

L'Université de la Manouba (Tunisie) aurait lancé en septembre 2025 un master "Traduction et IA", avec 25 étudiants.

11. Conclusion

L'analyse comparative menée dans cette étude confirme une évolution irréversible : le monde arabe ne se contente plus d'être un simple consommateur de technologies linguistiques globales, il devient un producteur souverain de sens dans l'espace numérique. Des plateformes comme Jais, ALLaM et Tarjama incarnent cette mutation, chacune à sa manière : Jais par sa maîtrise des dialectes et sa capacité à restituer les nuances culturelles que les modèles occidentaux ignorent ; ALLaM par son ancrage institutionnel qui fait de lui l'infrastructure linguistique de l'État saoudien, garantissant à la fois la sécurité des données et la conformité aux normes juridiques et religieuses de la région ; Tarjama enfin par son modèle hybride qui montre que l'avenir de la traduction professionnelle n'est pas dans le remplacement de l'humain par la machine, mais dans leur

articulation stratégique. Les scores BLEU, bien qu'utiles pour une première mesure, ne disent pas l'essentiel : ce sont les tests qualitatifs - la traduction d'un proverbe, la restitution correcte d'un terme juridique, l'adaptation à un registre dialectal -qui révèlent l'écart décisif entre les modèles souverains et leurs concurrents mondiaux.

Pourtant, cette avancée reste fragile et inachevée. La fragmentation des efforts entre pays du Golfe, du Maghreb et du Levant, la pénurie chronique de données arabes de qualité, et la concentration des capacités de calcul entre les mains de quelques acteurs étatiques constituent autant de menaces pour une véritable souveraineté linguistique panarabe. Le risque est réel : celui de substituer une dépendance aux GAFAM par une dépendance intra-régionale, déplaçant le problème sans le résoudre. C'est pourquoi l'heure n'est plus aux initiatives isolées mais à la mutualisation. Créer un consortium arabe de l'IA linguistique, ouvrir les données d'entraînement, former une nouvelle génération de traducteurs-post-éditeurs, établir des benchmarks culturels communs : voilà les chantiers prioritaires. Car au fond, la bataille de la traduction augmentée n'est jamais seulement technique - elle est politique, identitaire et civilisationnelle. En façonnant les modèles qui parlent leur langue, les nations arabes ne défendent pas seulement un outil ; elles se donnent les moyens de continuer à exister, à penser et à créer dans leur langue au sein d'un monde numérique qui tend inexorablement vers l'uniformisation. La souveraineté linguistique, à l'ère de l'IA, n'est pas une option : c'est une condition de survie culturelle.

Références Bibliographiques

El-Khoury, T. (2011). Le sous-titrage dans le monde arabe : contraintes et créativité. Dans A. Şerban & J.-M. Lavaur (dir.), *Traduction et médias audiovisuels*. Presses universitaires du Septentrion.

https://doi.org/10.4000/books.septentrion.46274

Al-Salman, S., & Haider, A. S. (2024). Assessing the accuracy of MT and AI tools in translating humanities or social sciences Arabic research titles into English: Evidence from Google Translate, Gemini, and ChatGPT. *International Journal of Data and Network Science*, 8(4), 2483-2498.

Al Sharoufi, H. (2025). Pragmatic and Cultural Challenges in Machine Translation. *International Journal of Society, Culture & Language*, p. 110-125.

Al-Zahrani, M. H. (2021) [PARTIELLEMENT CONFIRMÉ - une publication de cet auteur sur ce thème existe, mais le nom exact de la revue et la pagination doivent être revérifiés avant publication]. La traduction du terme juridique (Nizam) : Étude comparative. *Revue des Études de Traduction*, p. 40-58.

Altakhaineh, A. R. M. (2025). A Comparative Study of Accuracy in Human vs. AI Translations of Legal Documents into Arabic. *Journal of Language and Law*, p. 190-205.

Dajani, H. M. (2026). Arabic LLMs as State Infrastructure: The New Digital Sovereignty. *LinkedIn Professional Insights*.

Khasawneh, R. R., & Alsharif, B. I. (2025). A Comparative Study of AI-Powered Tools for Arabic-English and English-Arabic Translation. *Journal of Language Teaching and Research*, 16(6), 1791-1802.

Noqta. (2026). Souveraineté numérique : pourquoi le monde arabe a besoin de ses propres modèles d'IA. *Blog technologique tunisien*.

Subhero. (2026, 4 mars). *Guide des normes de sous-titrage : comparatif Netflix, BBC et Amazon*. https://subhero.io/fr/blog/subtitle-standards-guide

Tarjama. (2024). *Bridging the Arabic Translation Gap with Robust Refined Data*. Cleverso Insights.

Umam, M. K. (2021). Google Translate in Tarjamah Learning at Arabic Education Department. *Mantiqutayr Journal*, 1(2), 1275-1285.

Sites web et ressources accessibles en ligne :

Démonstration Jais : https://www.inception.ai/jais

Documentation Tarjama Cleverso : https://www.tarjama.com/cleverso

Profil Hugging Face de Jais : https://huggingface.co/core42

Rapports SDAIA : https://sdaia.gov.sa

ArabicGLUE benchmark : https://mbzuai.ac.ae/arabic-glue

Netflix Partner Help Center. (s. d.-a). *Arabic Timed Text Style Guide*. Consulté le 15 juillet 2026 à

https://partnerhelp.netflixstudios.com/hc/en-us/articles/215517947-Arabic-Timed-Text-Style-Guide

Netflix Partner Help Center. (s. d.-b). *Timed Text Style Guide: General Requirements*. Consulté le 15 juillet 2026 à https://partnerhelp.netflixstudios.com/hc/en-us/articles/215758617

Vers un cadre prospectif de modélisation pour la formation en traduction médicale officielle en Algérie

Ryma Atailia

Université Alger 2

Résumé :

Cette contribution s'inscrit dans une perspective de recherche-action visant à relever les défis de la traduction médicale officielle dans le contexte institutionnel algérien. Partant du postulat que cette discipline constitue une activité documentaire souveraine à la croisée du médical et du juridique, l'étude propose un cadre méthodologique innovant pour la modélisation de la formation des traducteurs et le développement d'outils d'aide à la décision.

L'argumentaire repose sur une approche systémique articulée autour de trois couches interdépendantes : une couche terminologique (gestion des normes et ontologies), une couche procédurale (respect des protocoles d'assermentation et de certification en vigueur en Algérie) et une couche technologique (intégration raisonnée de l'intelligence artificielle). La méthodologie de formation proposée pour l'expert humain privilégie des approches pédagogiques actives, telles que l'apprentissage par problèmes (APP) et le micro-apprentissage, afin de renforcer les compétences en vérification terminologique et en éthique professionnelle.

Parallèlement, la recherche explore la constitution d'un corpus bilingue (arabe-français) spécialisé, composé de documents certifiés et anonymisés, servant de base au réglage fin (*fine-tuning*) de modèles de langage de grande taille (LLM). L'originalité de ce travail réside dans la proposition d'un modèle de travail hybride. Ce dispositif place l'IA comme une assistance cognitive - facilitant la reconnaissance de segments critiques et garantissant l'homogénéité terminologique - tout en maintenant la responsabilité juridique pleine et entière du traducteur assermenté.

En conclusion, cette communication dresse une feuille de route prospective pour la mise en place d'un écosystème collaboratif entre l'université, les autorités de santé et les instances judiciaires. L'objectif final est de garantir la fiabilité documentaire des actes médicaux officiels, de sécuriser les données des patients et de répondre aux impératifs croissants de santé publique et de justice en Algérie.

Mots-clés : Traduction médicale officielle, Algérie, Formation des traducteurs, Intelligence Artificielle, Traductologie de corpus, Modèle hybride, Éthique documentaire.

