

**Assessing METEOR and MQM Effectiveness in Measuring the Quality of Machine Legal Translation from English into Arabic:
A Corpus-based Study**

Ali Ali NAEIMAH AL SWUEE ^{1*}

Mohamed BABCHIKH ²

¹ *Laboratoire de Didactique de la Langue et des Textes
University Yahia Farès of Medea, Algeria
ali.naeimah@univ-medea.dz*

² *Laboratoire de Didactique de la Langue et des Textes
University Yahia Farès of Medea, Algeria
babachikh.mohamed@univ-medea.dz*

Received date: 06/07/2024

Revised date: 20/12/2024

Publication date: 22/12/2024

Abstract:

This study examines the role of Quality Assessment Metrics METEOR and MQM framework in assessing the accuracy of legal Amazon Translate output. The researchers assume that the full reliance on the use of automatic metrics in assessing Machine Translation (MT) quality cannot lead to achieving reliable results. Instead, they recommend assessing MT production based on MQM typology. The findings showed that the MQM assessment was mostly faithful in rating the MT production. However, in order that METEOR achieves precise evaluation, it should correlate with human translation, since the MQM framework revealed that Amazon Translate output has inferior quality, and the same is true for METEOR.

Keywords: Quality Metrics; MQM framework; METEOR; Legal translation. Machine Translation; Quality assessment; Amazon Translate.

1- Introduction

The advancement of MT has raised the bar of expectations in improving the translation landscape. This made translators obliged to

* **Corresponding author:** Ali Ali NAEIMAH AL SWUEE, ali.naeimah@univ-medea.dz

evaluate the quality of their productions. In this respect, quality assessment is considered as a key dimension that emerged as a controversial area within the field of automatic translation. It oriented researchers' attention towards measuring MT effectiveness in translation in general. However, adopting MT is always dependent on quality standards in one way or another.

According to Lommel et al. (2014: 1) to achieve a more effective automatic translation, translators should consider the type of the software, the domain and the assessment tool, regardless of its nature, being manual or automatic. These elements constitute the core of MT's work. If the chosen system is suitable for the field or area of translation, a high quality will be achieved, as different translation projects demand different assessment methods.

Recently, quality measurement has been the central preoccupation of MT research. The ability to measure the quality is a confirmation step in trusting MT's output. One may ask, how can this be achieved? Since many researchers have different experiences using manual or automatic assessment, each confirm the success of different tools considering their languages and cultural requirements. Whatever the language is, judging MT quality is based on various considerations, including the system of the language, the domain of specialty, and the culture suitability, which are all part of the appropriateness of the assessment method.

What about assessing translations into Arabic? Are automatic metrics more reliable than human assessment? Could they be accurate? To answer these questions, it is of major importance to highlight that the Arabic language has distinctive features that are different from those of English. Therefore, using MT systems to increase translators' productivity requires one to measure their quality and decide on the assessment tools. In legal translation, this is a capital requirement, since such translation is a highly specialized and the subject is very sensitive. Therefore, the challenges are different, and assessment requires a high-performance metric. The researchers outline an argument about how translators determine the assessment tools, and which will provide accurate quality measurement.

This research deals with MT quality metrics, i.e., MQM framework and METEOR and their effectiveness to assess English-Arabic legal translation over the correlation with human translation and judgment. These two acronyms stand for:

- MQM: "Multidimensional Quality Metric".
- METEOR: "Metric for Evaluation of Translation with Explicit Word Ordering."

Although human TQA models and MT assessment metrics are different, each seeks to meet the end user's requests and provide a precise evaluation of the MT-generated texts. Metrics are defined and used to evaluate different types of texts. The present study attempts first to apply METEOR and MQM to assess the quality of legal translated generated by Amazon Translate. It also seeks to judge the effectiveness of these metrics based on the analysis of their assessment.

The outcomes of this study may help users to choose the best assessment metric and open the gateway for more developing processes, particularly in evaluating English-Arabic legal translation. It also reflects on MT development by systematically assessing one of the new systems. This also positively affects the research database relating to law-language resources in adopting MT in different domains and assessing its quality.

To date, surveys have not tracked their liability of the quality assessment tool named above, particularly with Arabic legal translation. The researchers presuppose that METEOR needs to show more precision in evaluating the quality of legal translation due to limitations and challenges. Instead, MQM, as a manual framework, offers a precise evaluation as it takes into account the cultural and linguistic dimensions.

This study assesses the METEOR metrics and MQM framework performance by comparing MT to a human reference translation to determine the appropriate metric for assessing legal machine translation from English into Arabic. Therefore, generalizing the results of this research to other language pairs and metrics may be ineffective.

2- Methodology

For this study, the researchers will choose a divorce petition and its human and automatic translations as a corpus. This judicial document was filed to Bossier Parish Clerk of Court on Feb 14, 2023, at 10:42 am. It was translated into Arabic by a sworn translator. They will conduct four basic steps to achieve valid results, including error analysis. First, they shall compute the METEOR score of Amazon Translate's output, focusing on the exact match and measuring "the precision, recall, accuracy, fluency, style, the absolute penalty, and the total METEOR scores." In the second phase, they will calculate MQM scores, identify error types based on MQM error typology, and categorize them (natural, minor, major, or critical errors) according to their influence on the legal content. The last phase includes computing the coefficient and comparing the two scores, METEOR and MQM, to identify the strength of the two metrics. This stage is evaluated based on fluency, meaning, severity of errors, and absolute penalty.

3- Legal Texts and Translation Challenges

Legal language can be defined as specific-purpose language that includes a technical terminology. Moreover, it has special linguistic features that must be clarified to understand their meaning. Legal texts are then categorized according to their functions, being especially descriptive or prescriptive, and then making clear distinction between regularity documents such as laws and contracts or official requests and affidavits. Another categorization of a legal text is related to the producers themselves, namely, the legislative bodies and institutions such as parliaments, each with their own style. Other types of producers are the courts, which issue judicial documents.

Most of these texts share several characteristic features, regardless of their purpose. However, each has a unique characteristic that differentiates it from other types. For instance, international documents might encompass more precise terminology specific to the subject matter of particular States' issues than agreements or contracts, which also affects the understanding process. This variation of legal text types reflects that various individual

users, specialists, and entities deal with legal language nationally and internationally.

Like in any other field, legal translators may encounter challenges during translation. Apparently, to produce their translations, they rely on differences in law systems, language, and culture. According to Benson (1985), the legalese content of legal documents is challenging to understand, as they include long sentences, passive verbs, impersonality, and double negatives with stuffed ideas, illogical words, and poor headings. Each feature above requires special skills, a previous legal language style, and knowledge of text categories. Husni and Newman (2015) agree with Benson's idea that legal language has "overarching terms." Furthermore, Varó and Brian Hughes (2014, 152) demonstrated that "in any type of translation, the problems that crop up are likely to be not merely linguistic but sociocultural too... in a specialized field like the legal translation".

