CAN WE TRUST MACHINES? A CRITICAL LOOK AT SOME MACHINE TRANSLATION EVALUATION METRICS

Muhammad Zayyanu Zaki [1] and Nazir Ibrahim Abbas [2]

[1] Department of French, Usmanu Danfodiyo University, Sokoto, Nigeria
[2] Department of Nigerian Languages, Usmanu Danfodiyo University, Sokoto, Nigeria

ABSTRACT

The growing interconnection of the globalised world necessitates seamless cross-lingual communication, making Machine Translation (MT) a crucial tool for bridging communication gaps. In this research, the authors have critically evaluated two prominent Machine Translation Evaluation (MTE) metrics: BLEU and METEOR, examining their strengths, weaknesses, and limitations in assessing translation quality, focusing on Hausa-French and Hausa-English translation of some selected proverbs. The authors compared automated metric scores with human judgments of machine-translated text from Google Translate (GT) software. The analysis explores how well BLEU and METEOR capture the nuances of meaning, particularly with culturally bounded expressions of Hausa proverbs, which often have meaning and philosophy. By analysing the performance of the translator's datasets, they aim to provide a comprehensive overview of the utility of these metrics in Machine Translation (MT) system development research. The authors examined the relationship between automated metrics and human evaluations, identifying where these metrics may be lacking. Their work contributes to a deeper understanding of the challenges of Machine Translation Evaluation (MTE) and suggests potential future directions for creating more robust and reliable evaluation methods. The authors have explored the reasons behind human evaluation of MT quality examining its relationship with automated metrics and its importance in enhancing MT systems. They have analysed the performance of GT as a prominent MT system in translating Hausa proverbs, highlighting the challenges in capturing the language's cultural and contextual nuances.

KEYWORDS

Machine Translation, Quality Metrics, Evaluation, BLEU, METEOR

1. INTRODUCTION

The increasing interconnectedness of the globalised world has created an unprecedented demand for seamless cross-lingual communication. As Sajo et al. (71) emphasise, translation is vital for effective communication across diverse cultures. This surge in cross-lingual interaction has propelled Machine Translation (MT) to the forefront of technological innovation. MT plays a crucial role in bridging communication gaps from facilitating international business transactions and fostering cross-cultural understanding to providing access to information and knowledge across linguistic barriers. Its applications are diverse, ranging from instant translation of web pages and social media content to powering multilingual customer support systems and enabling real-time interpretation in international conferences. Consequently, the development and refinement of robust and accurate MT systems have become paramount, driving significant research and investment in the field. This is further evidenced by the development of AI-powered

translation systems, as noted by Zaki et al. (23), which integrate Artificial Intelligence (AI) technology to enhance both translation efficiency and accuracy.

Additionally, the growing importance of MT is not merely a reflection of increasing globalisation, but also a catalyst for it. MT fosters greater collaboration and exchange on a global scale by lowering the barriers to communication. It empowers individuals and organisations to connect with wider audiences, access diverse perspectives, and participate in international discourse, irrespective of their native language. Furthermore, the advancements in MT technology, particularly with the advent of deep learning, have led to significant improvements in translation quality, making it an increasingly viable and reliable tool. This continuous improvement further reinforces the significance of MT, making it an indispensable component of an interconnected world and a key enabler of future global interactions.

Evaluating the quality of MT presents a complex and multifaceted challenge. As Zaki (24) explains that Machine Translation, a branch of Computational Linguistics (CP) or Natural Language Processing (NLP), focuses on using software for cross-lingual text conversion. Unlike evaluating computational tasks with clearly defined correct answers, translation quality is inherently subjective and depends on factors like the intended audience, the translation's purpose, and specific linguistic nuances. Human judgments, while considered the gold standard, are time-consuming, expensive, and susceptible to inconsistencies due to individual biases and varying interpretations of “good” translation. Furthermore, capturing the full spectrum of translation quality - encompassing fluency, adequacy, accuracy, and style, requires nuanced evaluation metrics that move beyond simple word-for-word comparisons. This inherent subjectivity and multi-dimensionality make establishing universally agreed-upon benchmarks and objective measures for MTE particularly difficult.

This research focuses on automated metrics, specifically BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit Ordering). As Triveni et al. (2012) explain that automated tools offer rapid and reliable feedback at a relatively low cost. These automated quality metrics provide a numerical score representing the quality of the translation. In contrast, human evaluation metrics, while valuable, assess qualities like fluency (how natural and fluent the translated text sounds), adequacy (how well the meaning is conveyed), coherence (how well the text flows as a cohesive whole), and readability (how easy the text is to read and understand).

Beyond the inherent subjectivity, several technical challenges complicate MT evaluation. The dynamic nature of language, with its constant evolution and contextual variations, poses a significant hurdle. MT systems often struggle with idioms, proverbs, metaphors, and culturally specific expressions, making it difficult to assess the true meaning conveyed in the translated text. This is why the researcher tries to explain the meaning and philosophy of each selected proverb in the analysis. Since there is this struggle, it becomes difficult to trust MT. The hurdle is highly achieved by a human translator as opposed to a machine due to his unique nature as a human understanding of natural language and pragmatics.

Additionally, the lack of parallel corpora for many language pairs (Hausa-French or Hausa-English) limits the training and evaluation of MT systems, further hindering the development of robust evaluation metrics. The diverse linguistic structures and grammatical rules across languages also make it challenging to create a one-size-fits-all evaluation approach. Consequently, the field of MTE is constantly evolving, with researchers exploring new methods and metrics to better capture the nuances of translation quality and address these ongoing challenges.

