Journal of Languages & Translation
P-ISSN: 2716-9359 E-ISSN: 2773-3505
Special Issue December 2025 pp.59-73

Evaluating Arabic-English Neural Machine Translation: Challenges Across Different Text Types

Rami BOUOUDEN[1]

Translation and Language Didactics Laboratory – Algeria

ramibouou@gmail.com

0009-0005-2183-2855

Received: *01/06/2025*	**Accepted:** *08/11/2025*	**Published:** *01/12/2025*

Abstract

Artificial intelligence and natural language processing have gained widespread recognition recently, particularly as neural machine translation (NMT) has become indispensable for translation service providers. However, despite its unprecedented technological advancements, machine translation (MT) engines continue to struggle with Arabic-English translation due to linguistic complexities, structural differences, and the limited availability of high-quality training data. These challenges are particularly evident when translating nuanced content that requires a deep understanding of context and cultural sensitivity. This paper evaluates the performance of three MT systems—Reverso, Systran, and Microsoft Azure—by analyzing their translations of three distinct text types: general, technical, and journalistic. It uses the TAUS Dynamic Quality Framework error typology to assess key aspects, including accuracy, fluency, terminology, and style. The analysis is both qualitative and quantitative, offering a comprehensive view of each system's strengths and limitations. The findings indicate that while the three MT engines generate comprehensible translations, they consistently struggle with recurring issues, such as domain-specific terminology, idiomatic expressions, and stylistic coherence. The study underscores the necessity of post-editing and highlights how MT can assist human translators in improving productivity while also emphasizing the irreplaceable value of human expertise. This research brings attention to the need to refine MT engines through domain-specific and higher-quality linguistic training resources for Arabic-English translation.

Keywords: Error typology; machine translation; neural machine translation; post-editing; TAUS framework.

[1] Corresponding author: Rami BOUDOUDEN / ramibouou@gmail.com

Introduction

Natural language processing (NLP) and artificial intelligence (AI) continue to flourish swiftly. Professionals in the translation industry continually strive to address the growing need for more reliable MT engines that can handle linguistically distant language pairs, such as Arabic and English, more efficiently. Regardless, despite their continuous breakthroughs, "automatic translation systems have not reached full independence in producing quality output that qualifies them to replace human translators." (Omar & Salih, 2024, p. 2). The advent of NMT in the era of AI has taken it to an entirely new level. However, issues such as word choice errors, spelling mistakes, and contextual inconsistencies in translations persist (Ali, 2020). Arabic, a central Semitic language of the Afro-Asiatic language family, is characterized by its rich morphology and varied syntactic patterns. It entails several challenges when using MT to translate texts, particularly from Arabic, into other languages. In particular, dealing with relative clauses presents major challenges for MT systems, especially when it comes to handling complex sentence structures (Nagi, 2023), while the limited availability of high-quality training data for Arabic, mainly when compared to other languages, hinders the development of reliable MT systems (Sajjad et al., 2020). Hence, MT evaluation has garnered considerable attention in recent years to enhance the quality of output from these engines and facilitate translators' post-editing processes. This study aims to evaluate the performance of three well-known MT systems—Reverso, Systran, and Microsoft Azure—in translating texts between Arabic and English. The researcher aims to investigate the benefits, limitations, and potential advantages of MT engines for human translators. The idea is to examine these engines' accuracy, fluency, and terminological consistency by analyzing the translation outputs of three distinct text types (organizational reports, legal administrative documents, and marketing content). The researcher employed the Translation Automation User Society (TAUS) MT Evaluation Guide intending to answer the following question:

- How effectively do Arabic-English neural machine translation systems handle different text types, and how useful are their outputs for human post-editing?

1. Background

1.1. Arabic and English MT

Arabic and English are two linguistically dissimilar tongues. Due to their structural, semantic, and grammatical differences, they pose particular challenges for MT systems. Arabic still lags behind English, as well as other European and Asian languages, in the field of Machine Translation. MT research originated in Western countries and primarily focused on English and other European languages. However, Arabic MT remains underdeveloped for several reasons, including structural and cultural differences (Almahasees, 2022). In addition, there is a lack of investment in developing more reliable Arabic-based MT engines, which leads users to rely more on already existing systems that are not fully optimized for this language. There is also a lack of training data for Arabic, and the materials used are primarily of low quality, which results in many common errors. Although many might overlook these errors, a myriad of high-quality Arabic corpora can be used to train modern MT engines instead of the reused, low-quality data scattered across the internet. Other apparent issues in this pair include word sense disambiguation, named entity recognition, and Arabic's rich morphology and complex vocabulary (Alkhatib & Shaalan, 2018; Almaaytah & Alzobidy, 2023). While innovations in NMT have surpassed older approaches, the overall quality of Arabic-English translations remains moderate and needs more refinement (Zakraoui et al., 2021). Almahasees (2022) points out that progress in enhancing Arabic machine translation has been slow. This is because the public and private sectors have not invested sufficient resources, and there has been a lack of collaboration and coordination among researchers working in the field.