As Cabanellas (2017) signals, the diversity of the legal system affects the translation decision since each has created rights and obligations suitable to the States's users. Wiesmann (2019), citing Matthiesen (2017:6-44), argues that legal texts have a range of characteristics related to the legal system, text type, and law branches. These are considered the significant challenges encountered in Machine Translation. Altarabin (2021, 90) states that one cannot spread language from culture and the law from culture. These factors are scaffolding on each other in causality relation. Therefore, legal translation can have severe consequences if translators are unfamiliar with it.

Due to such factors and to satisfy the end users, legal translators must judge the text quality, be it human or produced automatically. In this case, they are required to select the proper evaluation method, manual or automatic, which is a challenging task.

The following section will provide details on the characteristics of the METEOR and MQM frameworks as metrics for assessment.

4- Quality Evaluation Metrics

Prior research in translation quality assessment has predominantly recommended different metrics for evaluating MT output, such as BLEU and METEOR, which are described as the most usable. According to Papineni et al. (2002), manual methods are tool-slow and time-consuming. Therefore, automatic assessment gives quick feedback on the level of quality. Depraetere (2011) notes in this regard that " This kind of evaluation is based on the assumption that the closer the MT is to a human translation, the better it is" (p.113). According to Dobre (2015), the primary purpose of using metrics is to compare different translations in terms of their quality by calculating the scores of each translation. She confirms, most importantly, that it is highly probable that MT systems will be developed using NLP metrics.

4-1. METEOR Metric

METEOR was created to overcome the weakness of BLEU "The Bilingual Evaluation Understudy" as a standard metric. Researchers focused on METEOR as the best quality metric for evaluating MT output. According to Agarwal & Lavie (2008), METEOR shows a radical development in the correlation compared with BLEU and TER "Translation Error Rate," which was adjusted to an optimized correlation with a limited capacity to capture adequacy and fluency aspects. METEOR is an automatic metric created and launched in 2004 to upgrade the existing metrics. What sets it apart is that it includes the F-mean score utilized to balance the precision and recall scores.

METEOR's assessment processes are based on several steps, all conducted in comparison to human translation. Alignment is the first assessment step. It refers to the mapping between strings and pairs of words. It relies on the exact word match; if they are the same, two words include stem, synonyms, and even the phrase match. In other words, it measures the extent to which the MT software can transfer the content of human translation and achieve a high level of correlation (Lavie et al., 2004; Lavie & Agarwal, 2005). According to Mouratidis et al. (2020), METEOR assigns a good level score of correlation with human translation.

Second, METEOR rates the f-means and the fragmentation of words, i.e., the strongest match, by calculating the matched words and chunk numbers. It also computes the penalty for each alignment, thus dividing the unigrams matched sequence into small chunks, then adjacent these unigrams together and order them in an identical way in both strings. Consequently, the absolute penalty is calculated. Overall, METEOR parameter scores range from 0 to 0.5. If it includes more than one reference translation, it scores the generated text against each one separately and then uses the best pairing scores (Lavie & Agarwal, 2005).

4-2. MQM Framework

Attaining an objective evaluation of the quality of any MT output requires a precise tool with a high measuring level of performance. A study by White et al. (1994) affirms that human evaluation weighs many aspects of translation, such as accuracy, fluency, and adequacy. One can say that the MQM allows translators to measure all these aspects in the candidate's translation based on a specific error typology, making it more reliable compared to other methods. Overall, it can be utilized to judge the quality of a human, MT, or even AI-generated translation. On top of that, the MQM does not require a reference translation; instead, humans evaluate translation analytically and determine the typology of the issue type based on specific criteria related to the domain specifications.

MQM is used to analyze and measure translation quality products based on the typology of the issue type. It is a modern framework recommended by many researchers and experts. It was released in 2014 as a part of the QT LaunchPad group project funded by the EU. According to Mariana et al. (2015), the design of this framework depends on many quality evaluation metrics, more precisely on the LISA QA model. It can be applied to identify quality issues in the whole text as a holistic or specific error analysis. As explained on the MQM website, using it effectively requires evaluators to divide the original and target texts into smaller segments based on a sentence or chunk level and then align them. As

indicated earlier, the MQM offers analytical methods. It links errors with specific words and phrases in the text.

Lomel et al. (2013, 2015) consider MQM a flexible assessment method; it incorporates more than five stages: the primary step is a revision of the whole text, which is required to guarantee that the translation meets its specification. Based on this stage, evaluators can select or even create a suitable metric for performing translation assessment, then determine the specific value, i.e., the "Threshold Value" for each segment that helps to decide whether the outcomes meet the required criteria. Further, word count tools are evaluated. In the second stage, the evaluator judges the translation to the original text's specifications, where he identifies meanings and assigns types of errors, severity levels, and penalty points. The last stage is for calculating the overall scores, using Evaluation Word Count Software, and then comparing the translation outcome with the Threshold Value decided in stage two for evaluating the translation as successful or failing.

4-2.1. MQM Error Typology and Scoring Formula

Determining the logic error requires minimal effort with the MQM scale since its website <https://themqm.org> offers a list of possible errors categorized under seven dimensions, each including distinct subtypes. Overall, it defines over 100 issue types deriving from different quality assessment systems, whether they manually rely on reviewers' opinions or automatic error detection. All issue types are represented in a hierarchal framework. This maintains the granularity at any level. For instance, mistranslation is a subcategory of accuracy, and ambiguous translation is a subcategory of mistranslation, and so on. The other extremes are Verity and Fluency; they also include many subcategories represented in a hierarchal way, as Table (1) below illustrates.

The top-ranking issues are grouped based on the MQM website into several dimensions; for instance, accuracy, grammar, terminology, linguistics, verity, and others constitute the main dimensions. To make them clearer, short explanations and examples are provided (See Table 1). Lommel et al. (2013) argue that it is unexpected to use all the categories of

issue types rather than choose the most relevant issues to the assessment requirement.

Table (1) : MQM Error Categories

MQM Assessing Framework							
s.	Error Types name	The categories of error types					
1.	Terminology	Inconsistent with terminology resource (organizational and third-party terminology)	Inconsistent use of terminology (Multiple terms in ST or TT or abbreviation)			Wrong term (terminological collocation)	
2.	Accuracy	Mistranslation (Technical relation) (Ambiguous source content) (false friend, number, date, entity, overly literal)	Overt translation (MT heluation e.g. overt translation, addition, omission	Addition /Omission	Do not translate	Untranslated	Improper exact match
3.	Linguistic conventions (Fluency)	Grammar e.g., word forms, word order function words, agreement)	Punctuation e.g., whitespace, unpaired quote marks	Spelling Diacritics and transliteration	Cohesion	Coherence	Ambiguity
4.	Style	Organizational style	Legal style	Inconsistent with external reference	Language register e.g. Grammatical register	Awkward style	Unidiomatically style

Source: based on MQM website: <https://themqm.org>)

What makes the MQM framework more flexible is that, besides the classifications of errors, it offers a detailed explanation with examples for each type with numerical values. Lommel et al. (2013) MQM formula : " $TQ = 100 - AP (FPT - FP_s) - (VP_T - VP_s)$ " (p.6). Means: The overall Quality =100 - Penalties for Accuracy (Fluency Penalties for original text- the Fluency Penalties for the candidate translation)- (Verity Penalties for the candidate - Verity Penalties for the original)" (p.6). Another formula utilized to determine the severity of errors that affect meaning differently is: " $P = (Issues_{minor} + Issues_{critical} \times 10) \div Word\ count$ " (p.6).

