Aya Sayed Omran Elsayed*

When machines meet gavel: a case study of the English–Arabic machine translation of the Egyptian arguments before the International Court of Justice (2024)

https://doi.org/10.1515/lass-2025-0054
Received May 14, 2025; accepted May 29, 2025

Abstract: The legal field heavily relies on audio–visual content such as witness testimonies and trials, making accurate transcription and translation crucial, especially in cross-border cases. This study examines the performance of neural machine translation (NMT) in handling such material, using the DQF-MQM harmonized error typology to categorize errors by type, including terminology, accuracy, and fluency. Legal translation demands precision, as minor errors can impact legal outcomes. Thus, this analysis focuses on English-to-Arabic translations of Egyptian oral arguments before the International Court of Justice, sourced from DawnNews (Feb 21, 2024). It investigates whether errors stem from the ASR-generated transcript or the Google NMT system. The findings aim to guide machine translation post-editors (MTPEs) in identifying lexical and syntactic patterns that typically result in errors, ultimately supporting more accurate and legally sound translations.

Keywords: audio-visual translation (AVT); English to Arabic translation; legal discourse; International Court of Justice (ICJ); neural machine translation (NMT); the DQF-MQM harmonized error typology

1 Introduction

Over the past two decades, translation studies (TS) have undergone a significant transformation, often referred to as the "technological turn" (Díaz-Cintas 2013). This shift has empowered translators, interpreters, and linguists with tools that improve both speed and efficiency, opening new career paths that require human creativity and critical thinking. Tasks that were once labor-intensive, such as transcribing audiobooks or legal proceedings, can now be completed in moments using advanced technologies.

***Corresponding author: Aya Sayed Omran Elsayed**, Ain Shams University, Cairo, Egypt, E-mail: Aya.omran.97@alsun.asu.edu.eg. https://orcid.org/0009-0000-7347-2007

In the legal field, traditionally cautious and rooted in procedural rigor, technological integration is gradually taking hold. Two key innovations in this transformation are automatic speech recognition (ASR) and machine translation (MT). These tools have the potential to streamline legal workflows, enhance accessibility, and support cross-border legal communication. MT can facilitate the translation of essential legal documents, such as contracts and court opinions, making them accessible to clients with limited proficiency in English and aiding international legal processes.

However, despite its advantages, MT in the legal domain presents challenges. Legal language is highly nuanced and relies heavily on precision, and inaccuracies – particularly involving legal terminology or idiomatic expressions – can lead to serious consequences. While MT can assist in increasing efficiency, human oversight remains essential to ensure the reliability of translations. Subtle errors in translated legal texts may affect interpretation, precedence, or even judicial outcomes.

To use MT effectively in legal contexts, it must first be thoroughly and systematically evaluated. Assessing MT quality is a complex process, as errors may vary in nature and severity. Several typologies have been proposed to classify such errors – ranging from factual inaccuracies and clarity issues to linguistic and fluency errors. One of the most comprehensive and widely accepted frameworks for MT quality assessment is the DQF-MQM harmonized error typology, developed by the Translation Automation User Society (TAUS) and the German Research Center for Artificial Intelligence (DFKI).

This study employs the DQF-MQM framework to evaluate the Arabic translation of an English ASR-generated transcript. The transcript in question features Egyptian oral arguments presented before the International Court of Justice (ICJ) during its third session, as published on the DawnNews English YouTube channel on February 21, 2024.

2 Review of literature

With the soaring demand for subtitling all types of audiovisual content, the subtitling industry has experienced a gigantic leap in terms of quantity and production quality. However, this momentum could not keep pace with the rapid increase in the number of videos uploaded on the internet, every single day. For instance, on YouTube, a web portal located in around 25 countries across 43 languages, up to 20 h of video are uploaded every single minute, which results in nearly 8 years of content uploaded every day in hundreds of different languages. More than 3 billion of these videos are viewed daily, as per a study conducted by Diaz-Cintas, back in 2013. This undoubtedly indicates that these figures have dramatically increased, along with the significance of audiovisual material on the internet in bridging the gap between cultures, communicating, learning new languages, and reaching trustworthy sources of news and information, especially related to worldwide controversial issues and major humanitarian conflicts.

Stemming from this outstanding necessity to rely on tools instantly providing accurate and fluent subtitles, it has become fundamental to gauge the performance of the commonly used machine translation (MT) systems with the aim of improving their quality after investigating the root causes and patterns of translation errors and pinpointing the weaknesses of such systems. Despite the significance of this goal, there is still a great lack of consensus and standardization in the field of translation quality assessment, due to the complicated cognitive, linguistic, social, cultural, in addition to technical process this supposes (Popović 2018). When translation quality assessment (TAQ) is mentioned, it holistically refers to two approaches: the first is a macro-analytical approach, wherein questions of ideology, function, gender or even register are considered, while the latter is a micro-analytical one where the value of collocations and individual linguistic units are examined (Castilho et al. 2018). Various typologies and models have been established to distinguish translation errors at different levels, such as spelling, vocabulary, grammar, and discourse; some of them have further divided errors into subcategories that include, for instance, errors relating to concordance, style, confusion in word meaning with various exceptions, etc.

During the last decade, a myriad of projects has emerged aiming at standardizing these tools to better facilitate the adaptation of different tasks and language pairs and to reduce both effort and inconsistencies when developing an error typology. Examples of such projects include the multidimensional quality metrics (MQM) framework, created by the QTLaunchPad and dynamic quality framework (DQF) project, and established by the Translation Automation User Society (TAUS); this project started independently then was integrated in the “DQF/MQM Error Typology” in 2014 (Popović 2018).

Scrutinizing the significance of employing machine translation and investigating its potential in the legal domain, it can be seen that assessing MT systems presents a unique and critical demand for numerous reasons. First, since errors in legal translations can have considerable consequences, potentially affecting legal rights as well as judgements, examining MT systems aids in identifying areas for improvement to guarantee they produce reliable translations (Mackenzie 2019). Significance of MT evaluation also increases due to the complexity of the legal language that is full of specific terminology, double meanings, and subtle nuances, in addition to the high demand for almost all kinds of legal translations (Przybyszewski 2017).

The uniqueness of the legal language has been highlighted by Jackson (1985, p. 47), when he stated that the differences between the legal language and the ordinary language reside in the lexicon and structure of both jargons: “[the legal language] having a lexicon constituted in a manner different from that of the ordinary language, and involving terms related to each other in ways different from those of the ordinary language, must be autonomous of the ordinary language”. Jackson has also reiterated that it is difficult to comprehensively understand the entire meaning of legal texts, due to “lack of knowledge of the system, rather than lack of knowledge of individual lexical items” (1985, p. 48).

Secondly, according to Przybyszewski (2017), given that traditional human translation is time-consuming and costly, determining the potentials of MT systems to streamline translation processes potentially and identifying their limitations in handling legal language saves time and resources and helps in assessing their cost-effectiveness compared to traditional methods. When trained on vast amounts of legal data, MT systems can easily produce accurate translations and further reduce ambiguity in all sorts of legal communication (Carl 2018). Nonetheless, Mackenzie (2019) reiterates in his study that human oversight remains essential, as legal professionals must make informed decisions about when and how to depend on MT; even the best MT tool can barely produce flawless translation.

Through examining MT systems in the legal domain, their potential could be leveraged to improve efficiency and accessibility of legal services and ultimately pave the way for responsible integration of MT technology into the legal landscape. Yet, it is still pivotal to recognize limitations and post-edit MT outputs to guarantee the achievement of accurate, sound, and reliable legal translations.

Even though legal languages are universally characterized by a unique and complex legal lexicon, every single language has its own distinctive vocabulary due to the legal system which it represents. For example, legal English terminology is different from the legal Arabic one, since the first is demonstrative of the Common Law system, while the second is representative of the Islamic Law tradition. Additionally, the English legal lexicon is full of archaic words, common words with uncommon meanings and words of over-precision, among others (Tiersma 1999). Latinism is also a common feature of the English legal discourse; it refers to any word or expression that is borrowed from Latin, such as *bona fide*, “made in good faith”, and *res nova*, “a case or issue that has never before been decided by a court”, as per *Merriam Webster*.