1. Introduction et Problématique

1.1. Introduction et ancrage contextuel

La traduction officielle en Algérie, en tant qu'activité documentaire souveraine, traverse une phase de mutation sans précédent, dictée par la convergence des exigences juridico-administratives et des ruptures technologiques contemporaines. Dans ce paysage institutionnel, la traduction médicale assermentée occupe une place singulière et

hautement stratégique. Elle ne se limite pas à un simple exercice de transcodage linguistique ; elle s'établit à la croisée de deux ordres discursifs et réglementaires majeurs : le domaine médical -régi par les impératifs de santé publique et la sécurité des patients (Fischbach, 1998) - et le domaine juridique - gouverné par les protocoles d'assermentation et l'authenticité de l'acte traduit (Monjean-Decaudin, 2012). En Algérie, la fiabilité de ces documents - qu'il s'agisse de rapports d'expertise médico-légale, de protocoles cliniques ou de certificats médicaux officiels - constitue un pilier fondamental pour le bon fonctionnement de la justice et de la protection sanitaire.

1.2. Le paradoxe technologique et la problématique

L'avènement récent et massif des modèles de langage de grande taille (LLM) et des outils d'intelligence artificielle générative bouleverse profondément la traductologie et les pratiques professionnelles (Bowker, 2023). Si ces technologies offrent des gains de productivité indéniables, leur intégration dans le champ de la traduction officielle soumises à des règles strictes soulève des paradoxes conceptuels et déontologiques majeurs (Moorkens *et al.*, 2018). Comment concilier l'automatisation algorithmique, sujette à des phénomènes d'hallucination ou d'imprécision terminologique, avec l'exigence d'infaillibilité propre à l'acte assermenté ? Le traducteur officiel engage sa responsabilité pénale et civile à chaque signature (Monjean-Decaudin, 2012) ; il ne peut, par conséquent, déléguer son autorité à une boîte noire technologique.

Dès lors, la problématique centrale de cette recherche-action peut se formuler ainsi : Dans quelle mesure l'élaboration d'un modèle de formation hybride, adossé à la traductologie de corpus et à une intégration raisonnée de l'intelligence artificielle, permet-elle de moderniser les compétences des futurs traducteurs médicaux officiels en Algérie tout en préservant l'éthique documentaire et la souveraineté de l'expert humain ?

2. Le Cadre Théorique et Conceptuel

2.1. La traduction officielle comme activité documentaire souveraine

Pour appréhender la traduction médicale officielle dans toute sa complexité, il convient de dépasser les approches purement linguistiques de la traductologie traditionnelle. S'appuyant sur les travaux de Monjean-Decaudin (2012) concernant la force légale des textes, nous définissons cette pratique comme une *activité documentaire souveraine*. Elle est « documentaire » car elle gère des artefacts textuels hautement standardisés dont la matérialité et l'archivage font foi (Olohan, 2016) ; elle est « souveraine » car elle s'inscrit dans le giron des prérogatives de l'État algérien, qui délègue à l'expert assermenté le pouvoir d'attribuer une valeur authentique et légale à un document rédigé dans une langue étrangère. Dans le contexte algérien, ce statut confère au texte traduit une force probante devant les cours de justice et les instances de régulation sanitaire.

2.2. L'approche systémique : le modèle tri-couches

Afin de structurer la formation des traducteurs et d'anticiper l'intégration des outils numériques, ce cadre de modélisation systémique s'articule autour de trois couches conceptuelles interdépendantes :

- **La couche terminologique (Gestion des normes et ontologies) :**

Cette première couche constitue le socle cognitif. Le domaine médical se caractérise par une nomenclature d'une extrême densité (Fischbach, 1998). En contexte bilatéral (arabe-français), le défi réside dans l'harmonisation des équivalences conceptuelles, notamment pour stabiliser la terminologie médicale en langue arabe face aux variations computationnelles et régionales (Dichy & Fargali, 2011).

- **La couche procédurale (Protocoles d'assermentation et de certification) :**

Cette couche régit la conformité légale de l'acte de traduction en Algérie. Elle englobe la maîtrise des textes réglementaires, les protocoles d'authentification (apposition du sceau, formules d'usage) et le respect de la déontologie (secret professionnel, inviolabilité du document source).

- **La couche technologique (Intégration raisonnée de l'IA) :**

Située au sommet du système, cette couche représente l'écosystème numérique (outils de TAO, corpus électroniques et grands modèles de langage) (Bowker, 2023). L'intégration est dite « raisonnée » car la couche technologique reste bridée et validée par les verrous éthiques de la couche procédurale et l'exactitude de la couche terminologique.

3. Méthodologie : Constitution du Corpus et Fine-Tuning

3.1. Approche méthodologique et typologie du corpus

L'opérationnalisation du modèle tri-couches repose sur les préceptes de la traductologie de corpus appliquée (Zanettin, 2012). Pour pallier le

statut de «faibles ressources» de la langue arabe dans les modèles de langage génériques (Dichy & Fargali, 2011), cette recherche initie la constitution d'un corpus bilingue (arabe-français) hautement spécialisé et fermé (Bowker & Pearson, 2002). Le corpus cible deux grandes catégories de documents institutionnels algériens : d'une part, les textes sources d'expertises médico-légales (rapports d'autopsie, évaluations de préjudices corporels) et, d'autre part, les documents cliniques officiels (comptes-rendus opératoires et certificats médicaux descriptifs).

3.2. Le protocole critique d'anonymisation et d'éthique documentaire

La manipulation de données médicales et judiciaires en Algérie exige le respect absolu du secret professionnel. Avant toute intégration dans l'environnement computationnel, les documents collectés subissent un protocole d'anonymisation stricte, conformément aux dispositions de la Loi n° 18-07 du 10 juin 2018 relative à la protection des personnes physiques dans le traitement des données à caractère personnel en Algérie. Ce protocole opère à double niveau : une dé-identification nominale (suppression des identités directes) et une altération sémantique périphérique (modification des indices contextuels secondaires), empêchant toute ré-identification par recoupement (Pustejovsky & Stubbs, 2012).

3.3. L'ingénierie des données : Alignement et spécialisation des LLM (*Fine-tuning*)

Une fois le corpus standardisé, les paires textuelles (arabe-français) sont alignées au niveau de la phrase (Zanettin, 2012). Ce corpus parallèle

sert de base de connaissances pour le réglage fin (*fine-tuning*) de modèles de langage (LLM) open-source (Pustejovsky & Stubbs, 2012). Contrairement à l'utilisation d'une IA générative grand public, ce processus permet de spécifier les poids du modèle pour qu'il maîtrise les équivalences rigides de la nomenclature médicale franco-arabe et de réduire le taux d'hallucination en contraignant le modèle à ne générer des suggestions qu'à partir des structures documentaires validées du corpus (Moorkens *et al.*, 2018).

4. Ingénierie Pédagogique et Modèle Hybride

4.1. Didactique de la traduction médicale officielle : Les approches pédagogiques actives

La complexité de la traduction médicale assermentée exige une rupture avec les modèles d'enseignement traditionnels linéaires. Pour opérationnaliser notre modèle, l'ingénierie didactique s'appuie sur le constructivisme social (Kiraly, 2000) et privilégie deux approches actives :

- **L'Apprentissage par Problèmes (APP) :** Les apprenants sont placés en immersion face à des scénarios médico-légaux réels et complexes. L'étudiant doit explorer le réseau conceptuel et naviguer dans les contraintes procédurales de l'écosystème algérien (Kiraly, 2000).
- **Le Micro-apprentissage (*Micro-learning*) :** À travers des modules courts, les étudiants s'exercent spécifiquement à la détection des erreurs à haut risque, à la vérification terminologique fine et à la gestion des protocoles d'authentification.

4.2. Le modèle de travail hybride et le concept de « l'humain dans la boucle»

L'originalité de cette modélisation réside dans la formalisation d'un environnement de travail hybride, structuré autour du paradigme de *l'Humain dans la boucle* (*Human-in-the-loop*) (Bowker, 2023). Dans ce dispositif, l'intelligence artificielle n'intervient jamais en tant qu'agent décisionnel autonome, mais comme une *assistance cognitive* au service du traducteur. L'IA assure la pré-traduction et garantit l'homogénéité terminologique globale (Moorkens *et al.*, 2018). Ensuite, l'expert humain intervient pour réguler le texte, évaluer la pertinence contextuelle et corriger les nuances. Le modèle réaffirme ainsi un principe juridique cardinal : seul le traducteur officiel assermenté valide l'acte, assume la responsabilité pénale et civile, et lui confère sa dimension légitime (Monjean-Decaudin, 2012).