What makes MQM different from METEOR is, as mentioned above,

that it is a manual method. Therefore, it requires more effort than the automatic metrics. This reflects a month of detailed data it provides about the obtained scores, and offers a guideline on how to improve them. On the contrary, automatic metrics provide just numbers and percentages of the scores without explaining the reason for which translation has obtained such scores, neither the way this translation is enhanced. In addition, automatic metrics always require a reference translation and work without actionable guidance on the output suitability.

5- Data Collection

The data set consists of 20 excerpts. Each sentence has a reference translation. METEOR is used to assess both “accuracy” and “fluency” based on the evaluation of precision and recall. The range scores start from (0) as the poorest grade and 1 as the highest. Further, the severity errors are assigned, then the METEOR absolute penalty and the overall quality average of Amazon production are computed.

Phase two includes MQM reliability evaluation by identifying the severity of errors and their effect on the entire quality of the text. This is followed by calculating the total penalty employing all the mentioned elements in assessing language errors at the level of: terminology, accuracy, fluency, and meaning. To provide a detailed discussion and reliable results, the researchers conducted the following steps to make a valid assessment and illustrate an accurate comparison:

5-1. The METEOR Assessment of the Entire Corpus

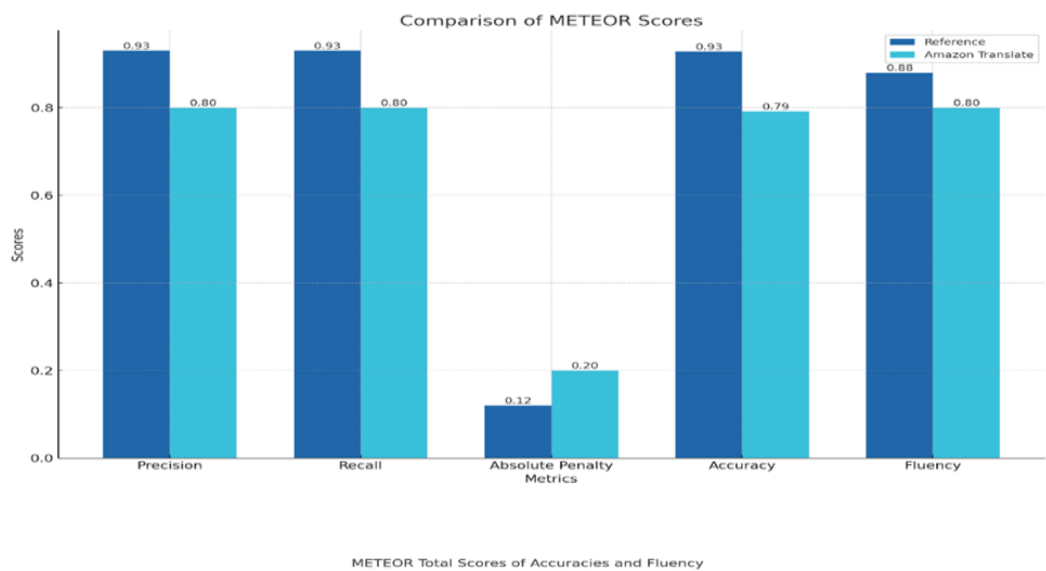
Table (2) shows that Amazon Translate's precision recorded 0.80 compared to the reference translation precision, which scored 0.93. This indicates that the METEOR scale tracks the precision of Amazon Translate well. Identical scores of correlations were recorded when appraising the capability of Amazon Translate to capture the reference translation content, i.e., in measuring the number of correct stems, words, and phrases reproduced by Amazon Translate, known as “recall”. The correlation coefficient of recall is 0.80 compared to reference translation, which is 0.93.

This also reflects the accuracy of METEOR in assessing how much of the content is reproduced compared to the reference translation.

Whereas METEOR penalty points assigned for the errors identified in Amazon Translate output are 0.20, which is not a higher rate compared to the reference translation penalty points, the degree to which Amazon Translate output conforms the standard translation reached 0.79 compared to reference translation accuracy scores, which is 0.93. This indicates that the assessment tool can identify the accuracy level for Amazon Translate output.

Furthermore, the researchers rate the candidate’s translation smoothness and naturalness, which includes grammatical correctness, word order, and synonyms-that is, the fluency of Amazon output. It scored 0.80 out of 1 for the entire document translation.

Table (2): METEOR Total Scores of Accuracies and Fluency

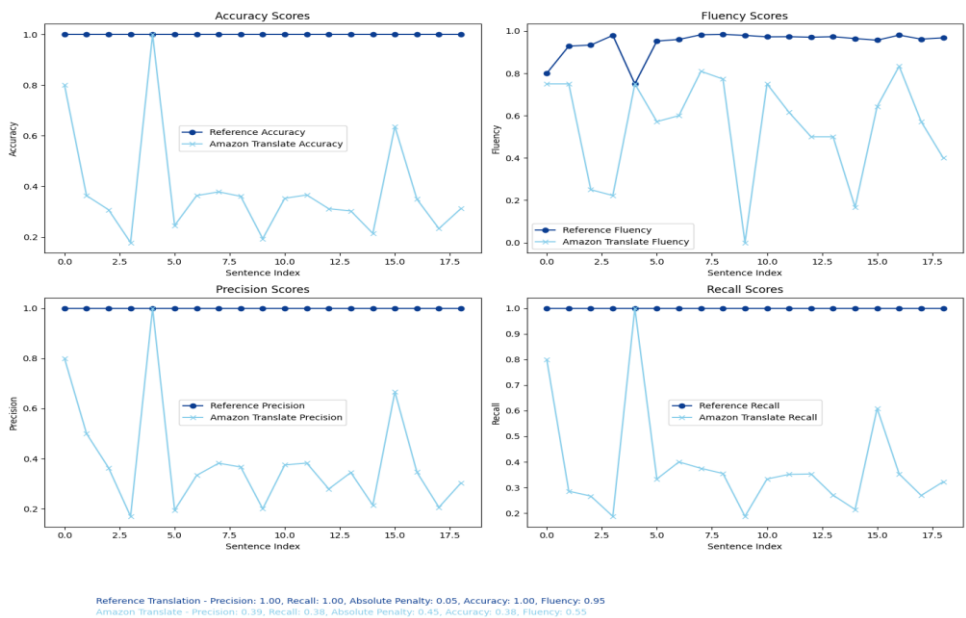


5-2. METEOR Assessment of the 19 Excerpts

Table (3) shows the METEOR assessment of Amazon’s translation of 19 legal excerpts. The recalled unigrams of Amazon Translate output scored 0.38, which can be considered low compared to the total scores for

human translation. The precision of the 19 Amazon translation extracts reached 0.39 out of 1.00, which can be considered as low. While the accuracy level recorded 0.38, these scores remained unchanged from the precision. Amazon's fluency level reached 0.55 compared to the fluency of the reference translation, scoring 1.00; one can posit that it is a remarkable rank. Conversely, the number of recalled unigrams in Amazon Translate output indicates low scores, i.e., 0.38, compared to the human translation scores, which obtained a total score of 1.00. The absolute penalty reached 0.45, higher than the 0.05 of the human translation.