Due to the sensitivity and complexity of legal proceedings, mistranslations in such a field have profound consequences, impacting the fairness and outcome of trials. In the judicial system, both accuracy and precision in language is crucial, since a misinterpretation of a legal document, witness testimony, or any sort of communication between legal parties can undoubtedly lead to violations of rights, or even wrongful convictions. Hence, when dealing with different language combinations, such as the English/Arabic combination, errors in translation can amplify such risks. Some high-profile legal cases have encountered mistranslations, which have caused significant issues.

Two prominent cases highlight the critical role of translation in international law. The first involves UN Resolution 242, where a linguistic discrepancy between the English and French versions led to divergent interpretations with significant political implications. The English version called for the withdrawal of Israeli forces from “territories occupied,” omitting the definite article “the,” which allowed Israel to argue for a partial withdrawal. In contrast, the French version included “des territoires occupés,” implying a complete withdrawal. This difference fueled ongoing debates about the resolution’s intent and legality. Although Arabic was not an official UN language at the time, many Arabic-speaking nations interpreted the resolution in line with the French version, reinforcing demands for full Israeli withdrawal and contributing to long-standing territorial disputes (Fathi n.d.).

The second case underscores the legal and financial consequences of a mistranslation in a contract involving the Egyptian government. A misinterpretation of the Arabic word “تسري” as “يسري” led to a fundamentally different understanding of contractual obligations, resulting in a multi-million-dollar arbitration ruling against Egypt. Further complications arose when the Egyptian Minister of Tourism signed the contract with the phrase “approved, agreed and ratified,” instead of simply “approved.” These additional words were interpreted as binding commitments, reinforcing Egypt’s liability in the dispute. Both cases demonstrate how minor linguistic differences, whether omissions, mistranslations, or additions, can have profound legal and political consequences, particularly in international law where precise language is paramount (Fathi n.d.).

3 Significance of the study

Whilst many scholars have tackled the machine translation quality assessment, proposed several models and matrices to distinguish translation errors, and even applied them on the output of machine translation systems, the analysis of automatically generated subtitles in legal contexts is not yet sufficiently tackled. This paper aims to fill this gap, while shedding light on the importance of integrating tools of technology in facilitating the access to public legal proceedings, especially regarding imminent cases and issues, such as humanitarian cases, war crimes, genocide, apartheid, etc.

Since technological trends emerge almost every single day, particularly in fields such as machine learning and natural language processing (NLP), it has become of utmost importance to investigate the potentials of neural machine translation (NMT) systems to further save time, effort and money, while providing audio–visual legal materials to whomever concerned, regardless of their location or mother tongue, achieving inclusion and democratization in media and information.

4 Objectives

This study aims at addressing the dire need for providing accurate live captions and fluent automatic translations to the specialized legal speeches, in broadcasted public hearings, especially those pertaining to sensitive humanitarian matters, such as the Crime of Genocide in the Gaza Strip, through assessing the Google neural machine translation (GNMT) system, activated and used by YouTube to render the automatically-generated English script into Arabic.

5 Research questions

The research endeavors to respond to the following research questions:

1. Considering the DQF-MQM Framework, what types of errors made by the GNMT?
2. Considering the DQF-MQM Framework, how many errors made by the GNMT?
3. What are the root causes of such errors made by the GNMT?
4. What is the score of the GNMT rendition of the entire session from English into Arabic, under terminology, accuracy, as well as fluency?
5. To what extent can legal entities and media channels depend on the GNMT system to render legal texts from English into Arabic?

6 Source of data

The data source of this study is Egypt’s oral arguments at the International Court of Justice’s (ICJ) hearings on the request for an advisory opinion on the Israeli occupation of Palestine, and the ongoing crime of genocide in the Gaza Strip, which was presented by Yasmine Moussa, Legal Advisor at the Cabinet of the Minister of Foreign Affairs of the Arab Republic of Egypt, on Wednesday, February 21st, 2024.

7 Theoretical framework

The error analysis in this paper is based on the DQF-MQM harmonized error typology, which is a joint effort by the Translation Automation User Society (TAUS) and the German Research Center for Artificial Intelligence (DFKI). This typology classifies errors into eight major issues as follows: accuracy, fluency, terminology, style, design, locale convention, verity, and an extra issue for any other unclassified errors.

Every category is then divided into further subcategories, reaching a total of 33 subcategories, as seen in Table 1. Moreover, this framework classifies errors on the basis of their severity level into four categories, which are critical, major, minor, as well as neutral errors, and assigns each severity level a penalty score: 10, 5, 1, and 0 penalty points, respectively. In this study, only three of these categories are used along with their subdivisions, namely accuracy, fluency, and style, in addition to three severity levels, which are critical, major, and minor errors.

It is not sufficient for raters to merely pinpoint the MT errors, since they must first be evaluated in terms of their severity, along with the nature of errors. In other words, errors must be assigned to certain severity levels with penalty points depending on how much a certain error affects the relevant performance. Indeed, the DQF-MQM framework has its own set of severity levels. DQF-MQM supports four severity levels as indicated below (Table 2).

Each severity level corresponds to penalty points used in scoring translations; depending on this score, the quality of the translated text is being judged. The default penalties are 10 points for critical errors, 5 points for major errors, 1 point for minor errors, and 0 for null. To calculate a score, the following formula is used:

$$\text{Score} = 1 - \left(\frac{\text{Penalties}}{\text{Word}}\text{Count}\right) \times 100$$

The above formula is clarified as follows: First, each severity level has a certain level of importance referred to as “weight.” Since some errors are more important than

Table 1: A subset of the DQF-MQM harmonized error typology.[a]

Number	Categories	Sub-categories
1	Terminology: Errors arising from using a term in the target content that is not equivalent to that of the source content or inconsistent with the relevant field. These errors might also arise from using incorrect terms.	(a) Inconsistent use of terminology (b) Wrong term: Using a term other than another correct one that should be used in a certain context or by an expert.
2	Accuracy: Errors arising when the propositional content of the source language is rendered distorted, omitted or with additional information in the target content.	(a) Addition (b) Omission (c) Mistranslation (d) Untranslated
3	Fluency	(a) Spelling (b) Grammer (c) Inconsistency

[a] For a full list of categories and sub-categories, visit https://themqm.info/typology/.

Table 2: DQF-MQM error severity levels.

Number	Severity level	Definition
1	Critical errors	Assigning an error to the critical severity level indicates that this error prevents a translation from fulfilling its purpose or even poses a risk for either serious physical, financial, or reputational harm.
2	Major errors	Assigning an error to the major severity level simply means that an error makes the intended meaning of the text unclear in a way that the intended user cannot recover the meaning smoothly but is unlikely to cause any real harm. It should be fixed before releasing the translation, since it may possibly annoy the user, because of a significant loss or change in meaning or because the error appears in a highly important part of the content
3	Minor errors	Assigning an error to the minor severity level indicates that the translation error has a limited impact on accuracy, fluency, stylistic quality, or even the general appeal of the content, but it does not, by any means, seriously hinder the usability, understandability, or reliability of the translated text.
4	Null	Assigning an error to the null severity level means that the evaluator considers that there is a more convenient solution. However, the word/phrase used by the translator is not an error.

others depending on the type of content being evaluated, the weight of such errors differ. For example, the “style” error might be less important than the “terminology” error in some contexts; thus, the former error would be of less weight than the latter. The DQF-MQM default weight is 1.0. Since this study adopts the analytic approach that tackles only three error categories, namely terminology, accuracy and fluency, the weight of these three categories remains 1.0. Second, “[e]ach severity level corresponds to” the framework default penalty points (Lommel 2018). Then, the penalty points are divided by the word count of the source content, and then the given value is subtracted from 1. Finally, since the final scoring is presented as a percentage, the recent given value is multiplied by 100.

8 Methodology

To address the research objectives, this study adopts a mixed-methods approach by qualitatively and quantitatively evaluating the Arabic output of Google neural machine translation (GNMT) using the DQF-MQM harmonized error typology. A four-step process is followed: extracting the English auto-generated transcript, obtaining the Arabic GNMT translation, annotating and analyzing translation errors, and scoring the MT output based on the identified error types. This analysis aims to assess the accuracy and fluency of the Arabic translation and identify the root causes of errors. The study ultimately seeks to determine the reliability of GNMT for translating public legal meetings into Arabic and to offer recommendations for improving the model's ability to handle complex linguistic features.