5. Conclusion et Perspectives

5.1. Synthèse des contributions

Cette recherche-action a permis de formaliser un cadre prospectif de modélisation pour la formation en traduction médicale officielle en Algérie face aux défis de la transition numérique. En structurant notre approche autour d'un modèle systémique tri-couches, nous avons démontré qu'il est possible d'intégrer les avancées de l'intelligence artificielle sans aliéner les fondements déontologiques de la profession. La constitution d'un corpus bilingue arabe-français fermé, sécurisé par un protocole d'anonymisation strict (Loi 18-07), offre une solution viable pour limiter le risque

d'hallucination des modèles et garantir l'exactitude terminologique indispensable aux actes médicaux légaux.

5.2. Perspectives et feuille de route institutionnelle

Ce travail dresse une feuille de route prospective pour l'établissement d'un véritable écosystème collaboratif en Algérie. L'intégration réussie de ces technologies exige le renforcement des passerelles institutionnelles entre trois acteurs clés : l'Université (pour piloter l'ingénierie linguistique et la formation), les Autorités de Santé Publique (pour valider la conformité des ontologies médicales) et les Instances Judiciaires (pour veiller à l'adéquation des outils avec les cadres réglementaires de l'assermentation). En unissant ces compétences, l'Algérie se dotera d'un dispositif moderne, capable de sécuriser les données des patients, de garantir la fiabilité absolue des actes officiels et de répondre aux impératifs conjoints de la justice et de la santé publique

Bibliographie:

Bowker, L. & Pearson, J. (2002). *Working with Specialized Discourse: A Practical Guide to Using Corpora*. London/New York: Routledge. (La base méthodologique pour justifier la constitution de corpus fermés et alignés pour la formation).

Dichy, J. & Fargali, A. (dir.) (2011). *Les technologies de la langue arabe*. Paris: L'Harmattan. (Crucial pour le public du séminaire et le Conseil Suprême de la Langue Arabe : cet ouvrage traite des spécificités morphosyntaxiques de l'arabe face à l'informatique).

Zanettin, F. (2012). *Translation-Driven Corpora: Corpus Linguistics and Translation Studies*. Manchester: St. Jerome. (Pour appuyer l'usage des corpus parallèles bilingues comme outils d'aide à la décision).

Bowker, L. (2023). *De-mystifying Translation: Bringing Artificial Intelligence and Language Technology to the Masses*. London: Routledge. (Un texte très moderne qui théorise parfaitement le rôle de l'IA comme assistance cognitive pour l'humain).

Moorkens, J., Castilho, S., Gaspari, F. & Doherty, S. (dir.) (2018). *Translation Quality Assessment: From Machine Translation to Human Evaluation*. Berlin: Springer. (L'ouvrage parfait pour l'Axe 4 du colloque, traitant des hallucinations technologiques, des prompts et de la post-édition critique).

Pustejovsky, J. & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning*. Sebastopol: O'Reilly. (Pour donner une assise scientifique solide à votre protocole d'ingénierie des données et de *fine-tuning*).

Kiraly, D. (2000). *A Social Constructivist Approach to Translator Education: Empowerment from Theory to Practice*. Manchester : St. Jerome. (La référence absolue pour légitimer l'Apprentissage par Problèmes (APP) et l'autonomisation critique de l'étudiant face aux outils).

Loi n° 18-07 du 10 juin 2018 relative à la protection des personnes physiques dans le traitement des données à caractère personnel (Journal Officiel de la République Algérienne). (À citer absolument dans votre article pour prouver l'ancrage légal algérien de votre protocole d'anonymisation stricte des données médicales).

D'un traducteur à un post-éditeur : Changement des rôles du traducteur pour une traduction précise et adaptée, cas des textes spécialisés traduits vers l'arabe par le biais des outils numériques

Zine Elabidine Benhaddi

Université de Mustapha Stamblouli -Mascara-

Laboratoire TICELET Université de Hassiba Ben Bouali,Chlef

Résumé :

La traduction automatique représente, depuis plusieurs décennies, un domaine de recherche vital, ayant connu des progrès technologiques notables grâce à l'évolution de l'intelligence artificielle. En outre, dans le domaine de la traduction spécialisée, la traduction automatique offre des avantages notables en matière de rapidité et d'efficacité, permettant même aux traducteurs spécialisés d'exploiter les systèmes automatiques pour gérer une grande quantité de documents spécialisés en un temps réduit. En réalité, la post-édition a une importance spécifique. C'est une méthode incontournable qui allie la performance de la machine à l'exactitude de l'homme. De cela, le traducteur doit obligatoirement se positionner comme un post-éditeur. Ce dernier aura pour mission de réviser, ajuster, corriger, remanier et relire le texte généré par un système de traduction automatique. C'est pourquoi, dans cette

analyse descriptive et analytique, nous cherchons à souligner le rôle du post-éditeur. L'objectif de cette mission de révision et d'ajustement des traductions automatiques est d'améliorer la qualité des traductions, surtout lorsqu'il s'agit de traduire des textes spécialisés du français vers l'arabe. Les résultats démontrent que la post-édition est essentielle pour assurer des traductions exactes et appropriées aux domaines spécialisés, mettant en lumière l'importance capitale du traducteur humain dans la procédure de traduction automatique. Ceci démontre que les rôles du traducteur sont renouvelables.

Mots clés : Rôles, traducteur, post-édition, traduction automatique, intelligence artificielle.

Introduction

La traduction occupe une place primordiale et joue un rôle vital dans la communication et l'interaction entre les nations et les civilisations à travers l'histoire. Elle constitue le fondement de toute civilisation et de tout progrès scientifique et culturel. L'humanité n'avait que la traduction comme un moyen pour savoir et découvrir l'autre.

Le domaine de la traduction dans le troisième millénaire est témoin d'une percée technologique sans précédent qui a remodelé les pratiques professionnelles et académiques. La traduction n'est plus une activité langagière statique fermée aux entretiens verbaux, mais plutôt une activité vivante et dynamique qui s'adapte très bien à des transformations rapides et renouvelle constamment ses mécanismes cognitifs et ses outils pour répondre aux exigences de l'ère numérique. L'émergence de

l'intelligence artificielle et des algorithmes de traduction automatique neuronale (machine neuronale) n'a pas seulement changé la vitesse de production et le volume des flux de texte, mais a également profondément impacté la structure (traduction professionnelle du traducteur remodelant les piliers de son identité et de sa place dans la société cognitive et économique.

Le plus pénibleproblème dans ce contexte est lorsqu'il s'agit de traiter des textes spécialisés, qu'ils soient médicaux, techniques, juridiques ou économiques, orientés vers la transmission à la langue arabe. Ces textes, par leur nature, vont au-delà de la structure linguistique superficielle pour s'engager dans des contextes épistémologiques et terminologiques très complexes, et nécessitent de prendre en compte les spécificités structurelles et sémantiques du système linguistique arabe, qui se caractérise par sa richesse dérivée et sa grande sensibilité contextuelle. D'où la question fondamentale : comment concilier la rapidité des technologies numériques avec l'exactitude de la performance humaine ? Et comment le rôle du traducteur est-il passé d'un créateur de texte à zéro à un « éditeur ultérieur » et à un adaptateur intelligent de sortie machine ?

Cette intervention vise à fournir une lecture critique analytique approfondie de cette transformation fonctionnelle et sociologique, basée sur des modèles pratiques et des applications : Montrer comment l'édition (postédition) ultérieure comble les lacunes sémantiques et lexicales laissées par les systèmes de traduction automatique, et comment le traducteur établit son statut d'élément souverain indispensable dans la production et le contrôle de qualité du savoir humain.