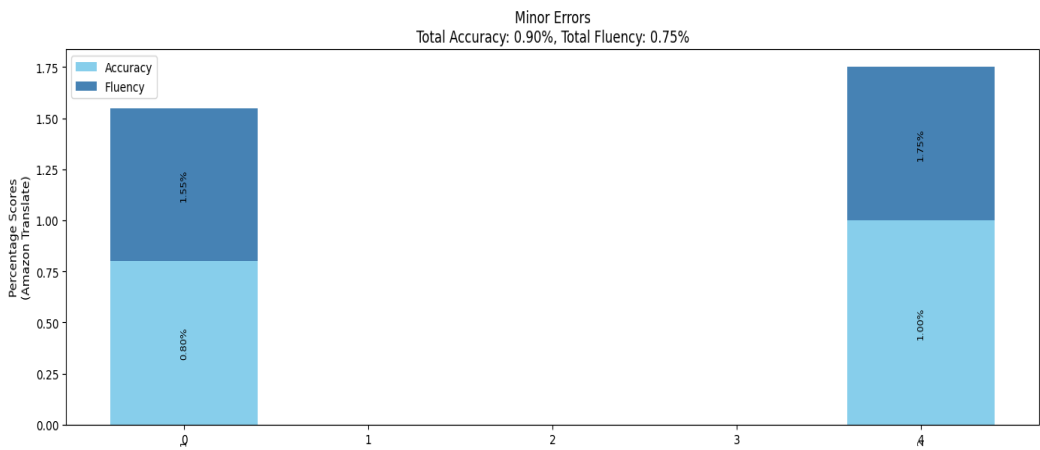
Table (3): METEOR Total Scores of Accuracies, Fluency Precision, and Recall Unigrams of the Chosen Sample



Tables (4,5,6) show METEOR classification of the issues type in the candidate's translation to measure its severity errors as minor, major, or critical. METEOR recorded two minor errors over 19 excerpts. Whereas the human translation reads:" *Maḥkamat al-Muqāṭa'ah al-qaḍā'iyah al-sādisahwa-al-'ishrūn*", Amazon Translate rendered it as:"*Maḥkamat al-Muqāṭa'ah al-qaḍā'iyah al-sādisahwa-al-'ishrūn*". Whereas METEOR ranked the candidate translation "لا شيء" at 0.80 because the error is minor,

compared to the human translation, which achieved a total score of 1.00. The accuracy of the minor errors reached 0.90%, while the fluency reached 0.75%. Therefore, the total METEOR score for these errors is (0.79%). Table (4) displays METEOR's total scores of minor errors.

Table (4): METEOR Total Scores of Minor Errors



On the other hand, the METEOR assessment revealed only one error as major, that is:

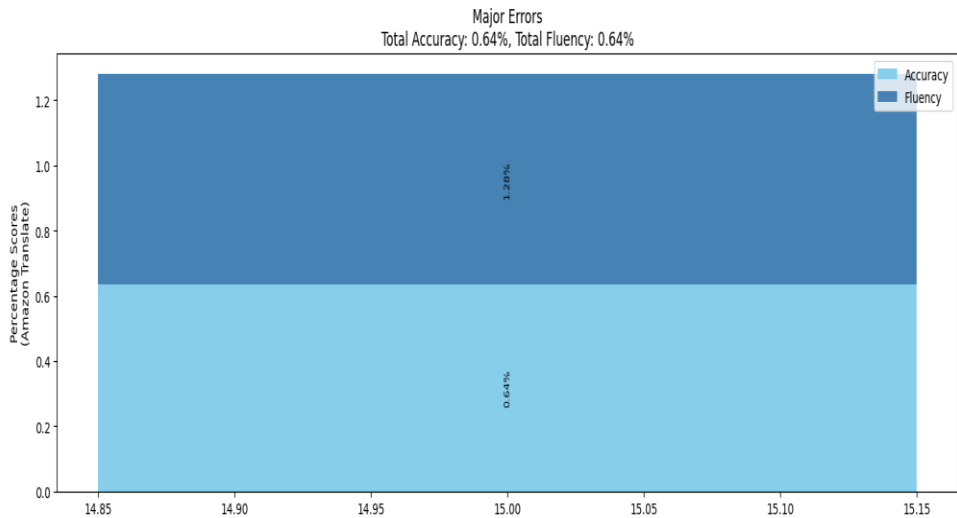
"Aqrra ltrfān b'nhmā qad thqqā min ṣiḥḥat ta'yīn al-'aqār wa-arqām ta'rīf al-markabāt"

Amazon Translation was

"Yshtrṭ altrfān al-Taḥaqquq min ṣiḥḥat al-awṣāf al-qānūnīyah wa-arqām ta'rīf al-sayyārah"

The accuracy recorded 0.64%, the fluency attained 1.28%, and the total fluency reached 0.64%. METEOR's total score for this major error reached 0.60%. All this is illustrated in Table (5) below.

Table (5): METEOR Total Scores of Major Errors



In contrast to minor and major errors, METEOR recorded 16 excerpts containing critical errors in Amazon Translate output. The METEOR level of accuracy of these excerpts ranges between 0.18 to 0.37. Notably, most accuracy scores range from 30 to 37. For example, while the reference translation reads:

"يخضع الحكم الحالي وقسمة الأملاك لقوانين ولاية لوزيانا، ويجب رفع أي دعوى ضرورية من أجل تنفيذ الأحكام المذكورة، أمام محكمة القسم القضائي السادس والعشرين، مقاطعة بوسير، لوزيانا، ويشكل هذا البند تنازلاً صريحاً عن أي اعتراض أو طعن حول الاختصاص الشخصي لمحكمة لوزيانا أو على تعيين أي محام لتمثيل الطرفين، عند الاقتضاء، وفقاً للمادة 5091 من قانون الإجراءات المدنية للولايات المتحدة لوزيانا."