9 Analysis

Defining and annotating errors in this paper are influenced by the guidelines introduced by Burchardt and Lommel for utilizing MQM in scientific research on translation quality, as they defined an issue as any error in "the translated text that either does not correspond to the source or is considered incorrect in the target language" (p. 12). It is worth mentioning that the annotation here flags translation errors based on their general category recommended in the theoretical framework to avoid confusion and better understand and explain the error at hand.

9.1 Terminology

9.1.1 Inconsistent use of terminology

It can be recognized that there is an inconsistency error when translating "advisory opinion" as in most instances, it is translated into "فتوى", while in this example, it is translated into "رأي كوفو الاستشاري". A fluency error also occurs in this rendition; it should have been clearer and way less awkward to translate it into "الرأي الاستشاري بشأن إعلان استقلال كوسوفو". Here, it is worth noting that "Kosovo" was transcribed inaccurately by the ASR system adopted by YouTube, which is a reason for the inability of the MT system to fluently render the entire phrase. When correcting the spelling mistake in the source text, the GNMT system rendered this phrase into "فتوى كوسوفو".

Example 1:

Source Text:
"Let me recall that the court has repeatedly and consistently rejected such arguments in the **kovo Advisory opinion**."
GNMT Output:
"اسمحوا لي أن أذكر أن المحكمة رفضت مرارًا وتكرارًا مثل هذه الحجج في رأي **كوفو الاستشاري**."

Here in the following sentence, for example, “advisory opinion” is translated as “فتوى”.

Source Text:
“this **advisory opinion** of historical importance”
GNMT Output:
”هذه **الفتوى** ذات الأهمية التاريخية“

According to Dag Hammarskjöld Library, the advisory opinion is “legal advice provided to the UN or a specialized agency by the International Court of Justice (ICJ), in accordance with Article 96 of the UN Charter” (United Nations 2023). In addition, as per UNTerm, this term has two accepted and commonly used renditions into Arabic (look at the below screenshot of UNTerm) (Figures 1–4).

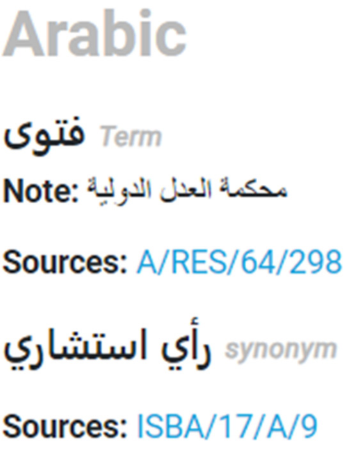
Talking about inconsistency in translate “advisory opinion” into Arabic, in the following two examples, “the wall opinion”, referring to the Advisory Opinion of the ICJ on the Construction of a Wall in the Occupied Palestinian Territory, is translated as “فتوى الجدار”, in one sentence, while in the other, it is translated into “رأي الجدار”. Here emerges a fluency error, since this rendition makes “the wall” seem as a person who has an opinion, which undoubtedly adds a layer of ambiguity to the entire paragraph.

Source Text:
“Allow me to recall that this very court in **the wall opinion** affirmed the UN’s permanent responsibility for the question of Palestine.”
GNMT Output:
”واسمحوا لي أن أذكر بأن هذه المحكمة بالذات أكدت في **فتوى الجدار** مسؤولية الأمم المتحدة الدائمة عن قضية فلسطين.“

Example 2:

Source Text:
“In **the wall opinion**, the court found no merit in the proposition echoed by some in these proceedings that the ongoing negotiations constituted a compelling reason to decline its competence.”
GNMT Output:
”في **رأي الجدار**، لم تجد المحكمة أي ميزة في الاقتراح الذي ردده البعض في هذه الإجراءات بأن المفاوضات الجارية تشكل سببًا مقنعًا لرفض اختصاصها.“

As shown in the below screenshot of translations extracted from a number of UN documents, “the wall opinion” is usually translated in this context as “الفتوى بشأن الجدار”.



Arabic

فتوى Term

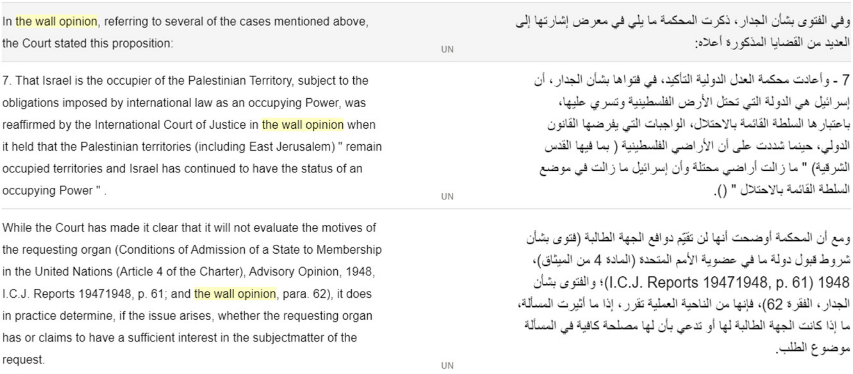
Note: محكمة العدل الدولية

Sources: A/RES/64/298

رأي استشاري synonym

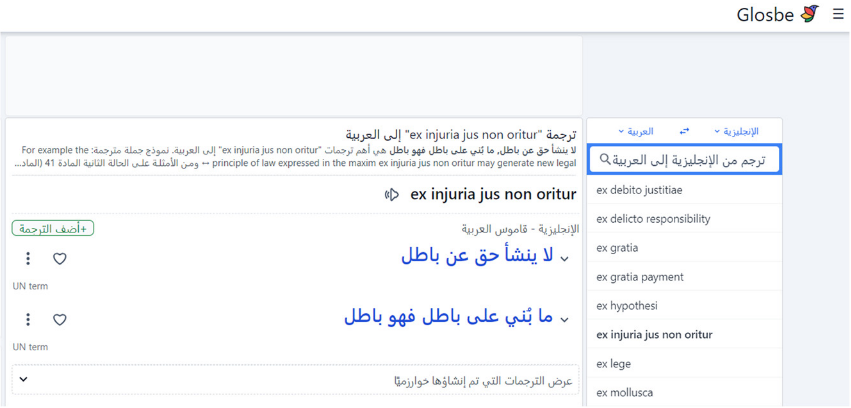
Sources: ISBA/17/A/9

Figure 1: The Arbaic translation of “advisory opinion” on UNTerm.



In the wall opinion, referring to several of the cases mentioned above, the Court stated this proposition:	UN	وفي الفتوى بشأن الجدار، ذكرت المحكمة ما يلي في معرض إشارتها إلى العديد من القضايا المذكورة أعلاه:
7. That Israel is the occupier of the Palestinian Territory, subject to the obligations imposed by international law as an occupying Power, was reaffirmed by the International Court of Justice in the wall opinion when it held that the Palestinian territories (including East Jerusalem) " remain occupied territories and Israel has continued to have the status of an occupying Power " .	UN	7 - وأعادت محكمة العدل الدولية التأكيد، في فتواها بشأن الجدار، أن إسرائيل هي الدولة التي تحتل الأرض الفلسطينية وتسري عليها، باعتبارها السلطة القائمة بالاحتلال، الواجبات التي يفرضها القانون الدولي، حينما شددت على أن الأراضي الفلسطينية (بما فيها القدس الشرقية) " ما زالت أراضي محتلة وأن إسرائيل ما زالت في موضع السلطة القائمة بالاحتلال " ().
While the Court has made it clear that it will not evaluate the motives of the requesting organ (Conditions of Admission of a State to Membership in the United Nations (Article 4 of the Charter), Advisory Opinion, 1948, I.C.J. Reports 19471948, p. 61; and the wall opinion, para. 62), it does in practice determine, if the issue arises, whether the requesting organ has or claims to have a sufficient interest in the subjectmatter of the request.	UN	ومع أن المحكمة أوضحت أنها لن تقيّم دوافع الجهة الطالبة (فتوى بشأن شروط قبول دولة ما في عضوية الأمم المتحده (الماده 4 من الميثاق)، 1948 (I.C.J. Reports 19471948, p. 61)؛ والفتوى بشأن الجدار، الفقره 62)، فإنها من الناحية العملية تقرر، إذا ما أثيرت المسألة، ما إذا كانت الجهة الطالبة لها أو تدعي بأن لها مصلحة كافية في المسألة موضوع الطلب.

Figure 2: The translation of “the wall opinion” in UN documents.