Zine Elabidine Benhaddi

Cette intervention scientifique aborde la transformation fondamentale et profonde de l'identité et des rôles fonctionnels du traducteur contemporain dans le contexte de l'accélération de la marée numérique et de la révolution de la traduction automatique neuronale. Le traducteur n'est plus seulement un transporteur linéaire de textes entre les langues, mais il joue désormais un rôle central en tant qu'éditeur ultérieur et adaptateur contextuel afin d'assurer l'exactitude des textes médicaux, juridiques et techniques spécialisés ainsi que leur pertinence pour le système linguistique et culturel arabe.

Sur la base de la littérature de la traduction moderne et de la comparaison des systèmes automatisés avec les capacités humaines, la recherche met en évidence les problèmes de l'utilisation excessive de la technologie, et examine les nouvelles compétences requises sur le marché du travail, tout en soulignant la solidité de l'élément humain, qui demeure la seule garantie pour la production de connotations contextuelles conscientes au-delà des significations lexicales superficielles de la machine.

1- Le processus de traduction automatique et les pièges de l'excès et de l'abandon :

La traduction automatique a traversé des étapes historiques et des développements structurels critiques. Les systèmes traditionnels basés sur la grammaire sèche et les systèmes statistiques identiques produisent des traductions littérales déformées manquant de cohérence synthétique. Avec l'avènement des modèles neuronaux basés sur les réseaux d'apprentissage profond et l'intelligence artificielle, la machine a fait un

bond en avant dans la simulation de modèles linguistiques humains et la génération de textes apparemment cohérents et sans faille. Ce développement a fait des outils numériques un partenaire quotidien dans les milieux académiques et professionnels.

L'idée de la mécanisation de la traduction remonte au XVIIe siècle (Descartes, Leibniz, Wilkins) et plusieurs modèles de « machines à traduire » ont été conçus dans les années 1930 (Artsouni en France et Smirnov-Troyanskii en URSS), mais c'est indiscutablement l'invention de l'ordinateur qui a provoqué l'essor de la traduction automatique (TA) après la Seconde Guerre mondiale[1].

La traduction automatique, née à la fin de la seconde guerre mondiale vue le défi de l'importance stratégique de renseignements au sein de l'armée, est défini par Frédérique LAB comme un système informatique qui a[2] :

- Pour entrée un texte « t1 », ou texte source écrit dans une langue « L1 » ou langue d'origine, et n'ayant pas subi d'aménagements spéciaux préalables au traitement automatique qu'il va subir.
- Pour sortie un texte « t2 », ou un texte traduit écrit dans une langue « L2 » ou langue cible, tel qu'il n'ait pas à subir de transformation pour être reconnu par les utilisateurs comme une traduction du texte t1.

1-1 L'illusion de la dispense et le phénomène de l'excès :

Selon Anne Marie LOFFLER LAURIAN le développement de la

[1] Harold SOMERS, Computers and Translation: A translator's guide, Benjamins Publishing Company, Amsterdam/Philadelphia, 2003, p11.

[2] Frédérique LAB , La traduction Automatique, le bulletin de L'EPI , N°52, pp1-2.

traduction automatique se fonde sur une possibilité technologique nouvelle à exploiter et sur une nécessité politique et économique d'échanges[3].2 La TA est née au milieu du vingtième siècle aux états unis sous l'impulsion de la défense américaine soucieuse de posséder des systèmes de cryptographie et de traduction susceptibles de faciliter le renseignement en langues étrangères dans le contexte de la guerre froide naissante dés la fin des années 1940, Warren WEAVER produit un Mémorandum qui pose la question de la faisibilité de la « Mechanical Translation[4] ».

Les études vont donner lieu à la réalisation à l'université de Washington de la première « machine à traduire » dédiée au traitement automatique des langues anglaise et russe, the US Air Force Automatic Language Translator, désigné sous le nom de code Mark[5]. A cet égard l'histoire de la traduction automatique est devisée en quatre grandes périodes : les années 50-60, le rapport ALPAC (1966), et ses conséquences, la reprise des années 70-80 et la période actuelle

Cet afflux technologique fulgurant a conduit deux groupes à tomber dans de graves pièges cognitifs, une classe qui pratique la « sur-utilisation » (Principalement des non-spécialistes et des chercheurs dans d'autres domaines scientifiques, qui ont été impressionnés par la douceur

[3] Anne Marie LOFFLER LAURIAN, La traduction automatique, Septentrion presses universitaires, France, 1996, p13.

[4] Mathieu GUIDERE, La traduction arabe méthodes et applications, Ellipse, France, 2005, p106.

[5] Mathieu GUIDERE, Introduction à la traductologie, Traducto, deBoeck, Bruxelles, 2010, p148

apparente du texte de la machine neuronale, croyant en la capacité d'une machine complète à transmettre des connaissances sans avoir besoin de contrôle humain. Cet excès comporte de grands risques, en particulier dans des domaines sensibles tels que la médecine et le droit, où une fine erreur de nom ou de conception du contexte d'une phrase peut conduire à des désastres cognitifs ou pratiques, puisque la machine ne lit pas entre les lignes, et ne comprend pas les intentions implicites, mais s'appuie sur le calcul des probabilités statistiques pour prédire le mot suivant.

1-2- La tendance au déni et le phénomène de l'abandon :

D'autre part, il y a une autre catégorie qui pratique la « sous-utilisation » (ou le déni absolu), les traditionalistes qui voient dans la machine une menace imminente à l'authenticité culturelle et effacent le contact humain. Ils refusent d'intégrer des outils numériques dans leurs pratiques, ce qui les isole du rythme de l'époque et réduit leur compétitivité sur un marché de la traduction caractérisé par la production de masse, les contraintes de temps et le réalisme nécessitent une position intermédiaire qui considère la machine comme un assistant productif fort, à condition que sa production soit soumise à la règle de raison

Seul l'être humain peut penser de manière critique et déchiffrer au-delà du texte.

2-La post-édition et son importance dans les textes spécialisés :

La post-édition est définie comme un processus d'examen, de correction, d'affinage et d'évaluation des textes produits par les systèmes de traduction automatique, dont l'objectif principal est d'améliorer la

qualité du texte et de le rendre conforme aux normes linguistiques et idiomatiques approuvées, et d'assurer son flux naturel dans la langue cible telle qu'elle a été initialement écrite.

La post-édition est une discipline relativement jeune qui a vu le jour en conjonction avec le développement de la traduction automatique dans les années 1950. Comme les traductions produites par les premiers systèmes de TA n'étaient que d'une qualité modeste, l'intervention d'un réviseur humain était nécessaire pour rendre utilisables les résultats obtenus.

A l'aide d'un système de TA adapté, la documentation est traduite automatiquement pour fournir une première version en langue cible. Un traducteur humain, appelé aussi «postéditeur», va alors procéder à la révision de la version traduite et lui apporter les corrections nécessaires, dans le but de la rendre publiable[6].

Sachant que, le volume de documents à post-éditer ne cesse d'augmenter, puisque la post-édition traite un volume de mots supérieur à celui en traduction, pour répondre aux besoins du marché[7].

« La post-édition n'est pas seulement une correction d'orthographes superficielles ou de fautes grammaticales, mais une

[6] Blain. Frédéric BLAIN, Modèles de traduction évolutifs, Thèse de doctorat en informatique, université de Maine, France, 2013, p54.

[7] Anne-Marie ROBERT, Vous avez dit post-éditrice ? Quelques éléments d'un parcours personnel, The Journal of Specialised Translation, 2013, p4.

reconstruction cognitive et contextuelle qui assure l'adaptation du texte à l'environnement culturel et délibératif du destinataire. »

La post-édition devient plus importante et plus précieuse lorsqu'il s'agit de textes spécialisés traduits en arabe. Le texte médical ou juridique contient un réseau de termes à connotation stricte et monosémique dans sa langue d'origine, mais lorsqu'il est transféré en arabe, il peut y avoir plusieurs entretiens ou l'absence d'un terme unifié, et c'est ici qu'intervient le rôle de l'éditeur ultérieur qui possède une compétence objective. (Compétence disciplinaire) de choisir le terme le plus précis et approprié pour le contexte cognitif.