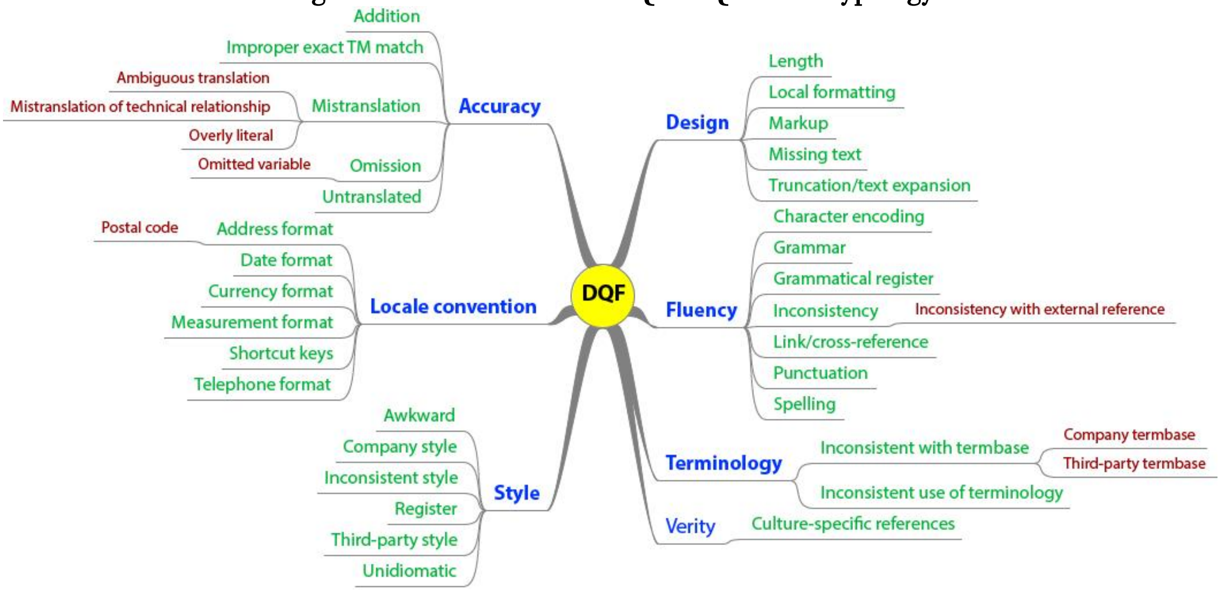
1.2 MT Evaluation

The performance of MT engines is evaluated and assessed in order to identify their strengths and weaknesses. Many methods and techniques have been used to study them from different angles. Such methods include human evaluation, automated evaluation, or a combination of both to obtain more detailed results. Every approach has its strengths and weaknesses, both of which are needed to evaluate the performance of MT technologies. Automatic methods include Metrics like BLEU (Bilingual Evaluation Understudy), METEOR, or TER (Translation Error Rate). According to Hadla et al. (2014), "The automatic evaluation of machine translation systems is based on a comparison of MT outputs and the corresponding professional human translations (Reference translations)" p.68.

On the other hand, language experts conduct human evaluations using methods like direct assessment, multidimensional quality metrics (MQM), and the TAUS Dynamic Quality Framework (DQF). These automatic metrics excel in several areas, including time savings, cost efficiency, and objectivity. However, human evaluation should always remain the benchmark, as it is needed for advancing translation and demonstrating meaningful progress (Mathur et al., 2020). Although costly and time-intensive, human evaluation appears to be the more reliable method for assessing the quality of machine translation output (Povilaitienė & Kasperė, 2022). Despite their inconveniences, both methods proved efficient in specific aspects and certain types of texts, especially when combined.

The TAUS (Translation Automation User Society) evaluation guidelines are part of a standardized framework for evaluating MT quality. These guidelines are widely recognized in the translation industry for their clear and systematic approach. The TAUS Dynamic Quality Framework (DQF), introduced in 2011, provides a flexible quality evaluation model for diverse content types and automated translation technologies (Valli, 2015). The current study uses the TAUS DQF error typology evaluation template (TAUS, 2021).[1] This error typology template is based on the harmonized DQF-MQM error typology (Figure 1). Error-typology evaluations are also employed in human translation, particularly during review procedures or to identify errors (Görög, 2014). The researcher chose this error typology evaluation method because it helps detect specific error types, allowing for the assessment of overall MT performance across different text types.

Figure 1: The harmonized DQF-MQM error typology



Source: (Valli, 2015, p.132)

[1] Template available on: https://info.taus.net/dqf-mqf-error-typology-template-download

Figure (1) displays a detailed classification pattern for all the main and subcategories of the harmonized DQF-MQM error typology the study will follow. Errors are classified into seven main types: Accuracy, Locale Convention, Style, Fluency, Terminology, Design, and verity. Each error type contains several other subcategories. This error typology may indicate that MT evaluation is complex, but it is also crucial in understanding and addressing the linguistic and non-linguistic aspects of translations. As TAUS (2021) outlined, this approach generally allows for a detailed evaluation of MT error, as its error typology ranges from broad classifications to highly detailed categories.

2. Methodology

The researcher adopted an exploratory-descriptive approach to evaluate the performance of three machine translation engines—Reverso, Systran, and Microsoft Azure—and the quality of their raw outputs across three distinct text types. Two texts were translated from English to Arabic, and one from Arabic to English. Additionally, the study aims to determine the extent to which human translators can benefit from these engines during the translation process. The evaluation and analysis process relies on the TAUS MT evaluation guidelines (TAUS, 2021), which offer a set of instructions and standards designed to assess machine translation engines based on the harmonized DQF-MQM error typology.

There are many MT systems available online. They are divided between free and paid platforms, yet some systems provide free and premium services. The advancements in AI, Neural Language Processing (NLP), and the need for more reliable systems have driven language service providers to become more competitive over the last few decades. The market domination and supremacy race have prompted developers to enhance the capabilities of their MT systems. Modern systems are now supposed to be able to deal with intricate linguistic patterns and cultural ambiguities. This study focuses on three MT systems: Reverso, Systran, and Microsoft Azure. Each has distinctive characteristics, and many language service providers adopt them in the modern translation market.