"yakhḍa' aalhukm al-ḥālī waqismah al'mlāk liqwānyn wilāyḥ lūzyānā, wayjib raf' ayyi da'wā ḍarwrīh min ajl tanfīdh al'ḥkām almdhkwrh, amām maḥkamah alqsm alqdā'yyi alssāds wāl'shryn, muqāṭ'ah būsy, lūzyānā, wa yshkkil hdhā albnd tnāzłā šryḥan 'an ayyi i'tirāḍ aw ṭa'n ḥawl alākhtšāš aalshkhšī limḥkamah lūzyānā aw 'alā ta'yīn ayyi muḥām litmthīl alṭṭarfyni, 'ind alāqtḍā', wfqan lilmāddah 5091 min qānūn al'jrā'āt almdnyyah llūzyānā."

"Amazon's Translation was as follows:

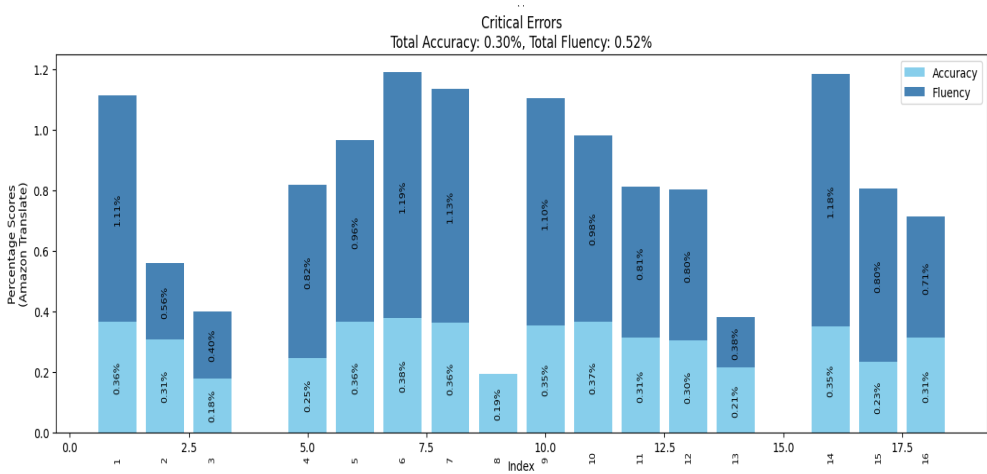
"يجب تفسير الحكم والتقسيم وحكمهما وفقاً لقوانين ولاية لوزيانا، ويجب اتخاذ أي إجراء قانوني ضروري لإنفاذ شروطه في محكمة المقاطعة القضائية السادسة والعشرين، بوسير باريش، لوزيانا، حيث يشكل هذا الحكم تنازلاً صريحاً عن أي اعتراضات أو استثناءات لممارسة الولاية القضائية الشخصية عليها من قبل محكمة لوزيانا أو على تعيين أي محام لتمثيلهم، إذا لزم الأمر، وفقاً لقانون الإجراءات المدنية في لوزيانا المادة 50 91؛"

" *Yajib tafsīr aalhukm wālttaqsym whkkinhmā wfqan liqwānyn wilāyh lūzyānā, wayjib ittikhādh ayyi ijrā' qānūny ḍarwrī li'nfādh shurwṭh fī maḥkamh al-Muqāṭa'ah alqḍā'yyah alssādsh wāl'shryn, būsyṛ bārysh, lūzyānā, ḥayth yushkkil hdhā aalhukm tnāzlā ṣryḥan 'an ayyi i'trādāt aw istthnā'āt limmārsh al-wilāyah alqḍā'yyah alshshakhṣyyah 'lyhā min qibl maḥkamh lūzyānā aw 'alā ta'yīn ayyi muḥām litmthylhm, idhā lazim al'mr, wfqan liqānwn al'jrā'āt almdnyyah fī lūzyānā almāddah 50 91 "*

In terms of accuracy, Amazon's output notably contains mistranslations, omissions, misinterpreting of technical concepts, literal translation, improper exact match, and untranslated concepts. Highlighting such errors in the above extract demonstrates the accuracy of the METEOR assessment since it achieved 0.37 out of 1.00, which indicates a low-quality level. On the other side, the critical accuracy errors are 0.30%. This means that Amazon's translation of these excerpts is also inferior.

The fluency ranges between 0.38 to 1.00, which is higher than the accuracy scores. Notably, 13 scored between 0.80 and 0.1, which can be considered higher correlation scores. Therefore, the total fluency score is 0.52, which means that the coherence of Amazon Translate is recorded at a close level to human translation. Table (6) displays METEOR's total scores of critical errors.

Table (6): METEOR Total Scores of Critical Errors



METEOR's total scores for overall Amazon output quality are calculated as follows: the total scores for precision are 7.42, the overall recall is 7.26, and the penalty is 8.54. The overall accuracy recorded is 7.27, while the overall fluency of Amazon's output is 10.45.

In the next step, the researchers evaluated Amazon Translate's output using MQM-Core Error Typology with 4 Severity Levels analysis. They selected 8 excerpts and performed a deep analysis of them based on MQM framework error categories, illustrated in Figure (1) above. Amazon data displays four dominating errors, which recorded a high-frequency rate (56%). These four types include 14 errors involving inconsistent terminology use, 20 errors in accuracy, mostly mistranslations, 12 errors grammatical errors, and 10 errors estimated as inconsistent style.

Table (8) shows that the top-ranking problems are grouped into four primary dimensions: accuracy, fluency, terminology, and style. It also computes the severity of each excerpt's errors.

The first extract contains a minor error in the expression "العشرين" and "العشرون" *"al-‘ishrūn" wa "al-‘ishrīn"*. Linguistically and grammatically, this error is related to the incorrect usage of Arabic compound numbers since formulating such numbers demands gender

agreement of the first part of the word. Amazon Translate effectively rendered the correct Arabic term.

On the contrary, in the second sentence, Amazon Translate produced a direct and simple translation that was almost literal, resulting in a critical error as it impacted the extract's intended meaning and coherence. Employing the Arabic connection to the conjunction (و)(*wa*) leads to changing the natural structure of the sentence to make it a list of general issues rather than depending on structures such as phrases to highlight the intended meaning. Furthermore, the human translation uses the preposition (with) to confirm that the petition includes the causes behind the divorce alongside the consequences of such a request. The translation generated by Amazon focuses on the divorce decision itself, besides some subsidiary or incidental issues.

Additionally, there is an inconsistent use of terminology in Amazon translation as (تحديد الأمور العرضية) (*Taḥdīd al-umūr al'rḍyh*). The language of law is accurate in communicating meanings using precise terms and concepts to transfer the legal context as the human translation does (التماس الفصل في تبعات الطلاق) (*il-Tamāss al-faṣl fī Tabi'āt al-ṭalāq*). The style is also a matter in this translation. Compared to reference translation, which is more formal and contains legal expressions suitable to judicial documents, Amazon's output shows a very simple and direct translation. In general, it displays an informal style that can be used in unofficial writing. METEOR also classifies this translation as low quality since it contains critical errors.