Par ailleurs, les acteurs de la post-édition sont multiples tels que[8] :

• Grands comptes et multinationales amenés à gérer d'énormes volumes de traduction dans un grand nombre de combinaisons linguistiques.

• Grandes institutions internationales comme l'Union européenne qui a lancé dès 2003 un appel d'offres concernant des travaux de post-édition pour couvrir ses énormes volumes de traduction.

• Editeurs et fournisseurs d'outils adaptés aux besoins spécifiques de la post-édition.

• Agences de traduction et de localisation proposant des services de post-édition.

[8] Anne-Marie ROBERT, La post-édition : l'avenir incontournable du traducteur ? , Société française des traducteurs, p138.

• Post-éditeurs, terminologues, traducteurs, réviseurs et linguistes-développeurs qui participent en tant que langagiers à cette activité.

La post-édition se situe à mi-chemin entre la traduction et la révision. La post-édition ne consiste en effet pas à traduire puisqu'une pré-traduction sert de base au travail à effectuer, ni à réviser puisqu'il ne s'agit pas d'une traduction humaine, mais d'une pré-traduction[9].

Le post-éditeur effectue quatre types d'activités[10]:

- Il révise les phrases issues de la mémoire de traduction préalablement traduites par des traducteurs.

- Il met à jour les phrases bénéficiant de remontées de mémoire partielles de la mémoire de traduction.

-Il post-édite les phrases ne bénéficiant pas de remontées de mémoire totales ou partielles qui ont été produites par des moteurs de pré-traduction automatique sophistiqués.

-Il relit le tout pour harmoniser, articuler et finaliser l'ensemble.

3- Les erreurs mécaniques dans le transfert de textes spécialisés vers l'arabe :

Une définition plus large de la traduction spécialisée peut être perçue chez Daniel GOUADEC à travers sa spécification des différents types de matériaux traités par un traducteur considéré comme spécialisé, il dit à cet égard : «... est spécialisé, par conséquent, tout traducteur traitant

[9] Voir : Anne-Marie ROBERT, Op.cit, p5.

[10] Anne-Marie ROBERT, Op.cit, p5.

exclusivement ou prioritairement un matériau qui relève d'un genre ou d'un type spécialisé : contrats, nomenclatures, brevets,...et/ou se rapporte à un champ ou domaine spécialisé pointu : traduction de matériaux dont les sujets renvoient aux domaines du droit, de la finance, de l'informatique, des télécommunications, ect. Se présente dans des formes et sur des supports particuliers supports multimédia, film, vidéo..,etc. Appelle la mise en oeuvres de procédures et/ou d'outils, de protocoles ou de techniques spécifiques : traduction de logiciels, traduction de matériaux multimédias[11]»

Quant à Christine DURIEUX elle considère que la traduction spécialisée se caractérise par son utilité ainsi que par son rôle de transmettre l'information contenue dans le texte source en utilisant des termes qui soient de nature à permettre à l'utilisateur de la traduction de lire et comprendre le texte et de réagir par la suite, car le destinataire de ce genre de traduction n'est pas seulement un lecteur mais aussi un utilisateur de l'information fournie[12].

Les pratiques de la traduction spécialisée évoluent de pair avec la révolution numérique qui façonne un monde où la technologie à un véritable impact, notamment en termes de performance, de gain de temps et d'assurance de la qualité.

La structure unique de racine et de dérivation de la langue arabe et sa dépendance totale au contexte et aux géminations arabes, souvent absents

[11] Daniel GOUADEC, Faire traduire, La maison du dictionnaire, Paris, 2004, p 78.

[12] Christine DURIEUX, Les Langues de Spécialité comme outil d'apprentissage dans la formation en traduction spécialisée , AL - MUTARGĪM, Vol. 18, N° 1, juin 2018 p285.

des textes écrits, mettent les systèmes de traduction automatique en difficulté permanente. Les manifestations les plus notables de ces erreurs peuvent être observées à plusieurs niveaux :

3-1-Les erreurs sémantiques sont les plus graves, lorsque la machine ne parvient pas à déterminer le sens voulu du mot polysémie (polysémie) dans un contexte particulier. Par exemple, le mot « culture » dans un script biologique signifie « ferme bactérienne », tandis qu'une machine peut automatiquement le traduire en « ثقافة », corrompant ainsi complètement la signification scientifique.

3-2-Les erreurs syntaxiques et grammaticales : la machine a tendance à imiter l'ordre des mots dans la langue source (comme l'anglais), produisant des textes arabes hybrides qui commencent toujours par l'acteur ou incluent des structures lâches telles que « par » ou « من طرف أو من قبل » de manière excessive et contraire à l'esthétique du récit arabe original.

3-3-Les erreurs d'incohérence terminologique : la machine peut traduire un seul terme en différents mots au sein du même texte. Le terme en arabe "الاستمرارية" se traduit par « durabilité » et parfois par « continuité », ce qui déroute le lecteur spécialisé et perturbe l'unité de la connaissance du texte.

4- La transformation fondamentale du rôle du traducteur et ses dimensions multi sociologiques :

Face à ces faits, le traducteur contemporain s'est débarrassé du voile de la dépendance classique au texte. Le traducteur, tel qu'il a été historiquement décrit comme le serviteur de l'écrivain original, n'était plus un acteur sociologique et culturel doté du pouvoir de la prise de décision textuelle et

de la modification cognitive. Les rôles nouvellement créés de l'éditeur ultérieur ou le post-éditeur peuvent être limités aux dimensions suivantes :

4-1-Traducteur en tant que terminologue:

L'éditeur ultérieur ne traite pas le terme comme un modèle prêt à l'emploi, mais plutôt le dissout et examine ses connotations historiques, sociales et psychologiques. Il comprend comment le terme médical ou technique évolue au fil du temps, et choisit parmi des alternatives traditionnelles ou modernistes ce qui convient à la conscience du lecteur arabe contemporain, ce qui fait de lui un analyste archéologique des concepts.

Par exemple, la terminologie qui représente des connaissances spécialisées pour les disciplines techniques et scientifiques. Selon Cabré les termes sont des unités de forme et de contenu qui appartiennent au système d'une langue déterminée ils ont une fonction fondamentalement référentielle et il y a cinq facteurs principaux qui servent à distinguer les termes spécialisés du lexique commun[13] :

a) la fonction.

b) le domaine.

c) les utilisateurs.

d) les situations de communication.

e) les types de discours.

[13] Voir : Maria Teresa CABRÉ, La terminologie : Théorie, méthode et applications, Les Presses de l'Université d'Ottawa, 1998 , pp 149-192.

4-2-Le traducteur comme adaptateur (localisateur) :

La traduction automatique transmet des mots, mais le traducteur localise les concepts. Le rôle de l'adaptateur dans la reformulation d'exemples, d'expressions idiomatiques, de nombres et de mesures conformément à l'environnement délibératif et aux systèmes normatifs du monde arabe, atteindre l'exigence de lisibilité et de palpabilité, et éviter les chocs culturels ou la confusion conceptuelle.

4-3-Traducteur comme négociateur et médiateur politique et culturel

Le mot traduit n'est jamais neutre, mais porte avec lui des dimensions idéologiques et civilisationnelles. Ainsi, l'éditeur ultérieur exerce le rôle de « négociateur » qui protège l'identité culturelle et linguistique de la langue cible (l'arabe), et contrecarre les tentatives de dominer ou d'envahir la culture qui accompagne les flux de connaissances de la langue occidentale en sélectionnant des formulations qui préservent la dignité et la sobriété de la langue cible.