2.1 Selected Texts

The researcher chose three different types of texts, two in English and one in Arabic, to study different aspects of translation outputs and provide a performance evaluation according to each text type and language pair:

- **Text 1**: A 1,418-word English text excerpted from a United Nations (UN) report on multilingualism within the Joint Inspection Unit (JIU), titled "Multilingualism in Programmes and Operations" (pp. 25–27).[1] The text is an organizational report. It is often used to improve the UN system's efficiency by presenting inspections, investigations, and recommendations. The report evaluates strategies, policies, and practices related to multilingualism within the organization and was translated into Arabic.
- **Text 2**: Consisting of 791 words, this Arabic text is extracted from the "Employment of Foreigners in Algeria" web page on "The Algerian Agency for Investment Promotion" website.[2] This regulatory or administrative guide with a legal tone provides detailed procedural information and legal requirements for employing foreigners in Algeria, covering regulations based on employment duration, necessary documents, application processes for work permits, and social security obligations. It is intended to guide companies and foreign workers through the legal and administrative steps for employment in Algeria and was translated into English.
- **Text 3**: This text consists of 1,022 words and contains several marketing paragraphs promoting various brands and their products in English. It is part of a translation project previously undertaken by the researcher in collaboration with a translation company. This text type was

[1] Full report: https://www.unjiu.org/sites/www.unjiu.org/files/jiu_rep_2020_6_english.pdf

[2] Report web page: https://aapi.dz/ar/emploi-des-etrangers-en-algerie-ar/

chosen to test how machine translation engines handle marketing texts, particularly in translating cultural elements and idiomatic expressions from English to Arabic.

2.2 Evaluation and Analysis

After obtaining the machine-translated outputs for the three texts, the next step involved evaluation. The researcher compared these outputs and identified errors based on the TAUS Error Typology Guidelines. Errors will be categorized into primary and secondary types, and their severity—neutral, minor, or critical—will also be determined accordingly. Subsequently, these errors will be analyzed to assess the quality of the translations and the performance of the engines.

3. Results

Tables (1), (3), and (5) display the analysis results of the machine translation output of the three texts. The analysis focused on identifying the types, frequencies, and severity of translation errors.

3.1 Text 1

Table 1: MT errors in the first text (English-Arabic)

Error type	Reverso (1236 w)	Systran (1248 w)	MS Azure (1201 w)
Accuracy (total)	**13**	**22**	**10**
Addition	-	2	-
Mistranslation	12	14	9
Over-translation	1	4	1
Untranslated text	-	1	-
Fluency (total)	**15**	**9**	**12**
Punctuation	2	1	1
Spelling	-	2	1
Grammar	9	6	6
Terminology (total)	**3**	**-**	**1**
Inconsistent use of terminology	1	-	1
Style (total)	**15**	**8**	**10**
Awkward	8	3	7
Unidiomatic	7	5	3
Design (total)	**1**	**1**	**2**
Local formatting	1	1	2
Locale convention (total)	**1**	**-**	**2**
Date format	-	-	1
Total errors	**48**	**40**	**36**
Severity of errors	Neutral: 7 Minor: 33 Major: 9	Neutral: 5 Minor: 29 Major: 7	Neutral: 1 Minor: 30 Major: 5

Table 1 displays the MT errors in the “Multilingualism in the United Nations” text. Most errors fall under accuracy, fluency, and style, with additional issues involving terminology, formatting, and locale conventions. To maintain focus and ensure clarity, this research presents a carefully selected group of examples highlighting the most notable errors and issues in the MT outputs.

An accuracy error occurs when the translated text does not fully align with the source text, except for variations allowed by predefined specifications (TAUS, 2021). Table (2) shows some examples of accuracy errors across the three engines:

Table 2: Accuracy errors in the first text

Source text	Reverso	Systran	MS Azure	Suggested translation
The language skill set and **multilingual profile** required of staff…	... وعلى مجموعة المهارات اللغوية و**المظهر المتعدد** **اللغات** المطلوب من الموظفين...	... ومجموعة المهارات اللغوية و**النمط المتعدد** **اللغات** المطلوب من الموظفين...	... ومجموعة المهارات اللغوية و**الملامح المتعددة** **اللغات** المطلوبة من الموظفين...	... ومجموعة المهارات اللغوية **ومؤهلات التعددية** **اللغوية** المطلوبة من الموظفين...
language immersion courses.	**دورات لغمر اللغة**	**دورات تدريبية لغوية**	**دورات انغماس لغوي**	**دورات انغماس لغوي**
… the Secretary-General **sets the tone at the top** for multilingualism…	... **يحدد** الأمين العام **نبرة** التعدد اللغوي في **أعلى المستويات**...	... **يضع** الأمين العام **نغمة** التعددية اللغوية **في القمة**...	... **يحدد** الأمين العام **النغمة العليا** لتعدد اللغات...	... **يولي** الأمين العام **أهمية كبيرة** **لممارسة** التعددية اللغوية...

The first error is apparent in the translation of "multilingual profile." The word "profile" was translated as "مظهر" (appearance), "نمط" (pattern), and "ملامح" (features or traits), respectively, in the three outputs. These translations are lexically correct but fail to convey the intended meaning in this context. The phrase refers to the knowledge, abilities, and skills associated with multilingualism and language proficiency. In Reverso's translation of "language immersion courses," the phrase was rendered as "دورات لغمر اللغة," repeating the same error observed in Google Translate's earlier handling of the term "immersive," which was translated as "غامرة." Systran, on the other hand, generated an inaccurate translation, "تدريبية," while Microsoft Azure provided an accurate and contextually appropriate translation ", دورات الانغماس اللغوي." All three engines provided different yet inaccurate and literal translations for the phrase "sets the tone at the top." In this context, the phrase refers to the UN Secretary-General demonstrating the importance of multilingualism by using various languages in his speeches and ensuring his messages are available in all official languages. Consequently, the better translation could be "يولي أهمية كبيرة لممارسة," which is more contextually appropriate. The three MTs also contained various accuracy errors that did not fall into the said categories. Some minor or neutral errors did not significantly affect the overall meaning.