For the third excerpt, Amazon Translate gave an inferior-quality translation since it included wrong terms and structure, and the level of formality was also different. In Arabic, the verb (يلتمس) (*Yaltmis*) is employed for a formal situation, such as a legal request. On the contrary, the verb (يرغب) (*Yarghabu*) indicates an informal use. It refers to a personal desire rather than to a formal legal request.

Alongside, the translation of the phrase (Determination of incidental matters) as (الأمور العرضية) (*al-umūr al'rḍyh*) instead of (تبعات الطلاق) (*Tabi'āt al-ṭalāq*), which determines the divorce results, whether it has legal or

personal consequences, is another stylistic and semantic error. One can say that Amazon's translation traced the correct Arabic grammatical structure. However, the informal style of the verb used in Amazon's translation affects the accuracy and meaning of the legal translation.

Considering the "MQM" assessment, one concludes that the evaluation outcome of this extract also confirms the METEOR assessment of this sentence as a low-quality translation.

The fourth excerpt contains critical errors according to the MQM classification. These include errors in accuracy, fluency, inconsistent use of terminology, and sentence style. To illustrate more, although Amazon's output contains some specific structures, such as (ستستلم) (*Satstalim*) and (ستقوم بمنحها والمساومة عليها) (*Satqūm bimnḥihā wālmsāwmh 'lyhā*), it can be considered very simple and direct in comparison to the structure utilized in the reference translation, which contains more complicated structures with multiple actions such as (يتنازل) (*Ya tanāzal pātrik wynql wyby' wyṣrriḥ*) (باتريك وينقل ويبيع ويصرح). These actions are interconnected as a series covering legal content within a specific framework. Apart from that, the implicit meaning contains commands to perform this action in detail.

Further analysis showed that Amazon Translate's output contains critical linguistic structure, rules, and coherence errors. It uses a more complicated structure along with detailed expressions, such as سينقل الطرفان كل حقوقهما وملكهما ومصالحتهما في جميع الممتلكات) (*Sa ynql alṭṭarfān kulla ḥuqwwḥmā wmlkḥmā wmslḥthmā fī jamī' almmtlkāt*) as compared to (ينقل كل) (*Yanql kullu ṭaraf ilā alṭaraf al-ākhar*), resulting in incomprehensibility. Furthermore, the use of the sentence (باعتبارها ممتلكات) (*bi-I'tibārihāmmtlkāt munfaṣilatil-ṭaraf al-ākhar*) may increase ambiguity compared to reference translation, which employs a clear expression (لتكون ملكا له على النحو المبين أعلاه) (*li takun milkan lah 'alā al naḥw al-mubyn a'lāh*). This phrase clarifies the direct transfer of rights formally and specifically. Therefore, we can only classify them as critical. In this case, we confirm the METEOR assessment by classifying this translation as of inferior quality.

The same difficulties arise in the translation of sentence six. These include errors in Arabic grammatical structure, meaning, and style. This includes the use of the verb (سيضطر) (Syḍṭr) as an alternative for (يلتزم) (Yaltazim) (compelled) as a performative verb. The stronger tone in (Syḍṭr) implicitly carries the meaning of compulsion rather than commitment. It may also arise that the commitment will be performed by a third party other than the parties of the lawsuit. Another highlighted issue in this translation is the inconsistent use of terminology within the legal term base, language, and domain. Such issue types noted in Amazon translation justify and prove the METEOR classification of this excerpt as low quality with critical errors. All in all, this also reaffirms the pivotal role of automatic metrics in the MT quality assessment.

On the other hand, sentence seven includes complexity in its grammatical structure, attributed to the length of the sentences. In this example, Amazon Translate failed to grasp the implicit meaning of these phrases (لا يمكن أن يفسر التزام الطرفين... الخ) (*Lāyumkinu an yufassir iltizām al-Ṭarafayn ... ilakh*). It needs to set explicit links between them and the whole sentence. Besides, it includes an "Over translation" of the concepts (may alienate, mortgage) as (التصرف أو الرهن أو رهن أو رهن كل أو أي جزء من الممتلكات) (*Alttasarruf aw alrrahn aw rahn aw rahn kulli aw ayyi juz' min almmtlkāt*) compared to the reference translation (أن ينقل أو يرهن مجمل الملكية) (*An yanqul aw yarhan mujmal almlkyyah*). In relation to the mentioned critical errors, there is no doubt that METEOR, as an automatic metric, appropriately placed it as of inferior quality.

Table (7): Errors' Typology

S.	Reference Translation	Amazon Translate	MQM Issue types				
			Terminology	Accuracy	Fluency	Style	Severity
1.	محكمة المقاطعة القضائية السادسة والعشرون <i>Maḥkamat al-Muqāṭa‘ah al-qaḍā‘īyah al-sādisah wa-al-‘ishrun</i>	محكمة المقاطعة القضائية السادسة والعشرين <i>Maḥkamat al-Muqāṭa‘ah al-qaḍā‘īyah al-sādisah wa-al-‘ishrīn</i>	None	Improper exact match	العشرون ('ishrun) Grammar		Minor
2.	عريضة افتتاحية لدعوى طلاق طبقاً لأحكام المادة 102 مع التماس الفصل في تبعات الطلاق <i>‘Ariḍah iftitāḥīyah li-da‘wā Ṭalāq ṭibqan li-aḥkām al-māddah 102 wa-taḥdīd al-umūr al-‘arṣiyyah</i>	التماس الطلاق بموجب المادة 102 وتحديد الأمور العريضة <i>Iltimās al-ṭalāq bi-mūjib al-māddah 102 wa-taḥdīd al-umūr al-‘arḍiyh</i>	Inconsistent use of terminology and the termbase	Mistranslation of technical concepts	Grammar: The structure	Informal (missing legal style)	Critical
3.	8. حيث يلتزم المدعي أن يفصل الحكم الذي سيصدر في الدعوى الحالية في تبعات الطلاق. <i>8. ḥaythu yltms al-mudda‘ī an yfsl al-ḥukm alladhī sayashdur fī al-da‘wā al-ḥālīyah fī Ṭabi‘āt al-ṭalāq.</i>	8. يرغب المدعي في تحديد الأمور العريضة التالية في هذه الإجراءات.. <i>8. yarghabu al-mudda‘ī fī taḥdīd al-umūr al-‘arḍiyh al-tālīyah fī Hādhīhi al-ijrā‘āt.</i>	Inconsistent of legal domain	Improper exact match	Grammar the suitability of Verbs	Informal	Critical
4.	1. لا شيء <i>1. lāShay’</i>	1. لا شيء <i>1. lā Shay’.</i>	None	None	Punctuation	None	Minor