Daniel GOUADEC considère que la prestation de traduction inclut une prestation de la part des traducteurs, qui inclut à son tour un ou des processus de traduction. Il distingue quatre phases, à savoir[14] :

-Phase d'attente et de prospective: cette phase inclut le démarchage, l'information, l'auto-formation, la formation, l'acquisition et l'optimisation de ressources et savoir- faire, la définition et la promotion de l'offre de services, l'obtention de certifications. Et après la prestation de traduction la

[14]Daniel GOUDEC, Modélisation du processus d'exécution des traductions, Meta : Journal des traducteurs, vol.50, n° 2, 2005, p644.

mise en place des savoirs, savoir-faire, ressources et autres dérives qui peuvent influencer sur l'exécution de la traduction de façon directe ou indirecte.

-Phase de prétraduction : elle inclut toutes les tâches et interactions qui interviennent dans la «prestation de traduction» ou dans la «prestation de traducteur» déterminant ou encadrant la phase de traduction.

-Phase de traduction : elle débute avec la prise en charge de la mission de traduction, de localisation, ou autre et couvre le transfert d'un «système conceptuel-culturel et d'un système de représentation à un autre ».

-Phase de post-traduction : Cette phase comporte une partie spécifique de clôture de la mission de traduction et de préparation des traductions futures avec l'enrichissement des bases de données.

A cet égard nous constatons que les outils technologiques peuvent intervenir dans toutes ces phases. Par exemple le Groupe de recherche Tradumàtica de l'université de Barcelone a élaboré un schéma qui permet de mieux visualiser ce constat. Ce schéma qui s'inspire des étapes de traduction proposées par la théorie fonctionnaliste de la traduction, et en particulier, par les travaux de Katarina Reiss, décompose le processus de traduction en différentes phases informatisées[15].

-Phase d'obtention du texte : obtention du texte à traduire en format numérique.

[15] Daniel GOUDEC, Op.cit, p644.

-Phase d'analyse : évaluation du texte pour obtenir des informations diverses telles que le registre linguistique, la difficulté thématique et terminologique, le taux de redondance, le formatage, etc.

-**Phase de documentation :** recherche de solutions aux problèmes détectes lors de la phase d'analyse, principalement les problèmes de terminologie et de compréhension du texte.

-**Phase de traduction :** traduction du texte original dans la langue cible, en tenant compte des solutions et des propositions obtenues lors de la phase de documentation.

-**Phase de révision :** élimination des erreurs éventuelles et optimisation du texte sur le plan linguistique.

-Phase finale du traitement du texte : traitement du texte dans les aspects les plus formels de l'édition, tels que l'orthotypographie, le style, le traitement de l'image, etc.

Ces six phases requièrent la maitrise d'une ample gamme d'outils technologiques qui peuvent, dans certaines étapes, différer selon le type de traduction réalisée. C'est pour cela les formations actuelles doivent les intégrer dans leurs contenus pédagogiques afin d'apprendre aux futurs traducteurs non seulement l'élaboration d'un corpus spécialisé représentatif, mais aussi les stratégies de traitement de données pour qu'ils puissent formuler des requêtes pertinentes.1

5- La résilience de l'élément humain face au rampement numérique :

Les forums technologiques et économiques discutent souvent la disparition imminente de la profession de traducteur et des solutions d'intelligence artificielle pour remplacer complètement l'humain. Cependant, une lecture attentive et approfondie de la réalité de la pratique prouve la futilité de cet argument. La constance et la stature du traducteur humain découlent de l'unicité de l'esprit humain face aux « significations marginales, ombres sémantiques et contextes culturels et historiques entourant les textes, des dimensions qui vont au-delà du réductionnisme mathématique ou de la modélisation algorithmique.

La machine manque de « conscience contextuelle » et d'« intuition linguistique », elle ne ressent pas le ton stylistique et n'est pas consciente de l'ironie ou de l'allusion politique, et elle est incapable de rendre moral ou (tonalité) décisions juridiques envers le texte traduit, de sorte que la relation réelle est une relation d'intégration consciente et d'intégration fonctionnelle : la machine donne vitesse et abondance et l'humain donne esprit, précision et légitimité cognitive.

6-Analyse pratique de quelques traductions via l'outil numérique:

Notre analyse consistera à soumettre trois types de textes (médical, économique et juridique) à la traduction automatique, suivie d'une phase de post-édition, en confrontant les performances de deux outils : Google Traduction et Reverso.

1. Texte médicale[16] :

Nouveau pas vers un vaccin tueur de tumeurs

Rêvons à un âge ultérieur. En ce temps-là, il y aurait, contre les cancers, un traitement, un traitement simple et efficace. Une simple piqûre. Le médecin injecterait au malade une molécule chimique qui agirait comme le bon vieux vaccin antirabique. La substance aurait les vertus de déchaîner contre la source du mal - en l'occurrence, les cellules cancéreuses- les foudres du système immunitaire. Celui-ci, sous l'effet du produit injecté, saurait soudain reconnaître, entre toutes les cellules, les malignes à abattre. On appellerait ce traitement un vaccin contre-cancer. Vers ce rêve auquel travaillent des scientifiques depuis une bonne dizaine d'années, un nouveau pas vient d'être franchi.

Deux équipes de l'Institut Pasteur conduites par deux chercheuses Claude Leclerc (vaccinologue) et Sylvie Bay (chimiste), annoncent, dans la dernière livraison du Journal of Immunology , avoir mis au point un composé chimique capable de provoquer le rejet des tumeurs. Les essaies ont été faits sur les dizaines de souris développant des cancers. 80% d'entre elles ont réussi après injection de ce composé, à ce débarrasser de leurs tumeurs.

[16]Mathieu GUIDERE , Manuel de traduction français-arabe/arabe-français, Ellipse, Paris, p 170

Exemple n°01 : un vaccin tueur de tumeurs

Nous avons remarqué que les deux systèmes de TA « Google Traduction » et « Reverso » ont traduit littéralement en arabe le mot « tueur » par " لقتل " et " قاتل ". En fait, le mot « tueur » est un adjectif qualificatif du nom « vaccin » donc il fallait trouver un équivalent approprié au sens du contexte médical qui est le mot « contre ». Donc nous pouvons proposer la traduction suivante :

"-خطوة جديدة لإيجاد / نحو لقاح ضد الأورام."....

Notamment, « Google Traduction » a traduit le mot « les essaies » en arabe par " "الاختبارات qui est une traduction erronée qui réduit complètement le sens du contexte car nous parlons ici des essaies qui ont été effectués sur des souris, cependant, « Reverso » a réussi à le traduire par " التجارب ". Donc nous pouvons proposer la traduction suivante :

"-أجريت تجارب على عشرات الفئران المصابة بالسرطان."...

Exemple n°02 : « Deux équipes de l'Institut Pasteur conduites par deux chercheuses Claude Leclerc (vaccinologue) et Sylvie Bay (chimiste), annoncent, dans la dernière livraison du Journal of Immunology ».

Traductions vers l'arabe :

-أعلن فريقان من معهد باستور بقيادة الباحثين كلود لوكلير(عالم اللقاحات) سيلفي باي (الكيميائي)، في العدد الأخير من مجلة علم المناعة.

Les deux systèmes ont commis d'autres erreurs en traduisant par exemple « par les deux chercheuses » par " بقيادة الباحثين " alors qu'en arabe la forme féminine existe donc on devrait dire " بقيادة الباحثتين ".De plus, en

traduisant « vaccinologue » par " عالم اللقاحات " et " عالم اللقاح " ainsi que « chimiste » par " الكيميائي " au lieu de les traduire correctement en arabe par "عالمة كيمياء " et " "أخصائية في اللقاحات: car se sont des noms de métiers. Cela explique le fait que les systèmes de TA ont parfois du mal à faire la distinction de genre, notamment lorsqu'il s'agit de la traduction du français vers l'arabe.

Exemple n°03 : « La substance aurait les vertus de déchaîner contre la source du mal en l'occurrence, les cellules cancéreuses- les foudres du système immunitaire[17] ».