According to TAUS (2021), Fluency errors are Issues linked to the form or content of the text. In terms of punctuation, all three MT engines retained the original English text's punctuation in their Arabic translations despite differences between the two languages in punctuation conventions. However, some errors occurred, particularly in Reverso's output, including an unnecessarily added period between two words in one case and omitted after numbering a heading in another. Spelling errors were also observed in Systran and Microsoft Azure translations. Systran misrendered "احتياجات" and "اتجاه" as "إحتياجات" and "إتجاه," introducing an unnecessary Hamza (the glottal stop). Microsoft Azure translated "تماشيا" as "تمشيا," omitting the elongation. Reverso, however, was notably free from spelling errors in this instance.

The majority of fluency errors were grammatical mistakes. A famous example of this category was noticed across all three translated texts: the unnecessary addition of the definite article "أل" at the beginning of certain words. For instance:

- "multilingual and multicultural contexts" → "السياقات المتعددة اللغات والثقافات"
- Suggested correction: "السياقات متعددة اللغات والثقافات"

Even though these errors might appear small, they have a noticeable impact on how smooth and natural the translated text feels. They showcase the difficulties MT engines face when processing Arabic grammar and their reliance on low-quality data found across the internet.

The reviewed output revealed consistent fluency issues stemming from literal translation. For example, in the phrase "the World Health Assembly requested the Director-General of WHO to…," all three engines used "طلبت إلى," which conveys a sense of plea rather than the professional and authoritative tone intended by "requested." A more accurate translation would be "طلبت من." Similarly, the sentence "… yet only English is a requirement when it comes to…" resulted in an awkward translation:

- "ولكن اللغة الإنجليزية فقط هي شرط عندما…"

The literal approach rendered the sentence awkward, and a smoother phrasing, such as " لكن الإنجليزية تعد شرطاً فقط" would better convey the intended meaning. These two examples underscore the tendency of machine translation engines to opt for literal translation, which often compromises fluency and coherence in the target language.

Terminology errors arise when a term or phrase is translated with a meaning unrelated or inconsistent with the original text's context, often resulting from the terms being inconsistent with the specified term bank or the arbitrary selection of target language equivalents (TAUS, 2021). One apparent terminology error in Reverso's output is the translation of "staffing" as "مُلاك الموظَّفين," which fails to capture the intended meaning of "staffing," a term that refers to the process of recruiting and selecting personnel based on their skills and expertise. A more accurate translation would be "التوظيف," a term correctly used by both Systran and MS Azure. Another problem is the inconsistency in how Reverso and MS Azure translate the word "classes." In the same translation, it was sometimes rendered as "دُروس" and other times as "فُصول," which introduces confusion. To avoid this, the term should be translated uniformly, adhering to a single option, such as "دُروس" or "حِصَص", throughout the entire text.

Style errors pertain to coherence, appropriateness, or adherence to specific stylistic guidelines. The analysis of translations generated by the three MT engines revealed several recurring stylistic challenges, many of which originate from literal translations of expressions, resulting in awkward and non-idiomatic styles in Arabic. A famous error observed across all engines was the overuse of the English passive constructions, influenced by the literal translation of auxiliary verbs like "is/are." For example:

- "All major global documents are translated into these three languages…"
- → "تتم ترجمة جميع الوثائق العالمية الرئيسية إلى هذه اللغات الثلاث"

While the translation conveys the basic meaning, it introduces the verb "تم" unnecessarily, which is less idiomatic in Arabic. Replacing it directly with the more idiomatic Arabic passive structure (e.g., "تُتَرجم") would result in a more fluent and natural translation. The same applies to "is confirmed" and "are assigned," which can be translated as "تُأَكّدُ" and "عُيّنُوا," respectively. Another common error is the literal translation of the verb "consider" as "تُعتبر", which, while grammatically correct, is less appropriate than "تُعد" or a context-specific alternative. For example:

- "The knowledge of a second working language is considered desirable."
- → "ولا تعتبر معرفة لغة عمل ثانية شرطًا"

Several examples demonstrated a lack of Arabic idiomaticity. For instance, "To draw attention to" was translated by all engines as "وجهت الانتباه", which, though correct, lacks the idiomatic quality of " لفتت الانتباه" or "استرعت الانتباه". In addition, the translation of the verb "nurture" illustrates how context can be overlooked:

- "…retain and nurture staff with the skill sets required."
- Reverso: "تغذيتهم بمجموعات المهارات …"
- Systran: "تزويدهم بمجموعات المهارات …"
- Microsoft Azure: "تنشئتهم بمجموعات المهارات …"

While “تزويدهم” (Systran) aligns best with the intended meaning of providing and enhancing skills, “تغذيتهم” (Reverso) and “تنشئتهم” (Azure) fail to convey the professional and developmental context, as they imply unrelated concepts of nourishment or upbringing. Contextually appropriate terms like “تعزيز” or “تطوير” would better reflect the intended meaning.

Locale convention errors involve inadequacies in adapting content to region-specific standards, such as date formats, currencies, addresses, phone numbers, abbreviations, and measurement systems (TAUS, 2021). While these errors were rare, one noteworthy example was the retention of Roman numerals (XXIII) in both Reverso and Systran translations, which were not translated into the Arabic numerical system (23).