5.	هم ينقل كل طرف إلى الطرف الآخر، كافة الحقوق وسندات الملكية ومنافعها، المعنية أعلاه، لتكون ملكا له على النحو المبين أعلاه، <i>h. Yanqul kull taraf ilā altaraf al- ākhar, kāffat al-Huqūq wa Sanadāt al- Malakīyah wa mnāf'hā, almu'ynh a' lā, ltkwn malikan la-hu 'alā al- nahw al-mubīn a' lā</i>	ها سينقل الطرفان كل حقوقهما وملكتهما ومصلحتهما في جميع الممتلكات الموصوفة أعلاه إلى بعضهما البعض، باعتبارها ممتلكات منفصلة للطرف الآخر، على النحو المبين أعلاه، مع إيلاء الاعتبار الكامل والكافي، الذي حصل عليه كل منهما الآخر؛ <i>h. Saynqul alṭarfān kulla ḥuqwqhmā wmlkḥmā wmşlḥthmā fī jamī' almmtlkāt almwşwşh a' lāh ilā ba' ḏhimā alb' ḏ, biā' tbārḥā mumtlkāt munḥslah lilṭarf al- ākhar, 'alā an-naḥw al- mubyn a' lāh, Ma' a ilā' alā' tbār al- kāmil wālkāfy, aaldhī ḥaṣala 'lyh kulli minḥmā min al- ākhar</i>	Inconsistent use of terminology	Improper exact match	Grammar Structure, spelling, coherence	Informal	Critical
----	---	--	---------------------------------------	-------------------------	---	----------	----------

6.	يلتزم كل طرف بالتوقيع على كافة التحويلات، أو الصكوك، أو الشهادات، أو فواتير البيع، أو غيرها من الوثائق اللازمة لتنفيذ أهداف هذا الاتفاق، بطلب منهما أو من الغير. <i>Yaltazim kull taraf bil tawqy' 'alá kaffat althwylāt, aw al-ṣukūk, aw al-shahādāt, aw fwāṭyr al-bay', aw ghayrihā min al-wathā'iq al-lāzimah li-tanfīdh Ahdāf Hādha al-ittifāq, bi-ṭalab minhumā aw min al-ghayr.</i>	سيضطر كل من الطرفين إلى التوقيع على أي وجميع التحويلات أو الأدوات أو الشهادات أو فواتير البيع أو غيرها من الأدوات اللازمة لتنفيذ مقاصد وأغراض هذه الاتفاقية على النحو المطلوب من قبلهم أو من قبل أي طرف ثالث. <i>Sayaḍṭar kull min al-Ṭarafayn ilá al-tawqī' 'alá Ayy wa-jamī' al-taḥwylāt aw al-adawāt aw al-shahādāt aw fawāṭyr al-bay' aw ghayrihā min al-adawāt al-lāzimah li-tanfīdh Maqāṣid w'ghrāḍ Hādhihi al-Ittifāqīyah 'alá al-naḥw al-maṭlūb min qblhm aw min qabla Ayy taraf thālith.</i>	Termbase, legal domain, legal language	Improper exact match	Grammar the use of the performative verb	Informal Tone	Critical
----	---	---	--	----------------------	--	---------------	----------

S	Reference Translation	Amazon Translate	MQM Issue types				
			Terminology	Accuracy	Fluency	Style	Severity
7.	ح. لا يمكن أن يفسر التزام الطرفين بدفع الديون المذكورة في اتفاق قسمة الأملاك المشتركة، بموجب الاتفاق الحالي، على أن هذه القسمة خاضعة ليند فاسخ صريح أو ضمني، بحيث أنه يسوغ لكل واحد من الطرفين أن ينقل أو يرهن مجمل الملكية التي يتحصل عليها نتيجة لهذه القسمة أو جزء منها. وفي هذا الصدد، قرر الطرفان أنهما يتنازلان عن الاستفادة من أي بند فاسخ، كان صريحا أو ضمنيا، أو امتياز البائع. <i>Lā yumkinu an yufassir iltizām al-Ṭarafayn bi-daf' al-duyūn al-madhkūrah fī Ittijāq qismat al- amlāk al- mushtarakah, bi-mūjib al- ittijāq al-ḥālī, 'alā an Hādhihi al- qismah khāḍ'<h>h lbnd fāskh ṣarīḥ aw ḍmny, Bḥayth annahu yasūgh li-kull wāḥid min al- Ṭarafayn an yunqalu aw yrhn Mujmal al-Malakīyah allatī yithṣl 'alayhā Naṭījat li-hādhihi al- qismah aw Juz' minhā. wa fī Hādḥā alṣdd, qrr</h></i>	ح. لا ينبغي تفسير أي متطلبات للأطراف هنا لدفع أي مديونية مذكورة في تقسيم ممتلكات المجتمع هذا بحيث يمكن أن يخضع هذا التقسيم لشرط حل صريح أو ضمني، ويقرر بشكل خاص أنه يجوز لأي من الطرفين التصرف أو الرهن أو رهن أو رهن كل أو أي جزء من الممتلكات التي حصل عليها هو أو هي في قسم ممتلكات المجتمع هذا. فيما يتعلق بذلك، يقرض على الأطراف التنازل بشكل خاص عن الاستفادة من أي شرط حل صريح أو ضمني، أو أي امتياز صريح أو ضمني للبيع. <i>H. lā yanbaghī tafsīr Ayy Mutaṭallabāt ll'ṭrāf hunā li- daf' Ayy mdywnyh mdhkwrh fī taqsīm mumtalkāt al- mujtama' Hādhā bḥyṭh yumkinu an yakhḍa' Hādhā al-taqsīm lshrt ḥall ṣarīḥ aw ḍmny, Wyuqrr bi-shakl khāṣṣ annahu yajūz li-ayy min al- Ṭarafayn al- taṣarruf aw al- rahn aw rahn aw rahn kull aw Ayy Juz' min al- mumtalakāt allatī ḥaṣala</i>	Inconsistent use of terminology	overt translation	Grammar	Informal Tone	Critical