ترجمة "ex injuria jus non oritur" إلى العربية

لا ينشأ حق عن باطل, ما بُني على باطل فهو باطل هي أهم ترجمات "ex injuria jus non oritur" إلى العربية. نموذج جملة مترجمة: For example the principle of law expressed in the maxim ex injuria jus non oritur may generate new legal ↔ ومن الأمثلة على الحالة الثانية المادة 41 (الماد...

ex injuria jus non oritur

الإنجليزية - قاموس العربية

لا ينشأ حق عن باطل

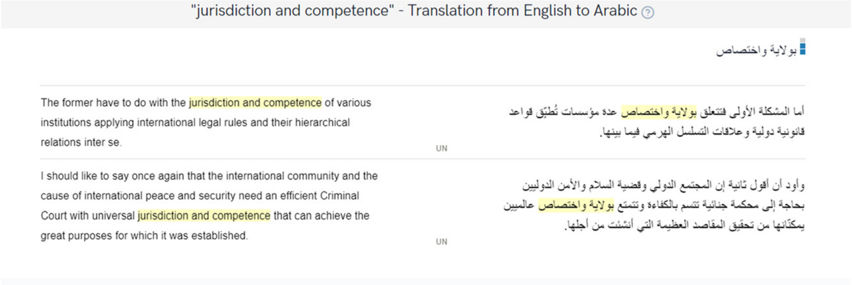
UN term

ما بُني على باطل فهو باطل

UN term

عرض الترجمات التي تم إنشاؤها خوارزميًا

Figure 3: The Arabic translation of “ex injuria jus non oritur” on Glosbe.



"jurisdiction and competence" - Translation from English to Arabic

بولاية واختصاص

The former have to do with the jurisdiction and competence of various institutions applying international legal rules and their hierarchical relations inter se.	UN	أما المشكلة الأولى فتتعلق بولاية واختصاص عده مؤسسات تُطبّق قواعد قانونية دولية وعلاقات التسلسل الهرمي فيما بينها.
I should like to say once again that the international community and the cause of international peace and security need an efficient Criminal Court with universal jurisdiction and competence that can achieve the great purposes for which it was established.	UN	وأود أن أقول ثانية إن المجتمع الدولي وقضية السلام والأمن الدوليين بحاجة إلى محكمة جنائية تتسم بالكفاءة وتتمتع بولاية واختصاص عالميين يمكنّانها من تحقيق المقاصد العظيمة التي أنشئت من أجلها.

Figure 4: The Arabic translation of “jurisdiction and competence” on Tarjamaan.

9.1.2 Wrong term

Although “Jus in Bello[1]” has been properly translated into “قانون الحرب” in multiple incidents (look at example 10), it has been incorrectly translated into “قانون بيلو” in the following sentence.

Example 3:

Source Text:
“Egypt respectfully submits that the court should advise the general assembly that number one the prolonged Israeli occupation is a continuing violation of international law for its breach of: number one **the Jus in Bello**,”
GNMT Output:
“وتدفع مصر بكل احترام بأن على المحكمة أن تبلغ الجمعية العامة بأن الاحتلال الإسرائيلي الذي طال أمده يمثل أولاً انتهاكًا مستمرًا للقانون
الدولي بسبب انتهاكه لما يلي: أولاً **قانون بيلو**،”

Additionally, the MT has rendered “jurisprudence” as “الحكمة القضائية”, in the following example, though it means “النظام القانوني”, “الفقه”, or “السوابق القضائية”, according to *Almaany Bilingual Dictionary*.[2] In this context, referring to previous similar cases where the ICJ advisory opinions have been delivered and have had relevance for the

[1] According to the International Committee of the Red Cross, “International humanitarian law, or jus in bello, is the law that governs the way in which warfare is conducted.” For more information, visit https://www.icrc.org/en/war-and-law/ihl-other-legal-regmies/jus-in-bello-jus-ad-bellum.
[2] Almaany English Arabic Dictionary. (accessed 2024-07-07).

ongoing debates in the general assembly, “this wealth of jurisprudence” should be translated as “هذه الثروة من السوابق القضائية”.

Example 4:

Source Text:
“the court could not possibly turn its back on **this wealth of jurisprudence**”
GNMT Output:
“لا يمكن للمحكمة أن تدير ظهرها لهذه **الثروة من الحكمة القضائية**”

9.2 Accuracy

9.2.1 Addition

In the following example, the engine has added an extra attached pronoun in “تحظرها”; this addition has negatively impacted the intended meaning. To render the meaning accurately, this phrase should be translated as “لا تحظر ... فحسب، بل تحظر أيضًا”

Example 5:

Source Text:
“it should be highlighted that article 49 not only **prohibits Force transfers but also in the Court’s own words and I quote any measure** taken by an occupying power in order to organize or encourage transfers of parts of its own population into the occupied territory.”
GNMT Output:
“**لا تحظر عمليات نقل القوة فحسب، بل تحظرها أيضًا على حد تعبير المحكمة نفسها، وأقتبس أي تدبير** 49 وتجدر الإشارة إلى أن المادة
تتخذه سلطة الاحتلال من أجل تنظيم أو تشجيع نقل أجزاء من سكانها إلى الأرض المحتلة.”

9.2.2 Mistranslating linguistic terms and phrases

In addition, the MT mistakenly rendered a well-known English phrasal verb “to second guess”, as in the following example:

Example 6:

Source Text:
“The court refuses to **second guess** the decision of the general assembly”
GNMT Output:
“ترفض المحكمة **التخمين الثاني** لقرار الجمعية العامة”

While “to second guess” means “to criticize someone’s decision after it has been made and say what you think is wrong with it”, in accordance with *Cambridge Dictionary*, the machine has rendered it literary, instead of rendering the intended meaning “قرارترفض المحكمة أن تنتقد” or “ترفض المحكمة أن تشكك في قرار”.

Moreover, according to *Collins Dictionary*, when we refer to anyone or anything as “august”, it means that this person or thing is dignified, notable or impressive. It is a formal term that is commonly used in the legal context; nonetheless, the MT has rendered it, as a variable noun not an adjective, into “أغسطس”, which is the eighth month of the year in the Western calendar. This word, being an example of polysemy, is inaccurately rendered by the MT system, since it cannot realize the context. A reason behind such an error may be that “August” has been written in upper case in the source transcript, which indicates that it is indeed a proper noun. This can be seen in this example:

Example 7:

Source Text:
“Distinguished members of the Court, the general assembly has turned to this **August** court”
GNMT Output:
“أعضاء المحكمة الموقرين، تحولت الجمعية العامة إلى محكمة **أغسطس** هذه”

Additionally, when facing one of the phrases commonly used in English: “which is not the case”, the system utterly distorted the meaning.

Example 8:

Source Text:
“Israel’s attack in 1967 was therefore not a defensive but an aggressive War even if the claim of self-defense were valid **which clearly is not the case**.”
GNMT Output:
“لم يكن دفاعيًا بل كان حربًا عدوانية حتى لو كان ادعاء الدفاع عن النفس صحيحًا، **وهو ما لا يحدث** 1967 وبالتالي فإن هجوم إسرائيل في عام
بوضوح.”

9.2.3 Mistranslation due to lack of context

Given that legal language relies heavily on precise wording and specific terms, a slight mistranslation due to lack of context may significantly alter the meaning and intent of the whole legal document. Legal documents also infrequently reference past rulings or precedents. Thus, without fully grasping the context, an NMT system may miss these nuances and produce an inaccurate translation. That is why analyzing mistranslation errors specifically related to lack of context is of utmost importance, when evaluating the output of any NMT system. As seen in the example below, the machine has mistakenly translated “members of the court” into “الإدارةأعضاء مجلس”.

Example 9:

Source Text:
“distinguished members **of the Court**”
GNMT Output:
“أعضاء **مجلس الإدارة** الموقرين”

Nonetheless, it was rendered correctly, in other instances, as follows:

Source Text:
“**Distinguished members of the Court**, the general assembly has turned to this August court”
GNMT Output:
“**أعضاء المحكمة الموقرين**، تحولت الجمعية العامة إلى محكمة أغسطس هذه”

Not comprehending the paratextual context has also led the machine to inaccurately render “for” in the following sentence into “من أجل”, instead of “وفيما يتعلق ب”.