3.2 Text 2

Table 3: MT errors in the second text (Arabic-English)

Error type	Reverso (997 w)	Systran (994 w)	MS Azure (990 w)
Accuracy (total)	**9**	**6**	**6**
Addition	2	1	1
Mistranslation	5	5	5
Untranslated text	2	-	-
Fluency (total)	**-**	**3**	**-**
Punctuation	-	3	-
Terminology (total)	**20**	**23**	**22**
Inconsistent use of terminology	5	5	5
Style (total)	**1**	**1**	**1**
Awkward	1	1	1
Design (total)	**-**	**1**	**-**
Total errors	**32**	**34**	**29**
Severity of errors	Neutral: 2 Minor: 16 Major: 14	Neutral: 2 Minor: 18 Major: 14	Neutral: 1 Minor: 13 Major: 15

The translations generated by the three engines contained various accuracy-related issues, including mistranslations, unnecessary additions, and untranslated text. A common error was the mistranslation of the phrase “إن وُجدت” (if any):

- “وعدد العمالة الأجنبية والوطنية المراد استقدامها لكل مرحلة من مراحل المشروع [إن وجدت].”
- → “... the number of foreign and national workers to be recruited for each stage of the project, [if any].” (Reverso, Systran, Microsoft Azure)

“If any” is commonly used in legal, commercial, or academic documents. It serves to indicate the potential absence or minimal presence of something. However, in this context, the more accurate translation of “إن وُجدت” would be “if applicable”, as it refers to the relevance of the documents to the project rather than their existence. Another error was the addition of the pronouns “he” or “she” in Reverso and Systran's outputs, despite their absence in the source text. This addition was a neutral error and did not significantly alter the meaning of the text. Reverso also failed to translate two sentences, resulting in critical errors that could introduce ambiguity into the target text.

Other examples of accuracy errors include the mistranslation of “الشركة المستخدمة” (the employing company) as “the used company” by Reverso, which is literal and contextually inappropriate. Similarly, Systran translated “استمارة تعريف الشركة” (company identification card) as “the company definition form”, conveying a different meaning. The same goes for MS Azure, which produced several mistranslations, such as translating “إجراءات منح تصريح العمل” (procedures for obtaining a work permit) as “Get the

Convention from where principle.” Another example is the redundant and incorrect translation of “الحصول على تأشيرة العمل” (obtaining a work visa) as “Obtaining Obtaining Visa Work.” Not all errors were consistent; some phrases were accurately translated in one instance but mistranslated in another. However, these errors suggest that the systems mostly opt for literal translation. Reverso and Microsoft Azure generated translations without any noticeable fluency errors, unlike Systran, which produced minor issues related to punctuation. The latter added unnecessary quotation marks (“”) in several places within the translated text. Although they do not significantly alter the overall meaning, they could mislead readers into interpreting the marked text as direct quotations or attributed statements.

The majority of errors in this study fell under the category of terminology. Given that the source text includes instructions and regulations related to employment and migration, most of the terms are legal in nature. Table (4) presents examples of how key legal terms were translated by the three MT systems, along with suggested equivalents:

Table 4: English translations of key legal terms in the Target Texts

Source Text	**Reverso**	**Systran**	**Microsoft Azure**	**Suggested Equivalent**
نظام	Regulation / Regime / System	Regulation / Regime / System	Regulation / Regime / System	**Regime**
الولاية المختصة إقليميا	the regionally competent state	the regionally competent state	the regionally competent state	**the regionally competent Wilaya**
انتداب أو مهمة قصيرة الأجل	a short-term assignment or assignment	a short-term assignment or assignment	a short-term assignment or assignment	**secondment or short-term missions**
العمل بأجر	Wage Activity	Paid Activity	Paid Activity	**Paid Activity**
السجل التجاري	Commercial Register	Trade Record	Commercial Registration	**Trade Register**

The term “نظام” can be translated as “regulation,” “regime,” or “system,” depending on the context. Considering this, “regime” is the most fitting term in the context of labour and migration laws. It describes a structured framework of rules or conditions. Words like “regulation” and “system” do not convey the same level of precision required in this legal setting. The term “الولاية المختصة إقليميا” refers to an administrative division in Algeria, similar to other administrative divisions worldwide, such as states, provinces, or governorates. Using “state” to translate “الولاية” can be confusing. It may suggest a fully sovereign state rather than an administrative division. A better option, like “the regionally competent Wilaya,” captures the term's administrative meaning, which is tied to specific areas in North Africa and the Middle East. Moreover, the phrase “انتداب أو مهمة قصيرة الأجل” refers to the temporary assignment of employees for short-term tasks. The MT systems' translations are repetitive and fail to distinguish between the two concepts. The suggested translation, “secondment or short-term missions”, reflects the intended meanings and is commonly used in legal and commercial contexts in Algeria.

Concerning the remaining terms, “العمل بأجر” refers to any work for which compensation is provided. “Paid Activity” seems to be the most appropriate translation in this context, while Reverso's “Wage Activity” is less common. “شهادة الإحصاء” is statistical records of individuals or businesses within a specific region or entity. Finally, “السجل التجاري” is an official record of a company's business activities. Both the “Trade Register” and “Commercial Register” are accurate translations, but the MT systems' outputs are less standardized. These examples demonstrate that MT engines continue to struggle with translating domain-specific terminology, resulting in ambiguity and misunderstanding.

Due to the reliance on literal translation, the style error category consisted of unnatural and out-of-place segments. Some phrases and sentences made the output appear to be a word-for-word translation from Arabic. Furthermore, only one design-related error occurred in Systran's translation. Systran's translation contained formatting and layout inconsistencies. This included unnecessary punctuation marks and difficulty reading some sentences with missing spacing.