	<i>alṭrfān anhmā ytnāzlān ‘an al-istifādah min Ayy Band fāskh, kāna ṣarīhan aw ḍmnyā, aw Imtiyāz albā’</i>	<i>‘alayhā huwa aw hiya fī Qism mmtlkāt al-mujtama’ Hādhā. fīmā yata‘allaqu bi- dhālik, yufrd ‘alā al-aṭraf al-Tanāzul bi- shakl khāṣṣ ‘an al-istifādah min Ayy sharṭ ḥall ṣarīḥ aw ḍmny aw Ayy Imtiyāz ṣarīḥ aw ḍmny lilbā’i’.</i>					
8.	أقر الطرفان بأنهما قد تحققا من صحة تعيين العقار وأرقام تعريف المركبات الواردة في هذه العريضة والحكم وجميع المستندات الأخرى ذات الصلة. <i>Aqara alṭrfān b’nhmā qad thqqā min ṣiḥḥat ta’yīn al-‘aqār wa- arqām ta’rīf al-markabāt al-wāridah fī Hādhīhi al- ‘arīdah wa-al- ḥikam wa- jamī’ al- Mustanadāt al- Ukhrā Dhāt alṣlh.</i>	يشترط الطرفان التحقق من صحة الأوصاف القانونية وأرقام تعريف السيارة الواردة في هذا الالتماس والحكم وجميع المستندات الأخرى ذات الصلة. <i>Yshtrṭ alṭrfān al-Taḥaqquq min ṣiḥḥat al- awṣāf al- qānūniyah wa- arqām ta’rīf al-sayyarah al- wāridah fī Hādhā alāltmās wa- al-ḥikam wa- jamī’ al- Mustanadāt al- ukhrā Dhāt al- ṣilah.</i>	Inconsistent use of legal terms	Improper exact match Literal translation	Grammar performative verbs	Informal	Major

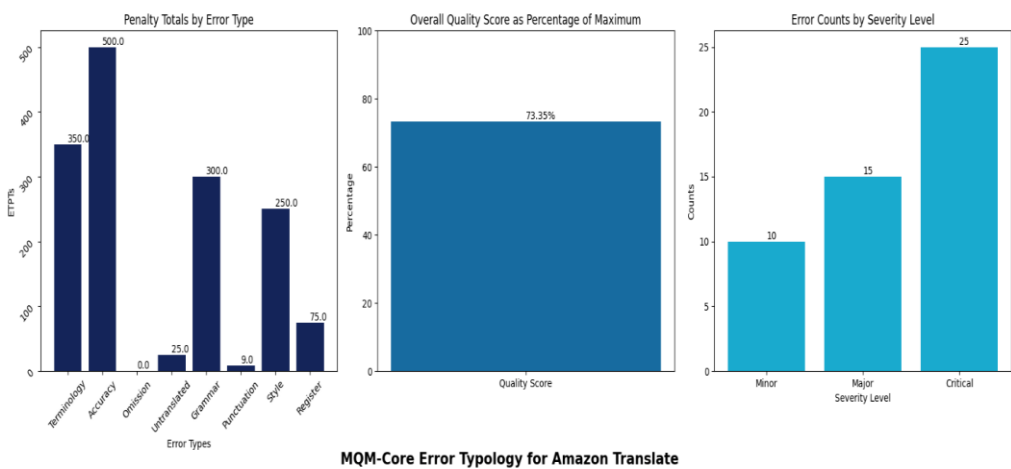
Amazon’s translation of the last excerpts shows structure, meaning, and terminology errors. Although these cannot be considered critical, the translation is not of good quality.

In essence, the misusein (alṭrfānyshtrṭ)changes the meaning of the legal concept (The parties stipulate). Actually, the original style of this excerpt utilizes an assertive style, in which the performative verb(aqrralṭrfān)explicitly expresses the action completion.

Table (7) shows the results from the analysis constructed from "MQM Core Typology". The assessment revealed that the penalty scores of

2714.00 distributed asterminology use, wrong terms, and inconsistent terminology resources reached 350.00. The accuracy level, i.e., mistranslations, over-translations, under-translations, omissions, and untranslated words, phrases, or sentences, reached 500.0. The percentage of untranslated words was 25.0, while the grammatical errors rate reached 300.0, with style errors reaching 250.0. Besides, tone and register errors reached 75.0,while the "Overall Quality Score" (OQS) of the selected extracts reached 73.35. The percentage of minor errors reached 10% compared to the critical errors at 25%, whereas 15% were considered major errors out of the total. Table (8) illustrates all of these details.

Table (8): MQM-Core Error Typology for Amazon Translate Output



6- Conclusion

The assessment experiment revealed that the MQM shows a detailed description of the issue types and the severity levels for each score, which is consistent with the hypotheses of this research that the MQM framework is more precise than the automatic assessment tool. The framework used here gave the best results in assessing MT performance in legal translation. Hence, measuring the quality is more convincing, satisfying, and reliable when relying on the MQM framework.

Even so, METEOR's performance was similar in classifying the translation as inferior or high quality. Its assessment was largely accurate

since it was based on categorizing the severity of errors as natural, low, minor, major, or critical. What makes its performance satisfying is its ability to measure the following aspects: “there call, precision, and the total of the absolute penalty of the translation”. Actually, using METEOR depends on what evaluators want to assess and can be adapted to achieve better results since the MQM framework revealed that Amazon Translate output has low quality, and the same is true for METEOR.

7- References

- Agarwal, A., & Lavie, A. (2008). *Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output*. pp. 115–118.
- Banerjee, S., & Lavie, A. (n.d.). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*.
- Depraetere, I. (2011). *Perspectives on Translation Quality*.
- Dobre, I. (2015). A Comparison Between BLEU and METEOR Metrics Used for Assessing Students within an Informatics Discipline Course. *Procedia - Social and Behavioral Sciences*, 180, pp. 305–312. <https://doi.org/10.1016/j.sbspro.2015.02.121>
- Lavie, A., & Agarwal, A. (2005). *Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*.
- Lavie, A., Sagae, K., & Jayaraman, S. (2004). *The Significance of Recall in Automatic Metrics for MT Evaluation*.
- Lommel Arle & Melby Alan. (2015). *Assessing Translation Quality with Multidimensional Quality Metrics (MQM) Task 2: Select Dimensions*. <http://www.qt21.eu/launchpad/content/multidimensional-quality-metricsoremailinfo@qt21.eu>
- Lommel, A. (n.d.). *Multidimensional Quality Metrics (MQM): A New Framework for Translation Quality Assessment Arle Lommel (DFKI)*.
- Lommel, A. R., Burchardt, A., Uszkoreit, H., Berlin, D., & De, H. U. (n.d.). *Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality*. <http://www.qt21.eu/launchpad>

- Lommel, A. R., Burchardt, A., Uszkoreit, H., Berlin, D., & De, H. U. (2013). Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality. *Proceedings of Translating and the Computer* 35. <http://www.qt21.eu/launchpad>
- Mariana, V., Cox, T., & Melby, A. (2015). The Multidimensional Quality Metrics (MQM) Framework: a new framework for translation quality assessment. In *The Journal of Specialised Translation Issue* (Vol. 23). <http://www.qt21.eu/mqm-definition/>
- MQM website : <https://themqm.org>
- Mouratidis, D., Kermanidis, K. L., & Sosoni, V. (2020). Innovative Deep Neural Network Fusion for Pairwise Translation Evaluation. *IFIP Advances in Information and Communication Technology*, 584 IFIP, 76–87. https://doi.org/10.1007/978-3-030-49186-4_7
- White, J. S., O'Connell, T., & O'mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *The First Conference of the Association for Machine Translation in the Americas*, pp. 1–25.