Traduction vers l'arabe :

-ومن شأن المادة أن تتمتع بفضائل إطلاق العنان لغضب الجهاز المناعي ضد مصدر الشر - في هذه الحالة، الخلايا السرطانية.

Les erreurs stylistiques se produisent lorsque le choix des mots, la structure des phrases ou d'autres aspects linguistiques ne correspondent pas au style attendu pour un contexte donné et c'est le cas de l'expression « les vertus de déchaîner contre la source du mal » traduite par :

فضائل إطلاق العنان ضد مصدر الشر ."

Cela montre l'absence totale de l'effet esthétique produit dans le texte original.

Le rôle du traducteur humain autant que post-éditeur ne peut être remplacée par la machine car cette dernière ne pourra jamais déchiffrer les

[17] Mathieu GUIDERE , Manuel de traduction français-arabe/arabe-français, Ellipse, Paris, p 170.

caractéristiques linguistiques, les connotations terminologiques et les charges sémantiques liées aux textes spécialisés.

2. Texte économique[18] :

Domaine économique

Les marchés asiatiques résistent pour l'instant à la déprime boursière

Malgré la faiblesse de Tokyo, les places asiatiques affichent des performances positives depuis le 1er janvier.

Le décalage horaire a parfois du bon. Mercredi, en début d'après-midi, quand les investisseurs européens ont commencé à se demander s'ils ne vivaient pas le krach tant redouté, nombre de gérants étaient sans doute soulagés de savoir les marchés asiatiques endormis. La crise n'aura finalement pas eu lieu et hier, les places asiatiques auront connu une journée étonnamment calme – Taipei a fini en hausse de 1,5 %, à 5.742.74 points, Séoul en légère baisse (moins 0,3 % à 541,83 points) – à l'image de leur comportement depuis le début de l'année. Alors que le CAC hexagonal et le S&P 500 new-yorkais affichent des reculs supérieurs à 10 %, ces deux marchés ont l'impudence d'afficher des progressions de respectivement 21 % et 7 % en 2001, sans parler des 10 % gagnés par la Bangkok ou des 50 % du marché chinois coté en devises. Les spécialistes n'en restent pas moins extrêmement prudents.

Chasse aux bonnes affaires. « Par rapport à 1998, ces pays n'ont que peu d'inflation, ils diminuent même leurs taux d'intérêt, le danger vient cependant de leur surexposition à l'économie américaine », explique Nicholas P. Sargen, stratège chez JP Morgan Private Banking à New York. « Les investisseurs internationaux ont tellement liquidé l'an dernier qu'ils se permettent aujourd'hui du *bottom fishing*, de la chasse aux bonnes affaires », ajoute Lan Wang Simond, analyste chez Pictet & Cie à Genève.

Hong Kong en comparaison affiche une performance décevante (moins 10 % depuis le début de l'année), déséquilibrée par quatre sociétés, HSBC, China Mobile, Hutchison Whampoa et Cheung Kong Holdings, qui représentent 62 % de l'indice Hang Seng. « China Mobile, que les investisseurs ont longtemps conservé en dépit de la baisse du secteur des télécoms, subit aujourd'hui des ventes de capitulation », explique la spécialiste. Il n'est pas certain que de tels mouvements de vente ne touchent à leur tour Taipei ou Séoul, si la croissance américaine s'avère incapable de relancer la demande de composants électroniques, pilier de ces économies.

La Tribune du 16/03/2001

-Exemple n°01 :

-Source (FR): « Mercredi, en début d'après-midi, quand les investisseurs européens ont commencé à se demander… »

-Reverso: فحين بدأ المستثمرون الأوروبيون يوم الأربعاء بعد الظهر يتساءلون...

-Google traduction : « الأربعاء، في وقت مبكر من بعد الظهر، عندما بدأ المستثمرون الأوروبيون يتساءلون...»

[18] Jean Delisle, Les manuels de traduction, essai de classification, TTR, : Traduction, Terminologie, Rédaction, Erudit, p17

En français, les compléments circonstanciels du temps sont fréquemment placés au tout début de la phrase.. En arabe standard, une structure de phrase verbale est préférée (verbe + sujet + temps adjoint), déplaçant naturellement ces repères temporels (يوم الأربعاء بعد الظهر) plus profondément dans la proposition.

-Post-édition : "...إذ استهل المستثمرون الأوربيون ظهيرة يوم الأربعاء بالتساؤل عن

-Exemple n°02 :

-Source (FR): « ... explique Nicholas P. Sargen, stratège... » / « ... ajoute Lan Wang Simond, analyste... »

Reverso :

"يشرح السيد نيكولا سارجن، المخطط... / وتضيف لان وانغ سيموند، المحللة.....

-Google traduction :

نيكولاس ب. سارجن، الاستراتيجي.....يشرح." / لان وانغ سيموند، المحلل.....تضيف

Le récit français place souvent les verbes de rapport (par exemple, explique, ajoute) au milieu ou à la fin d'une citation. La syntaxe élégante de l'arabe favorise fortement l'introduction du locuteur avant son discours en utilisant des particules cohésives comme أكّد أن (confirmé) ou مضيفاً أن (ajouté), plutôt que de diviser ou de terminer les phrases avec eux.

Google traduction fonctionne principalement sur des modèles d'apprentissage profond qui traitent le texte en morceaux. Lorsqu'il est confronté à des citations entre parenthèses ou à des devis incorporés (comme indiqué dans l'exemple 02), GT rencontre des difficultés avec le flux au niveau du paragraphe. Il traduira explique Nicholas... littéralement

comme يشرح نيكولاس ... à la toute fin d'une phrase de 50 mots. En arabe, cela crée une charge cognitive sévère pour le lecteur.

Reverso utilise souvent des bases de données de mémoire de traduction localisées (corpus bilingues). Si elle correspond à un texte humain précédemment traduit, elle tirera parfois correctement une structure arabe inversée. Cependant, si aucune correspondance n'est trouvée, il revient à une copie mot-pour-mot de la syntaxe française, laissant les clauses de rapport pendantes maladroitement.

-Post-édition : pour une traduction précise, le traducteur est appelé à affiner la traduction comme suit :

"يشرح الخبير الاستراتيجي اليد نيكولا سارجين..... في حين تُضيف المحللة الإقتصادية السيدة وانغ سيموند ....."

-Exemple n°03 :

-Source (FR): « Taipei a fini en hausse de 1,5 %, à 5.742,74 points... »

-Traduction (AR):

... أغلقت سوق تايبي على 5.742,74 نقطة، أي بارتفاع نسبته 1.5%

Reverso :

«تايبيه أنهت السباق بارتفاع 1.5% لتصل إلى 5742.74 نقطة...» **Google traduction :**

Dans le journalisme financier arabe, les mouvements sont exprimés explicitement. Au lieu de dire « jusqu'à 1,5 % » littéralement, l'arabe standard utilise des structures nominales/verbales comme بارتفاع نسبته

(« avec une augmentation dont le pourcentage est… » pour préserver la lisibilité.

Google traduction rencontre fréquemment des difficultés avec la mise en page spatiale de textes mixtes arabe-français. Lorsque les nombres, les décimales et les pourcentages (par exemple, 1,5 % à 5.742,74 points) sont traduits, GT brouille souvent la direction des parenthèses ou retourne le signe de pourcentage du mauvais côté du nombre, rendant les données financières trompeuses.

Reverso fonctionne légèrement mieux avec les nombres car son interface est spécifiquement conçue pour l'alignement bilingue. Cependant, il ne parvient toujours pas à ajouter automatiquement les mots contextuels nécessaires comme نسبته (son ratio) ou بواقع (par la quantité de) avant les pourcentages, laissant le texte arabe sec et décousue.

-Exemple n°04 :

-Source (FR): « Le décalage horaire a parfois du bon. » / « Chasse aux bonnes affaires. »

-Reverso: قد يكون الفارق الزمني فأل خير / صيد الصفقات المربحة (ou اقتناص الفرص)

-Google traduction : " فارق التوقيت يكون جيدا في بعض الأحيان." / "البحث عن الصفقات."