Example 10:

Source Text:
“And **for the fundamental prohibition of racial discrimination segregation and subjugation**, it is against this legal framework that the legality of Israel’s policies and practices in the occupied Palestinian territories must be assessed first with respect to the Jus in Bello.”
GNMT Output:
”ومن أجل الحظر الأساسي للتمييز العنصري والفصل والإخضاع، فإنه يتعارض مع هذا الإطار القانوني ويجب تقييم
شرعية سياسات إسرائيل وممارساتها في الأراضي الفلسطينية المحتلة أولاً فيما يتعلق بقانون الحرب.“

9.2.4 Untranslated

It is arduous for the machine to translate legal maxims; legal maxims are concise and precise statements often encapsulating specific legal concepts or historical references within a particular legal system. Hence, translating them literally might lose their cultural and legal meaning and hinder capturing the full meaning or legal weight of the maxim in the target language. As seen in the chosen data, the speaker used one of the Latin legal maxims, “ex injuria jus non oritur”, which is not commonly used in everyday English. The GNMT system has failed to properly render this maxim into Arabic, so it has been kept as it is in English, as in the following example:

Example 11:

Source Text:
“This cannot be justified as a safety measure taken by Israel in the exercise of its prerogatives as an occupying power according to the legal Maxim **Ex injuria jus non oritur**. One should not be able to profit from one’s own wrongdoing.”
GNMT Output:
“ولا يمكن تبرير ذلك باعتباره إجراءً أمنيًا اتخذته إسرائيل في ممارسة صلاحياتها كقوة احتلال وفقًا للمبدأ القانوني **Ex injuria jus not oritur**: لا ينبغي للمرء أن يكون قادرًا على الاستفادة من أخطائه.”

According to *Glosbe*, this legal maxim has two Arabic equivalents, as demonstrated in the image below:

Legal phrases like “ex injuria jus non oritur”, which means that “one cannot rely on a violation of law to establish a new legal right or to confirm a claimed right”, as per the definition provided by *UNTerm* (Guide to Latin in International Law. 2nd ed. Oxford: Oxford University Press, 2022), carry a very specific legal weight that could not be captured by the NMT system, since it primarily focuses on translating natural language and is not properly equipped to handle such highly specialized phrasings.

9.3 Fluency

9.3.1 Arabic sentence parsing

Arabic sentence parsing (الإعراب) refers to the change that occurs in the final letters of words, namely nouns and verbs, when assigned diacritical marks or parsing marks. In example (12), a syntactic error is observed in “الموقرون”, which comes after “أعضاء المحكمة”, the vocative construction in this sentence, as follows:

Example 12:

Source Text:
“**Distinguished members of the court**, Palestine has been subjected to the longest protracted state of occupation in modern history”
GNMT Output:
”**أعضاء المحكمة الموقرون**، تعرضت فلسطين لأطول حالة احتلال في التاريخ الحديث“

To make the noun in the form of a sound masculine plural, “ون” ought to be added to the end of the word if it is a subject, and “ين” is rather added if it is in an object position, comes after a preposition or a vocative construction, or when it is the second noun of a genitive construction, according to The CJKI Database of Arabic Plurals (DAP), the first up-to-date database covering all types of regular and irregular Arabic plurals. Thus, it should be written as “الموقرين”, since it is a sound masculine plural noun جَمْعُ المُذَكَّرِ السّالِم describing and following a mudhaaf in nasb, the vocative construction.

9.3.2 Misuse of prepositions

Another translation error has been observed, where the speaker corrected herself, as in the following example:

Example 13:

Source Text:
“It is my great honor and privilege to appear on behalf **of Egypt of the Arab Republic of Egypt**”
GNMT Output:
“إنه لشرف وامتياز عظيم لي أن أمثل نيابة **عن مصر في جمهورية مصر العربية**”

Instead of simply omitting “of Egypt” or even rendering the entire phrase as “عن مصر عن جمهورية مصر العربية”, the machine used the wrong preposition “في”, meaning “in”, adversely impacting the meaning of this sentence.

9.3.3 Duplication

Duplicate words can also be recognized when analyzing the selected MT output. One of the causes of such error is the fact that two English words may be used in Arabic as synonyms that have the exact same equivalent in data training the NMT system.

Example 14:

Source Text:
“First, on the matter of **jurisdiction and competence**,”
GNMT Output:
“أولاً، فيما يتعلق بمسألة **الاختصاص القضائي والاختصاص** ”

In the above example, the term “الاختصاص” is duplicated in the Arabic MT rendition, when translating “jurisdiction and competence”. According to the explanation of *Britannica* on the difference between both jurisdiction and competence, on the one hand, “Competence refers to the legal ‘ability’ of a court to exert jurisdiction over a person or a ‘thing’ (property) that is the subject of a suit”. On the other hand, jurisdiction, which a competent court may exert, is “the power to hear and determine a suit in court.” Therefore, in most cases, these two terms are translated together into “الولاية والاختصاص”, as in the below image of the UN translations stored in a platform, called Tarjamaan,[3] which includes around 41,893,460 Arabic–English contextual translations and up to 49,218,730 bilingual Arabic–English example sentences.

Another example of unidiomatic duplication is the repetition of the word “وقائي” in the following sentence:

Example 15:

Source Text:
“In 1967, international law recognized neither **preemptive nor preventive self-defense** and the terms of the UN Charter on this matter are clear requiring an armed attack to occur in order to trigger the right of self-defense.”
GNMT Output:
”في عام 1967، لم يعترف القانون الدولي **بالدفاع الوقائي أو الوقائي عن النفس**، وأحكام ميثاق الأمم المتحدة بشأن هذه المسألة واضحة تتطلب وقوع هجوم مسلح من أجل تفعيل حق الدفاع عن النفس.“

According to Joe Barnes, in his paper titled “Preemptive and Preventive War: A Preliminary Taxonomy”, the preemptive military attack means any military action taken before the enemy initiates his, when there is irrefutable evidence that the enemy is about to attack, while the preventive military attack rather refers to any military action initiated, when one believes that the enemy’s attack is inevitable, even if it is not imminent, and fears that the risk of such attack will be greater with the passage of time. Based on this comparison, “international law recognized neither preemptive nor preventive self-defense” should be translated into “لم يعترف القانون الدولي بالدفاع الاستباقي أو الوقائي عن النفس”. Moreover, in the following example, both siege and blockade are translated into “الحصار”.

[3] https://tarjamaan.com/en/about.

Example 16:

Source Text:
“One only needs to look at Israel’s vicious wholesale destruction of Gaza today after years of **imposing the medieval methods of Siege and blockade** to realize the extent of Israel’s transgression of this principle.”
GNMT Output:
”ولا يحتاج المرء إلا إلى النظر في تدمير إسرائيل الشامل الوحشي لغزة اليوم بعد سنوات من فرض أساليب الحصار
والحصار في العصور الوسطى لإدراك مدى تجاوز إسرائيل لهذا المبدأ.“

According to *Collins Dictionary*, siege refers to the military operation aiming at surrounding a place to force people to come out of the place and give up. Yet, blockade means preventing people or goods from getting into or out of a certain place. When someone blockades a road or a port, it means that they are hindering the use of that road or port. Consequently, while the first is military, the latter is rather economic (Figures 5–8).

This image demonstrates the suggested rendition of blockade into Arabic as “حصار اقتصادي”, or “حصار سلمي”. Thus, this above-mentioned example should have been translated into “فرض أساليب القرون الوسطى في الحصار العسكري والاقتصادي”. The same error also appears in the rendition of “aid and assistance” hereinafter, as both words are formal synonyms of “help”. Thus, both have been translated into “المساعدة”, affecting the style of the entire sentence.