3.2 Text 3

Table 5: MT errors in the third text (English-Arabic)

Error type	Reverso **(876 w)**	Systran **(880 w)**	MS Azure **(893 w)**
Accuracy (total)	**15**	**11**	**14**
Addition	-	1	-
Under-translation	1	1	1
Mistranslation	5	7	3
Untranslated text	9	2	10
Fluency (total)	**4**	**12**	**4**
Punctuation	-	2	1
Spelling	-	1	-
Grammar	-	3	2
Inconsistency	-	-	1
Terminology (total)	**5**	**6**	**4**
Inconsistent use of terminology	-	2	1
Style (total)	**7**	**9**	**5**
Awkward	3	5	2
Unidiomatic	4	4	3
Total errors	**31**	**38**	**27**
Severity of errors	Neutral: 1 Minor: 8 Major: 22	Neutral: - Minor: 12 Major: 26	Neutral: - Minor: 11 Major: 16

The three MT systems produced comparable accuracy-related errors, including additions, omissions, mistranslations, and untranslated segments. Only one error occurred in the addition category during Systran's translation. Systran added unnecessary elements, such as the dates and an explanatory note about a company's name in English, which were not mentioned in the source text. On the other hand, under-translations were present in all three translations. To exemplify:

- "... our catalogue which already contains more than 9000 products and counting."
- → Reverso: الكتالوج الذي يحتوي بالفعل على أكثر من 9000 منتج والعد
- → Systran: الكتالوج لدينا والذي يحتوي بالفعل على أكثر من 9000 منتج والعدد
- → MS Azure: كتالوجنا الذي يحتوي بالفعل على أكثر من 9000 منتج والعد

All three systems failed to translate the phrase "and counting" fully, which could have been rendered as "والعدُّ مُستمر" (and counting) or "والعدد في تزايد" (and the number is increasing). This omission weakens the statement's dynamic. In addition, several brand names and technical terms were left untranslated, often appearing in their original Latin script. The same goes for mistranslation errors. Consider the following example in Reverso, Systran, and MS Azure respectively:

- "Ziploc brand Sandwich Bags are the unbeatable bags to grab on the go now with Grip n Seal technology."
- (العلامة التجارية Ziploc ساندويتش أكياس هي أكياس لا تقبل المنافسة للاستيلاء على الذهاب الآن مع Grip n Seal التكنولوجيا)

- (ساندوتش براند زيبلوك هي الحقائب التي لا يمكن ضربها في التنقل الآن مع تكنولوجيا Grip n Seal)
- (أكياس الساندويتش من ماركة زيبلوك هي الحقائب التي لا تقبل المنافسة للاستيلاء عليها أثناء التنقل الآن باستخدام تقنية Grip n Seal)

These translations are awkward and, at times, nonsensical due to the literal rendering of phrases like "to grab on the go" and the improper handling of brand names and other words.

Fluency errors primarily stemmed from reliance on a literal translation. These errors disrupted the natural flow of the text and, in some cases, made the translations sound awkward or unnatural. Reverso's output was particularly prone to literal translations. For example:

- "That's why we're continuously extending our help to the communities and finding ways to reduce our environmental footprints to achieve greener earth."
- لهذا السبب نحن نوسع باستمرار مساعداتنا للمجتمعات المحلية ونجد طرقا للحد من بصماتنا البيئية لتحقيق أرض أكثر خضرة).

This translation is literal and lacks fluency, particularly in the latter part of the sentence. The phrase "extending our help" should be translated as "تقديم المساعدة" (offering help) or "مد يد العون" (extending a helping hand), rather than the literal "نوسع مساعداتنا". Similarly, "reduce our environmental footprints" should be translated as "الحد من الآثار البيئية" (reducing environmental impacts) instead of the awkward "الحد من بصماتنا البيئية", which implies leaving a positive impact. At the same time, Systran's translation contained numerous grammatical errors, including the omission of definite articles (أل) in several words and incorrect verb conjugations, while incorrect pronoun usage and punctuation errors carried over from the source text in MS Azure's output.

The MT systems produced a limited number of terminology errors. Table 6) illustrates some of these errors, along with suggested corrections:

Table 6: Terminology Translations

Source Text	Reverso	Systran	MS Azure	Suggestion
Parchment Paper	ورق الرق	ورق الرقائق	ورق الزبدة	**ورق الزبدة**
Couriers	سعاة	سعاة	شركات النقل	**شركات التوصيل**
Facial Tissue	أنسجة الوجه	نسيج الوجه	مناديل الوجه	**منادلي الوجه**
Clingfilm	الكلينغفيلم	فيلم كلينغهام	فيلم التشبث	**ورق التغليف**

The translation of "Parchment Paper" deferred across the systems, with Reverso and Systran providing inaccurate translations. At the same time, MS Azure correctly rendered it as "ورق الزبدة", accurately reflecting its use in baking and cooking. For "Couriers," Reverso and Systran translated it as "سعاة", which is less precise. MS Azure's "شركات النقل" (transport companies) was closer but still less accurate than "شركات التوصيل" (delivery companies). The term "Facial Tissue" was mistranslated by Reverso and Systran as "أنسجة الوجه" (biological facial tissues) and "نسيج الوجه" (facial fabric), respectively, both of which are inaccurate. MS Azure correctly translated it as "مناديل الوجه", which refers to disposable paper tissues. Finally, "Clingfilm" was poorly translated across all systems, with Reverso transliterating it as "الكلينغفيلم", while Systran and Microsoft Azure provided inadequate translations (" فيلم كلينغهام" and "فيلم التشبث", respectively). The correct translation is "ورق التغليف / بلاستيك" (plastic wrap), used for food preservation.

Since it is a marketing material, the MT engine's outputs contained several stylistic errors. This issue is caused by literal translation and the lack of reliance on non-idiomatic expressions or phrases. Three systems once again incorrectly used the verb "تمَّ" instead of the passive voice, resulting in awkward constructions like "يتم فحص" instead of the more natural "يُفحَصُ". Another example of stylistic awkwardness is the translation of the phrase "Falcon Pack is a global player." Systran and MS Azure rendered it literally as "لاعب عالمي" (global player), which is correct yet out of context. Reverso, however, provided a more idiomatic translation: "شركة رائدة عالميا" (a globally leading company), which is an adequate and stylistically appropriate expression. Similarly, the phrase “Fine Filipino cooking” was translated differently by each system:

- Reverso: “الطبخ الفلبيني الجيد” (Good Filipino cooking).
- Systran: “الطهي الفلبيني الممتاز” (Excellent Filipino cooking).
- MS Azure: “الطبخ الفيليبيني الفاخر” (Luxurious Filipino cooking).