Traduire des expressions idiomatiques aboutit littéralement à des absurdités. « Chasse aux bonnes affaires » (littéralement : « chasse aux bonnes affaires ») est localisée dans les pages d'affaires en arabe sous le

nom de اقتناص الفرص (saisir des opportunités) ou صيد الصفقات المربحة (pêcher pour des affaires profitables).

Google traduction est très **grammatical mais trop littéral**. Il est excellent pour traduire des mots individuels, mais il ne parvient pas à saisir la prose journalistique des journaux financiers. Il laisse des phrases structurées comme les phrases françaises, ce qui donne une expression arabe maladroite et non native.

Reverso est supérieur pour les **phrases idiomatiques et les collocations** en raison de sa nature de base de données contextuelle.

Le toucher humain : un traducteur doit posséder une base solide en finance. Aucun traducteur automatique ne comprend que la « pêche de fond », "bottom fishing" signifie acheter des actions sous-évaluées après une bais

3. Texte juridique[19] :

Action en justice

> L'action en justice est la possibilité donnée à toute personne de s'adresser à un juge pour obtenir le respect de ses droits en invoquant des prétentions. C'est le droit pour l'auteur de ces prétentions d'être entendu par le juge sur le fond de celles-ci afin qu'il dise si elles sont bien ou mal fondées. L'action en justice, c'est aussi le droit pour l'adversaire de discuter du bienfondé de ces prétentions ; elle concerne donc, à la fois, le demandeur et le

[19] Mohammed BENZOUBA, Lexique juridico-judiciaire, Français-arabe, Lexique édité dans le cadre du Programme d'Appui au Secteur de la Justice en Algérie, p 31.

défendeur. Souvent, on confond « action en justice » et « demande en justice ». Il est utile de faire la distinction entre ces deux notions. On vient de définir sommairement l'action en justice. Quant à la demande en justice, elle correspond en fait à l'acte par lequel une personne soumet au tribunal une prétention, c'est sa requête en quelque sorte, ou plutôt le contenu de cette requête. Dans leurs répliques, les défendeurs demandent souvent au juge de « rejeter l'action du demandeur», alors qu'il vise en réalité sa demande.

-Exemple n°01 : « On vient de définir sommairement l'action en justice »

Google traduction :

"لقد قمنا للتو بتعريف الإجراء القانوني بإيجاز".

Reverso :

"تم تحديد الإجراء القانوني على نطاق واسع"

Nous avons remarqué que les deux systèmes de TA « Google Traduction » et « Reverso » ont traduit littéralement la phrase : « On vient de définir » sans respecter le temps de conjugaison du verbe source. Par exemple « Google traduction » a opté pour la traduction suivante : " لقد قمنا للتو بتعريف" quant a « Reverso » il a traduit par : " تم تحديد" au lieu de"لقد عرفنا".

-Exemple n°02 : « L'action en justice est la possibilité donnée à toute personne de s'adresser à un juge pour obtenir le respect de ses droits en invoquant des prétentions ».

Google traduction :

"الإجراء القانوني هو الإمكانية المتاحة لأي شخص للاتصال بالقاضي للحصول على احترام حقوقه من خلال رفع الدعاوى".

Reverso :

"والدعوى القانونية هي فرصة لأي شخص لتقديم طلب إلى قاض للحصول على احترام حقوقه عن طريق الاحتجاج بالمطالبات"

« Google traduction » a réussi à traduire l'expression « la possibilité donnée » par " الإمكانية المتاحة" contrairement à « Reverso » qui l'a traduit par " الفرصة " qui est moins utilisée dans le domaine juridique.

-Exemple n°03 : « On lit parfois même dans des dispositifs de certains jugements ce prononcé : " l'action du demandeur est rejetée »

. **Google traduction :**

"بل إننا أحيانا نقرأ في أحكام بعض الأحكام هذا النص":رفض دعوى المدعي."

Reverso :

"بل إنه يذكرأحيانا في بعض الأحكام أن دعوى المدعي مرفوضة"

Nous avons remarqué dans cet exemple : « On lit parfois même dans des dispositifs de certains jugements ce prononcé » que « Google Traduction » a reproduit deux fois le même mot " أحيانا نقرأ في أحكام بعض الأحكام هذا النص" : " أحكام " ce qui crée une imprécision au style de la phrase. Pour « Reverso » il a carrément supprimé le mot « prononcé » en traduisant la phrase ainsi :

"بل إنه يذكرأحيانا في بعض الأحكام"

Post-édition :

"-أحيانا نقرأ في منطوق بعض الأحكام هذه الصيغة" :رفض دعوى المدعي.. "...

Nous observons que les types d'erreurs les plus fréquentes qui figurent dans la traduction de ces trois textes spécialisés par les deux systèmes de TA « Google Traduction » et« Reverso» sont les faux-sens, les non-sens, les ajouts, les suppressions. Ces exemples analysés visent à prouver l'importance de la post-édition et qu'en dépit du développement remarquable de la traduction automatique et l'intelligence artificielle, la traduction humaine et le rôle du traducteur humain autant que post-éditeur ne peut être remplacée par la machine.

Conclusion

Sur la base de ce qui précède, nous pouvons dire que la transition d'un traducteur à un post-éditeur et à un conditionneur contextuel n'est pas une retraite ou une diminution du traducteur humain. c'est plutôt pour élever ses tâches à des niveaux plus élevés de critique, d'analyse et de contrôle cognitif que le succès du transfert de la science et des connaissances spécialisées vers la langue arabe à l'ère numérique dépend de l'étendue de la conscience qu'ont les traducteurs de ces nouveaux rôles, et armé des compétences informatiques et linguistiques nécessaires pour diriger et diriger la machine. La position du traducteur n'a pas été sapée ni perdue, mais est devenue plus vitale et authentique, de sorte que l'élément humain est toujours la garantie souveraine et la seule garantie pour la production de documents précis, des traductions sobres et adaptées qui atteignent la communication civilisationnelle.

Références

Anne Marie LOFFLER LAURIAN, La traduction automatique, Septentrion presses universitaires, France, 1996.

Anne-Marie ROBERT, Vous avez dit post-éditrice ? Quelques éléments d'un parcours personnel, The Journal of Specialised Translation, 2013.

Anne-Marie ROBERT, La post-édition : l'avenir incontournable du traducteur ?, Société française des traducteurs.

Anne-Marie ROBERT, Op.cit, p5.

Christine DURIEUX, Les Langues de Spécialité comme outil d'apprentissage dans la formation en traduction spécialisée, AL - MUTARĞIM, Vol. 18, N° 1, juin 2018.

Daniel GOUADEC, Faire traduire, La maison du dictionnaire, Paris, 2004.

Daniel GOUDEC, Modélisation du processus d'exécution des traductions, Meta : Journal des traducteurs, vol.50, n° 2, 2005.

Daniel GOUDEC, Op.cit.

Frédéric BLAIN (Blain. Frédéric BLAIN), Modèles de traduction évolutifs, Thèse de doctorat en informatique, université de Maine, France, 2013.

Frédérique LAB, La traduction Automatique.

Harold SOMERS, Computers and Translation: A translator's guide, Benjamins Publishing Company, Amsterdam/Philadelphia, 2003.

Jean Delisle, Les manuels de traduction, essai de classification, TTR, : Traduction, Terminologie, Rédaction, Erudit.

Maria Teresa CABRÉ, La terminologie : Théorie, méthode et applications, Les Presses de l'Université d'Ottawa, 1998.

Mathieu GUIDERE, La traduction arabe méthodes et applications, Ellipse, France, 2005.

Mathieu GUIDERE, Introduction à la traductologie, Traducto, deBoeck, Bruxelles, 2010.

Mathieu GUIDERE, Manuel de traduction français-arabe/arabe-français, Ellipse, Paris.

Mohammed BENZOUBA, Lexique juridico-judiciaire, Français-arabe, Lexique édité dans le cadre du Programme d'Appui au Secteur de la Justice en Algérie, p 31.