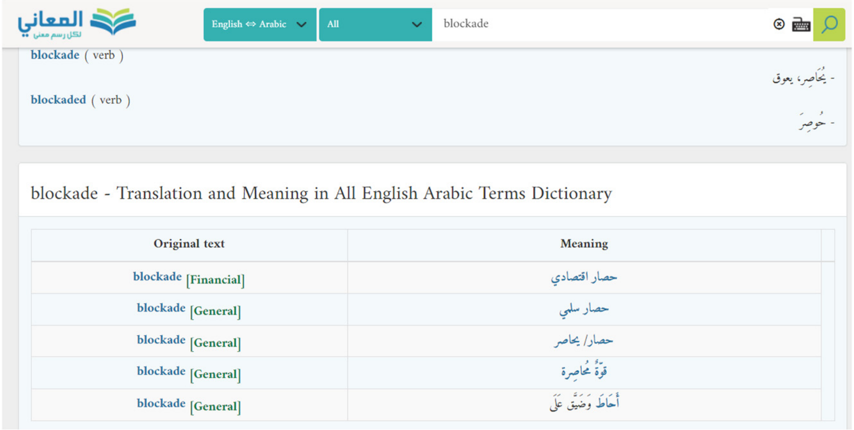
Example 17:

Source Text:
“Three, all states have a duty not to recognize the illegal situation created by Israel’s ongoing violation resulting from its prolonged occupation settlement and annexation of the occupied territory and not **to render Aid or assistance in maintaining that situation**.”
GNMT Output:
”ثالثًا، يقع على عاتق جميع الدول واجب عدم الاعتراف بالوضع غير القانوني الناشئ عن انتهاك إسرائيل المستمر الناتج عن احتلالها المطول
واستيطانها وضم الأراضي المحتلة، وعدم تقديم المساعدة أو المساعدة في الحفاظ على هذا الوضع.“

Based on the below screenshot taken from *Almaany bilingual dictionary*,[4] it is preferable to translate the above-tackled example into “عدم تقديم المعونة والمساعدة”, to avoid any sort of ambiguity or awkwardness.

Employing the DQF-MQM harmonized model to evaluate the performance of the GNMT system in translating the chosen data, the analysis holistically and

[4] Almaany English Arabic Dictionary. (accessed 2024-07-10).



blockade (verb)

- يُحَاصِر، يعوق

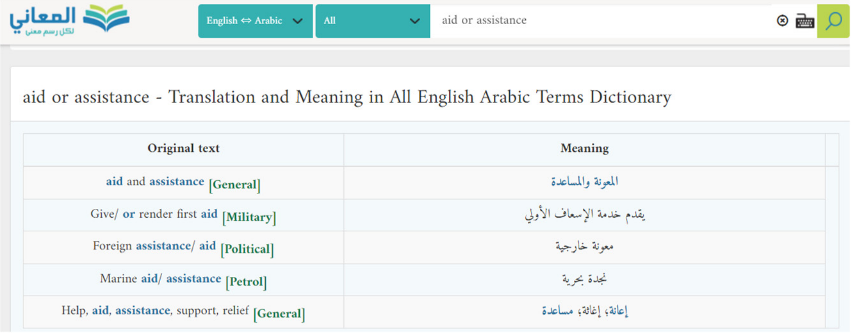
blockaded (verb)

- حُوصِرَ

blockade - Translation and Meaning in All English Arabic Terms Dictionary

Original text	Meaning
blockade [Financial]	حصار اقتصادي
blockade [General]	حصار سلمي
blockade [General]	حصار/ يحاصر
blockade [General]	قُوَّة مُحاصِرة
blockade [General]	أَحَاطَ وَضَيَّق عَلَى

Figure 5: The Arabic translation of “blockade” on Almaany Dictionary.



aid or assistance - Translation and Meaning in All English Arabic Terms Dictionary

Original text	Meaning
aid and assistance [General]	المعونة والمساعدة
Give/ or render first aid [Military]	يقدم خدمة الإسعاف الأولي
Foreign assistance/ aid [Political]	معونة خارجية
Marine aid/ assistance [Petrol]	نجدة بحرية
Help, aid, assistance, support, relief [General]	إعانة؛ إغاثة؛ مساعدة

Figure 6: The Arabic translation of “aid” and “assistance” on Almaany Dictionary.

qualitatively identifies the translation errors conducted by the machine and the overall weaknesses of the system used in handling legal terminology, accuracy, fluency, in addition to adherence to legal conventions. In the following section, all errors are analyzed quantitatively, to reach the overall score of the entire MT output and come up with a conclusion on whether the GNMT system has performed well in translating information, dealing with specific legal terminology, and rendering the intended meaning fluently from English into Arabic.

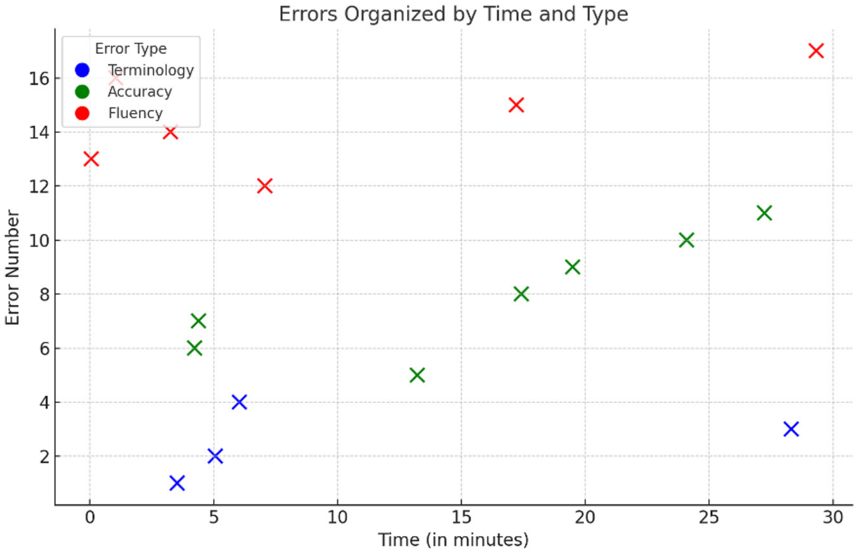


Figure 7: Errors by time and type (Terminology, accuracy, fluency).

This section presents an analytic evaluation of the machine translation (MT) output of the English transcript, focusing on the severity and overall scoring of identified errors. The evaluation involves classifying errors under specific issues, while assigning severity levels and corresponding penalty points (Tables 3–5).

Overall quality score (OQS) of the automatically translated Egyptian arguments under each issue

This section is dedicated to illustrating penalty points and scores under the three issues: Terminology, accuracy, and fluency.

Terminology scoring:

- Total of penalties = 26
- Score = 1 – (26/3553[5]) = 0.99
- Score in percentage = 99.2 %

Accuracy scoring:

- Total of penalties = 41
- Score = 1 – (41/3,553) = 0.98
- Score in percentage = 98.8 %

[5] Word count of the english source transcript.

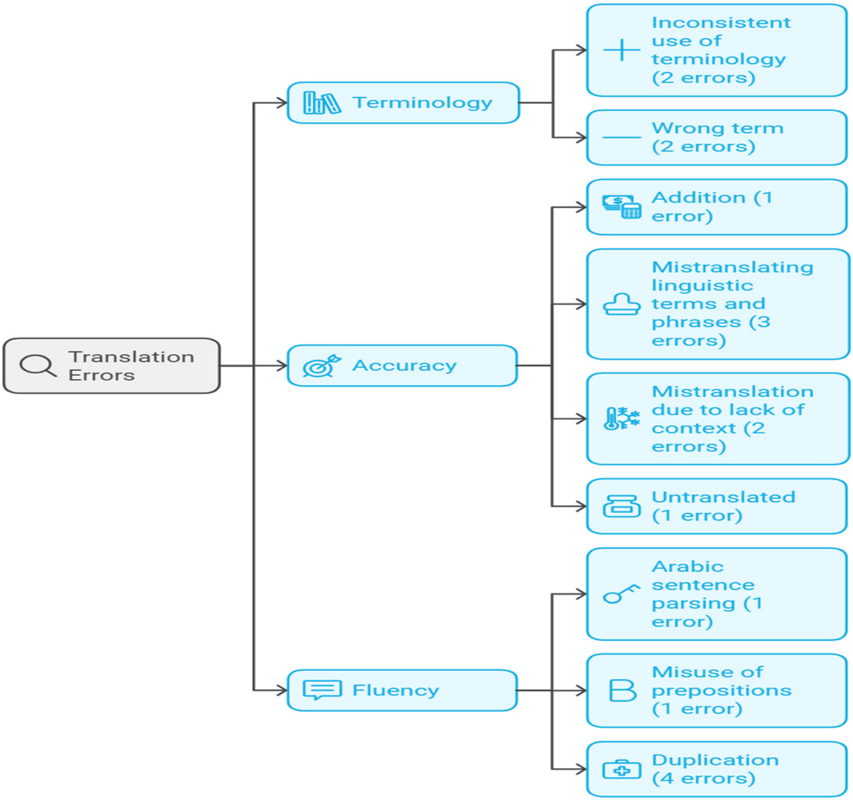

Translation Errors
Terminology
Inconsistent use of terminology (2 errors)
Wrong term (2 errors)
Accuracy
Addition (1 error)
Mistranslating linguistic terms and phrases (3 errors)
Mistranslation due to lack of context (2 errors)
Untranslated (1 error)
Fluency
Arabic sentence parsing (1 error)
Misuse of prepositions (1 error)
Duplication (4 errors)

Figure 8: Translation errors classified by types (Terminology, accuracy, fluency).