Reverso's translation is literal and basic. Systran and Microsoft Azure, on the other hand, provided more elaborate translations. Another appropriate rendering in this context would be “الطبخ الفلبيني الراقي” as these terms better capture the sophistication implied by the original phrase. Stylistic errors reveal the limitations of MT systems in producing natural and idiomatic translations in contexts that require persuasive or culturally sensitive language. As Ayadi (2017) notes, translating advertising slogans requires specific strategies to preserve their persuasive function, a task that machine translation alone cannot achieve. As a result, human involvement through post-editing can make a significant difference by enhancing quality and maintaining clarity.

4. Discussion

The results revealed several differences and similarities in the performance of the three MT engines and the quality of translations based on the (TAUS) error typology. Accuracy, fluency, terminology, and style were the most frequent error types. This method provided a clear framework for evaluating the quality of MT, particularly in contexts where reference translations are unavailable. MT engines generally perform moderately and deliver understandable translations to varying extents when dealing with news texts and reports, given their direct style and formal, unambiguous nature. In a study by Ismailia (2023), the machine translations of a news report were generally comprehensible. However, they contained several errors in verb tenses, passive voice, auxiliary verbs, and definite articles when translating from English to Indonesian. In the current study, certain linguistic and stylistic structures and expressions specific to UN news reports posed challenges. While Reverso and MS Azure handled the language of UN institutions more effectively, Systran struggled with specific phrases and context-specific terminology, resulting in errors in several instances.

NMT systems face persistent issues when translating domain-specific expressions due to their highly specialized vocabulary and insufficient contextual information (Arčan & Buitelaar, 2017). Such limitations are further complicated when dealing with domain-specific terminology, which requires a deep understanding of specialized knowledge and contextual awareness, an issue that modern MT systems still struggle with (Naveen & Trojovský, 2024). General-purpose MT systems tend to underperform due to their inability to handle technical terms or culture-specific expressions. Unlike the other two systems, Reverso performed better in accurately translating legal terms, likely due to its extensive database of legal corpora. Along the same lines, Alkatheery (2023) notes that MT faces several challenges when rendering specialized legal structures and terminology from Arabic to English. Thus, the development of customized translation systems trained on domain-specific datasets is essential to improve accuracy and reliability in these contexts (Chu & Wang, 2020).

Marketing texts require precise translation and cultural adaptation to maintain the message's intent and persuasive elements. MT engines often fail to capture the nuances and idiomatic expressions essential to marketing content, struggling to preserve stylistic and rhetorical components. Musaad and Al Towity (2023) pointed out that translating idiomatic expressions using MT engines between Arabic and English poses overarching problems for developers, as these systems often produce literal translations. The engines frequently delivered stiff, literal translations, lacking stylistic fit. Better training data, including idioms and stylistic conventions, is essential for natural, fluent results. In this study, all three MT engines leaned toward literal translations of idiomatic expressions, producing target texts that felt awkward and failed to meet the standards expected of marketing texts.

Accuracy errors were prevalent across all three engines. Reverso committed the highest number of errors in the category of incorrect translations, while Systran's translations also contained unnecessary additions and inaccuracies. MS Azure performed relatively well in terms of accuracy, but untranslated segments were frequently found in marketing texts. The engines succeeded in handling pragmatic texts well but faced issues with complex structures. Other common issues included misuse of definite articles and literal translations. In his study, Ali (2020) found that Google, Bing, and Ginger made significant errors in accuracy and fluency when translating a UN report from English to Arabic. Thus, these issues emphasize the need for post-editing as an indispensable step in refining MT outputs.

Term translations varied, but the number of errors in the terminology category revealed the engines' limitations when dealing with specialized texts. It can be argued that terminology specific to a particular legal system, such as Algerian legal terms in the second text, further complicates the task for these engines. Correspondingly, Vigier-Moreno and Macías (2022) found that MT engines struggle with specialized legal terminology and terms specific to certain legal systems when translating Spanish to English. Specialized MT engines always come in handy when dealing with legal texts, and opting for them might be a practical solution. According to Koponen and Salmi (2017), a detailed analysis of post-editing outputs can reveal the nature of errors that MT systems generate. Therefore, post-editing is essential for addressing accuracy, fluency, terminology, and style issues. The engines can also be enhanced by incorporating high-quality Arabic training resources. The objective is to ensure that the final output aligns with the target language's style and nuances.

Conclusion

Using the TAUS error classification method, this study evaluated the performance and quality of three MT engines in translating three types of texts between Arabic and English: a UN news report, a legal-administrative text on labour and migration, and a set of marketing paragraphs. The researcher assessed the quality of the translations and compared the engines' performance to explore the extent to which human translators can benefit from MT outputs. Most of the obtained results align with studies conducted on several MT systems (Ali, 2020; Vigier-Moreno & Macías, 2022; Ismailia, 2023).

MT engines perform reasonably well with pragmatic and explicitly worded texts but struggle with idiomatic expressions, culture-specific elements (e.g., marketing content), complex sentence structures (e.g., UN reports), and specialized legal terminology (e.g., labour and migration texts). This is evident in Arabic-to-English translations, where the linguistic differences between the two languages and the reliance on literal translation often lead to awkward and error-prone outputs. A notable factor contributing to these challenges is that MT systems are trained on contemporary Arabic texts from the internet, which are often filled with common errors and literal translations from foreign languages. MT fails to accurately translate cultural and idiomatic expressions from English into Arabic, producing literal translations that do not capture the original meaning (Alawi & Abdulhaq, 2021). Therefore, it is recommended that MT engines be trained on high-quality, error-free Arabic sources, such as older, well-written resources, which feature proper grammar and adequate idiomatic expressions. Bogush et al. (2019) argue that proper human intervention in post-editing is essential for effectively utilizing MT systems, clarifying that human translators must conduct linguistic analysis, verify terminology, and complete the final drafting. Furthermore, translators should be trained to work effectively with MT systems and critically evaluate their outputs to maximize their potential, especially during the post-editing phase. Good NMT engines can help skilled translators tremendously as post-editing becomes less daunting. There is a need to include MT post-editing in English-Arabic academic programs to capitalize on this matter. Research on integrating machine translation post-editing (MTPE) into translator training and education programs is limited in the Arab world, emphasizing the urgent need for academic institutions to implement responsive and forward-thinking policies and practices (Omar & Salih, 2024).

References

Alawi, N., & Abdulhaq, S. (2021). Machine translation: The cultural and idiomatic challenge. *Journal of Al-Azhar University*, *19*(2), 207–234.

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*(5). https://doi.org/10.4236/ojml.2020.105030

Alkatheery, E. R. (2023). Google Translate errors in legal texts: Machine translation quality assessment. *Arab World English Journal for Translation and Literary Studies*, *7*(1), 208–219. https://doi.org/10.24093/awejtls/vol7no1.16

Alkhatib, M., & Shaalan, K. (2018). The key challenges for Arabic machine translation. In K. Shaalan, A. E. Hassanien, & F. Tolba (Eds.), *Intelligent natural language processing: Trends and applications* (Vol. 740, pp. 139–156). Springer. https://doi.org/10.1007/978-3-319-67056-0_8

Almaaytah, S. A., & Alzobidy, S. A. (2023). Challenges in rendering Arabic text to English using machine translation: A systematic literature review. *IEEE Access*, *11*, 94772–94779. https://doi.org/10.1109/ACCESS.2023.3309642

Almahasees, Z. (2022). *Analysing English-Arabic machine translation: Google Translate, Microsoft Translator and Sakhr*. Routledge. https://doi.org/10.4324/9781003191018

Arčan, M., & Buitelaar, P. (2017). Translating domain-specific expressions in knowledge bases with neural machine translation. https://doi.org/10.48550/arXiv.1709.02184

Ayadi, A. (2017). Investigating the strategies used to translate English advertisement slogans into Arabic. *Revue Des Sciences Humaines*, *28*(3), 05–16.

Bogush, A. M., Korolova, T. M., & Popova, O. V. (2019). Teaching machine translation to the students majoring in the humanities. *Information Technologies and Learning Tools, 71*(3). https://doi.org/10.33407/itlt.v71i3.2724

Chu, C., & Wang, R. (2020). A survey of domain adaptation for machine translation. *Journal of Information Processing*, *28*, 413–426. https://doi.org/10.2197/ipsjjip.28.413

Görög, A. (2014). Quantifying and benchmarking quality: The TAUS Dynamic Quality Framework. *Tradumàtica Tecnologies de La Traducció*, *12*, 443–454. https://doi.org/10.5565/rev/tradumatica.66

Hadla, L. S., Hailat, T. M., & Al-Kabi, M. N. (2014). Evaluating Arabic to English machine translation. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *5*(11). https://doi.org/10.14569/IJACSA.2014.051112

Ismailia, T. (2023). Analysis of machine translation performance on translating informative text from English into Indonesian. *EBONY: Journal of English Language Teaching, Linguistics, and Literature*, *3*(2), 129–138. https://doi.org/10.37304/ebony.v3i2.9809

Koponen, M., & Salmi, L. (2017). Post-editing quality: Analysing the correctness and necessity of post-editor corrections. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *16*. https://doi.org/10.52034/lanstts.v16i0.439

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4984–4997. https://doi.org/10.18653/v1/2020.acl-main.448

Musaad, M. M. A. M., & Al Towity, A. A. (2023). Translation evaluation of three machine translation systems, with special references to idiomatic expressions. *Humanities and Educational Sciences Journal*, *29*, 678–708. https://doi.org/10.55074/hesj.vi29.700

Nagi, K. A. (2023). Arabic and English relative clauses and machine translation challenges. *Journal of Social Studies*, *29*(3). https://doi.org/10.20428/jss.v29i3.2180

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, *27*(10), 110878. https://doi.org/10.1016/j.isci.2024.110878

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

Povilaitienė, M., & Kasperė, R. (2022). Machine translation for post-editing practices. *Scientific Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development*, *24*. https://doi.org/10.31392/NPU-nc.series9.2022.24.04

Sajjad, H., Abdelali, A., Durrani, N., & Dalvi, F. (2020). AraBench: Benchmarking dialectal Arabic-English machine translation. *Proceedings of the 28th International Conference on Computational Linguistics*, 5094–5107. https://doi.org/10.18653/v1/2020.coling-main.447

TAUS. (2021). *TAUS Dynamic Quality Evaluation*. TAUS – The Language Data Network. https://info.taus.net/dqf-mqf-error-typology-template-download

Valli, P. (2015). The TAUS Quality Dashboard. *Proceedings of Translating and the Computer 37*, 127–136. https://aclanthology.org/2015.tc-1.17

Vigier-Moreno, F., & Macías, L. (2022). Assessing neural machine translation of court documents: A case study on the translation of a Spanish remand order into English. *Revista de Llengua i Dret*, *78*, 73–91. https://doi.org/10.2436/rld.i78.2022.3691

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic machine translation: A survey with challenges and future directions. *IEEE Access*, *9*, 161445–161468. https://doi.org/10.1109/ACCESS.2021.3132488