Table 3: Errors under terminology.

Number of examples	Severity level	Penalty point
1	Minor	1
2	Critical	10
3	Critical	10
4	Major	5

Fluency scoring:

- Total of penalties = 26
- Score = 1 – (26/3,553) = 0.99
- Score in percentage = 99.2 %

Table 4: Errors under accuracy.

Number of examples	Severity level	Penalty point
5	Major	5
6	Critical	10
7	Critical	10
8	Major	5
9	Major	5
10	Major	5
11	Minor	1

Table 5: Errors under fluency.

Number of examples	Severity level	Penalty point
12	Minor	1
13	Major	5
14	Major	5
15	Major	5
16	Major	5
17	Major	5

Distribution of errors by time and explicit comparison between segments of the argument

Here is the graph that visualizes the errors by their types (Terminology, accuracy, fluency) and arranged in ascending order by time. Each error is represented as a point, with different colors[6] indicating the error type:

While the *x*-axis represents the time in minutes, the *y*-axis shows the error numbers. This distribution aims at facilitating the recognition of when and what type of errors occur over time. This timeline provided in the graph for the MT errors made during rendering the speech provides necessary patterns in the frequency of the types of errors present in the final output.

To start with, regarding terminology errors, this type of error in the argument is quite scattered across the text but looking at the critical terminology errors, it is noted that they occur around the midpoint of the arguments section (e.g., 5:07, 28:31). This indicates that during the discussion of complicated legal terms, such as “Jus in Bello,” errors in translating or interpreting such specific legal terms arise.

[6] Blue: Terminology errors/Green: Accuracy errors/Red: Fluency errors.

For accuracy errors, the highest points of the concentration of critical accuracy errors occur at earlier times in the timeline, particularly during the introduction and the first argumentative stage. For example, critical errors occur as early as 4:23 and 4:39, during the critical changeover from the session delivering opening remarks to that of developing the argument. These errors are conceivably related to the mistranslations of the key phrases which prepared the legal dramatizations. Finally, fluency errors, in contrast to others, are dispersed, yet they tend to be less severe in terms of penalty points (many of them are marked as “minor”). These errors, however, are persistent across all sections, from the introduction, at 0:06, to the conclusion, at 29:32.

When dividing the speech into three sections (Introduction, from 0:00 to 5:00; arguments, from 5:00 to 25:00, and conclusion, from 25:00 to end, it can be recognized that the introduction shows fluency and accuracy errors early on, indicating that issues with phrasing and precise articulation occur here. Such errors include the misuse of prepositions, as the case when rendering the phrase “of Egypt of the Arab Republic of Egypt” at 0:06). Early critical accuracy errors, at 4:23 and 4:39, suggest misunderstandings of key legal concepts right at the beginning. These errors may undermine MT trustworthiness and clarity in rendering context.

In the second section, the highest rate of critical errors has been observed, especially in both terminology and accuracy. Since this part involves complex legal terms, concepts, and arguments, any mistakes can undoubtedly lead to a misrepresentation of legal points. For example, at 28:31, the inaccurate rendition of “Jus in Bello” has created a critical misunderstanding of the core legal arguments. Due to the sensitivity of speech in this section, these errors are even more damaging, as they affect the core logic of the case being made. Furthermore, the conclusion also features fluency and accuracy errors, though at a somewhat lower rate. While fewer critical errors occur here, any MT errors at this stage can deeply weaken the overall message, which is vital for summarizing and reinforcing the main points.

10 Discussion of findings

Since the International Court of Justice (ICJ) is tantamount to being the principal judicial organ of the United Nations, responsible for settling legal disputes between states, ensuring clear and accurate communication across languages in such a setting is paramount. This study has delved into the potential of English–Arabic MT, to support the ICJ in its multilingual proceedings, using an established evaluation framework, the DQF-MQM model, specifically designed for MT evaluation.

Out of 17 translation errors conducted by the GNMT system, four errors are under terminology, seven under accuracy, and six under fluency. Accuracy has the largest number of errors, specifically under mistranslation, with a total of five errors divided into linguistic errors and errors due to lack of context.

The analysis demonstrates that the machine has failed to maintain consistency when translating legal terminologies twice, translated a term incorrectly once and kept a legal maxim untranslated also once. These instances have surely affected the level of accuracy of the entire MT output, along with other instances where the machine has even failed to render commonly used linguistic terms or get the background context of the speech, leading to ambiguity. In addition, duplication, a sub issue under fluency, has the second largest number of errors among all other sub issues: four errors of repetition; using synonyms on some occasions has led to the machine repeating the exact same Arabic word, adversely impacting the flow of the entire MT output.

Based on the above analytical quantitative evaluation of the GNMT Arabic output, the overall quality score (OQS) of the automatically translated Egyptian arguments under terminology is 99.2 %, under accuracy is 98.8 %, and under fluency is 99.2 %. These results, unequivocally, show that the neural machine adopted by Google has had an outstanding performance translating the entire speech of the representative of the Arab Republic of Egypt, lasting for 29:59 min; that indeed boosts hopes for depending on machines to save time, effort and money when subtitling similar legal audio-visual material.

The distribution of errors by time also explains that the highest rate of critical errors appears in the arguments section, which indicates that high stakes in terminology and complexity of content jeopardize the MT performance. On one hand, legal arguments usually involve nuanced language, references to international law, and technical terms. Thus, translating or interpreting such content accurately requires a deep understanding of both languages and legal systems. On the other hand, misinterpreting a legal term, such as “Jus in Bello” can utterly alter the meaning of an argument. As the speech progresses, the focus shifts from general introductions to specific legal arguments, which makes errors more impactful and critical, as any small mistake can lead to major misunderstandings of the legal reasoning.

To start with, when examining the root causes for such translation errors, lack of context is a major hurdle for neural machine translation (NMT) systems, leading to mistranslations; although the NMT systems rely on statistical patterns learned from vast amounts of translated text, a single word or phrase can have multiple meanings depending on the context. Without understanding the surrounding text, the system might pick the wrong translation. Moreover, the sentence structure and word order can significantly alter the entire meaning, since NMT systems primarily go for word-level translation, potentially missing the nuances conveyed by sentence structure. Usually, idioms, and cultural references, among others, heavily depend on context. Thus, NMT systems often struggle to understand these references and translate them literally, leading to nonsensical or even offensive outputs. In crucial scenarios where accuracy is paramount, as the case in the legal domain, combining the GNMT system with human expertise guarantees the best results, as human translators can definitely leverage the machine's speed for repetitive tasks while ensuring contextually appropriate translations. By deeply understanding these above-mentioned limitations and exploring potential solutions, NMT systems will be more adept at handling context and producing accurate translations across diverse situations.

Furthermore, to avoid untranslatedness when dealing with Latin terms and maxims, some NMT systems allow users to flag specific terms or phrases for special treatment. Thus, the translator could then flag legal Latin phrases to ensure they are not translated automatically and are reviewed by a human translator. It is also crucial for a human translator to review the NMT output, identify untranslated terms, and provide the most accurate equivalent in Arabic, considering the legal context. A bilingual glossary of specialized legal terms with their Arabic equivalents can also be integrated with the NMT system used to automatically insert the correct translations for such terms.

Studying duplication errors, potential solutions to this issue may be improving synonym recognition. Advancements in NMT architecture and training data sets could play a pivotal role in better identification and differentiation of close synonyms within the source language. These advancements may encompass incorporating synonym dictionaries, or embedding knowledge graphs into NMT models, so they could improve their abilities to understand the nuanced context of a sentence, and opt for the most appropriate synonym, based on the surrounding context, words, and sentence structure.

Based on the above-stated findings and recommendations, MT systems offer a path towards increased efficiency and improved accessibility within the legal sphere. Nonetheless, integrating human expertise through pre-editing and post-editing remains crucial, particularly for critical legal translations where both accuracy and nuance are essential. It is also paramount to further assess the performance of such tools and acknowledge their limitations, when dealing with more technical and diversified formal legal texts.

11 Conclusions

This paper explores the potential of neural machine translation (NMT) to produce accurate and fluent subtitles for legal audio-visual material, specifically analyzing the Arabic output generated by Google neural machine translation (GNMT) for Egypt's oral arguments in support of South Africa's genocide case against Israel at the International Court of Justice (ICJ), uploaded on DawnNews English on February 21, 2024. Using the DQF-MQM harmonized error typology, the study evaluated the accuracy, fluency, and terminology in the machine-translated Arabic version of the official English transcript.

While the DQF-MQM harmonized error typology outlines eight major error categories, such as style, locale convention, and design, the present study focuses on three primary issues: terminology, accuracy, and fluency. This narrowing of scope is intentional and grounded in the specific nature of the source material, which consists of formal oral arguments delivered at the International Court of Justice (ICJ). In this context, terminology precision, factual accuracy, and linguistic fluency are the most critical elements for ensuring legal clarity and preserving the speaker's intent. While style and locale convention are undoubtedly relevant in broader legal translation contexts, they were less pronounced in this particular case, where the tone is formalized and globally oriented by default. Future work could certainly expand the analysis to include these additional categories, particularly when evaluating localized legal documents or more stylistically varied legal genres.

The analysis followed a structured process: extracting the English auto-generated transcript, obtaining the Arabic GNMT output, annotating translation errors, assigning severity-based penalty points under three key categories – terminology, accuracy, and fluency – and calculating final scores. This enabled a comprehensive assessment of whether the MT output met basic standards for legal translation. The study also investigated the underlying causes of translation errors and offered recommendations for improving NMT performance in legal contexts.

Findings showed that GNMT generally succeeded in conveying information but struggled with legal terminology, Latin expressions, idiomatic language, and complex sentence structures. These challenges stem from the limitations of general-purpose MT systems when applied to the specialized and nuanced language of legal discourse. Despite these issues, the study affirmed that MT systems like GNMT hold promise for enhancing legal translation workflows, provided they are used in conjunction with human post-editing.

To address the current limitations of NMT systems in legal contexts, this study recommends several strategic advancements. First, training NMT models on domain-specific legal corpora and involving legal experts in the development process can significantly improve the handling of legal terminology and stylistic conventions. Equally important is the refinement of evaluation metrics tailored to the complexities of legal discourse. The integration of context-aware systems is also crucial, particularly for maintaining coherence in lengthy documents. Additionally, the development of multi-source and multilingual translation models can enhance cross-linguistic consistency and reliability, both essential for legal accuracy. Special attention must also be paid to low-resource legal languages, where the scarcity of training data hampers translation quality. Leveraging techniques such as transfer learning, along with institutional collaborations to build specialized legal datasets, can help bridge this gap and promote more inclusive legal communication.

Alongside these technical improvements, the study highlights the need for comprehensive ethical and regulatory frameworks governing the use of NMT in legal settings. These frameworks should address issues of liability, data privacy, and clearly define the permissible boundaries for machine translation in legal contexts. Real-time speech translation also emerges as a priority for future research, especially in courtroom settings involving non-native speakers. Enhancing speech-to-text NMT systems, managing disfluencies, and exploring multimodal translation approaches are proposed as promising areas of development.

Crucially, the ethical implications of using NMT in legal environments deserve deeper consideration, particularly regarding access to justice. Mistranslations can lead to severe consequences, such as misleading clients, misrepresenting legal arguments, or influencing verdicts. These risks are amplified in low-resource languages, where underperforming MT systems may exacerbate existing inequalities. Therefore, NMT should be regarded as an assistive tool rather than a replacement for professional legal translators. To mitigate risks, the adoption of mandatory human post-editing and the establishment of strict quality assurance standards are essential components of any ethically sound NMT deployment in legal contexts.

Additionally, it is important to acknowledge that the study's analysis is based on a single transcript, the Egyptian oral arguments before the ICJ. This focused approach offers in-depth insight into the performance of NMT in high-stakes legal contexts, but it does limit the generalizability of the findings. Legal language and translation challenges can vary considerably across genres (e.g., contracts, court rulings, statutes) and across different source languages and dialects. Future research should therefore aim to replicate this methodology across a broader range of legal documents and scenarios. Doing so would help validate the findings and support more robust conclusions about NMT's utility and limitations in legal translation workflows.

While the conclusion touches on future research directions, further elaboration can clarify the technological and collaborative pathways ahead. For instance, advancements in large language models (LLMs) such as GPT-4 offer promising capabilities for context-aware and semantically nuanced translation. These models could enhance legal translation by better capturing syntactic structures, idiomatic expressions, and domain-specific terminology. Partnerships between legal institutions, academic bodies, and multilingual communities could also facilitate the annotation and validation of translated legal texts, helping to improve MT performance across diverse legal systems. Combined, such innovations could support a more inclusive and accurate multilingual legal landscape.

References

Carl, Michael. 2018. Machine translation in law: A promising future but challenges remain. *The Law Technology Journal* 13(2). 112–117.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way & Georgakopoulou Panayota. 2018. Evaluating MT for massive open online courses. *Machine Translation* 32(1). 255–278.

DAG Hammarskjöld Library. 2023. *Research guides*. United Nations. https://research.un.org/ (accessed 12 June 2025).

Díaz-Cintas, Jorge. 2013. The technology turn in subtitling. In Marcel Thelen, Barbara Lewandowska-Tomaszczyk, John Newman & Bernard Darras (eds.), *Translation and meaning – Part 9*, 345–352. Zuyd University of Applied Sciences. https://doi.org/10.4324/9781315749129.ch39 (accessed 12 June 2025).

Fathi, Mohamed Ahmed. n.d. A translation error that cost us millions. https://www.fathiahmed.com/arabic/articles/a-translation-error-that-cost-us-millions (accessed 12 June 2025).

Fathi, Mohamed Ahmed. n.d. Translation of SC Resolution 242. https://www.fathiahmed.com/arabic/articles/translation-of-sc-resolution-242 (accessed 12 June 2025).

Jackson, Bernard. 1985. *Semiotics and legal theory*. London: Routledge.

Lommel, Arle. 2018. Metrics for translation quality assessment: A case for standardising error typologies. In Joss Moorkens, Sheila Castilho, Federico Gaspari & Sharon Doherty (eds.), *Translation quality assessment: From principles to practice*, 109–127. Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_6 (accessed 12 June 2025).

Mackenzie, Hazel. 2019. The machine translation revolution and legal interpretation. *The Modern Law Review* 82(3). 542–568.

Popović, Maja. 2018. Error classification and analysis for machine translation quality assessment. In Joss Moorkens, Sheila Castilho, Federico Gaspari & Sharon Doherty (eds.), *Translation quality assessment: From principles to practice*, 129–158. Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_7 (accessed 12 June 2025).

Przybyszewski, Monika. 2017. Cost-effective solutions for legal translation: Can machine translation deliver? *International Journal of Law, Policy and Technology* 11(1). 78–92.

Tiersma, Peter M. 1999. *Legal language*. London: University of Chicago Press.

United Nations. 2023. What is an advisory opinion of the International Court of Justice (ICJ)? Ask UN. https://ask.un.org/faq/208207 (accessed 12 June 2025).

Bionote

Aya Sayed Omran Elsayed
Ain Shams University, Cairo, Egypt
Aya.omran.97@alsun.asu.edu.eg
https://orcid.org/0009-0000-7347-2007

Aya Sayed Omran Elsayed is an MA candidate and a teaching assistant at Faculty of Al-Alsun, Ain Shams University, Egypt. She has worked as a full-time translator, editor, and interpreter for 3 years, which pushed her towards examining ways of using technology to further help translators and enhance the quality of translation. She won the first place in ASU Innovates, a competition held every year by ASU Ihub, for developing an idea of establishing a platform to train and help both translators and interpreters, using artificial intelligence and automatic speech recognition. She has published a research paper in philology, a semi-annual, peer-reviewed academic journal issued by Al-Alsun Faculty and supervised by a number of academics and specialists from various universities around the world, entitled “MQM Evaluation of Arabic Machine Translation of the 1st Lecture of a Harvard Online Course” (2009).