Among the various metrics developed to assess Machine Translation Quality (MTQ), BLEU and METEOR have emerged as two of the most widely used and influential metrics. On one hand, BLEU, introduced by Papineni et al. (2002), revolutionised Machine Translation Evaluation by offering an automatic, objective, and relatively inexpensive method. It operates by comparing the N-gram (a contiguous sequence of "n" items (characters or words) from a given sequence of text or speech used in analysis) overlap between the machine-generated translation and one or more human reference translations. While simple in its approach, BLEU's ability to correlate reasonably well with human judgments, especially at the corpus level, has made it a staple in MT research and development. Its widespread adoption has facilitated comparisons across different MT systems and allowed for rapid progress in the field.

METEOR, developed by Banerjee and Lavie (2005), attempts to address some of BLEU's limitations by incorporating stemming, synonymy matching, and word order considerations. It aims to capture aspects of translation quality beyond simple n-gram overlap, rewarding translations that use paraphrases or synonyms and penalising those with significant word order deviations through explicit word alignment between candidate and reference translations. Although computationally more complex than BLEU, METEOR's enhanced ability to capture semantic similarity and word order information makes it a valuable metric, often used in conjunction with BLEU for a more comprehensive MTQ assessment. Both BLEU and METEOR, despite their strengths and weaknesses, remain central to the evaluation and advancement of Machine Translation Technology (MTT). As Zaki (180) notes, MTT could be used on various platforms, including computers and other devices, reducing translation time and ensuring contextually appropriate output.

This research focuses on providing a comprehensive overview of two prominent MTE metrics: BLEU and METEOR. It investigates the mechanics of each metric, explaining how they function and what aspects of translation quality they aim to capture. The research further analyses the strengths and weaknesses of both metrics, discussing their advantages and limitations in assessing different types of translations across various language pairs. Finally, it explores the potential applications of these metrics, highlighting their role in MT system development, research, and benchmarking, ultimately aiming to provide a clear understanding of their utility and impact in the field of MTE.

1.1. Background on Machine Translation Evaluation (MTE)

The evaluation of MT has been a central concern since the inception of the field. Early approaches relied heavily on human evaluation, which, while providing valuable insights, proved to be resource-intensive, time-consuming, and susceptible to subjective biases. As MT systems grew in complexity and output quality, the need for more efficient and objective evaluation methods became increasingly apparent. This led to the development of a range of automated metrics, aiming to approximate human judgments while offering greater speed and consistency. The initial focus was often on simple metrics based on word matching, but these proved inadequate in capturing the nuances of translation quality, such as fluency, adequacy, and meaning preservation.

The evolution of MTE metrics reflects the ongoing quest to better capture the multifaceted nature of "good" translation. Researchers have explored various approaches, from n-gram overlap and edit distance to more sophisticated methods incorporating syntactic parsing, semantic similarity, and even machine learning techniques. This pursuit has been driven by the increasing demand for high-quality translations across a wide range of applications, from simple document translation to complex tasks like cross-lingual information retrieval and dialogue systems. The development

and refinement of MTE metrics are crucial not only for benchmarking different MT systems but also for guiding research directions and fostering advancements in the field.

The objectives of the research are to:

1. Critically evaluate the effectiveness of BLEU and METEOR metrics in assessing machine translation quality.
2. Identify the strengths and weaknesses of BLEU and METEOR and explore potential alternatives and complementary evaluation metrics.
3. Investigate the feasibility of developing more robust and reliable evaluation methods for MT systems.

The research questions are:

1. To what extent do BLEU and METEOR metrics accurately reflect the quality of machine translation outputs?
2. What are the limitations and biases of BLEU and METEOR in evaluating MT systems?
3. Can alternative evaluation metrics be developed to improve the reliability and validity of MTE?

The significance of the research is to contribute to a deeper understanding of how human evaluators assess MTQ and how their evaluations relate to automated metrics. The research could help to improve the development of more accurate and reliable evaluation metrics, enabling translators to better assess the quality of MT outputs. The research could help translators streamline their evaluation processes, reducing the time and effort required to review and edit MT outputs. The translators could gain insights into how to effectively collaborate with MT systems, leveraging their strengths while mitigating their weaknesses. The research's findings could inform the development of training programmes and resources for translators, helping them stay up-to-date with the latest development in MTE and optimisation.

The research of human evaluation of MTQ is still in its infancy, and more research is needed to understand how these machines work, how humans evaluate translation quality and how these relate to automated metrics. It is based on this that the researcher tries to fill in and contribute knowledge to the understanding and efficacy of these machines and their limitations. The research is limited to the quality translation outputs of BLEU and METEOR using Google Translate (GT) software.

2. LITERATURE REVIEW

The literature on MTE is extensive and reflects the ongoing challenges in quantifying translation quality. Early work often focused on human evaluation, recognising its importance in capturing the nuances of meaning and fluency, Hutchins and Somers, (1992). However, the practical limitations of human evaluation, including cost and time constraints, spurred the development of automated metrics. One of the most influential early metrics was BLEU Papineni et al. (2002), which introduced the concept of n-gram overlap with reference translations. BLEU's simplicity and ease of computation contributed to its widespread adoption, though it was quickly recognised that it had limitations, particularly in capturing semantic similarity and handling paraphrasing, Callison-Burch et al. (2006).

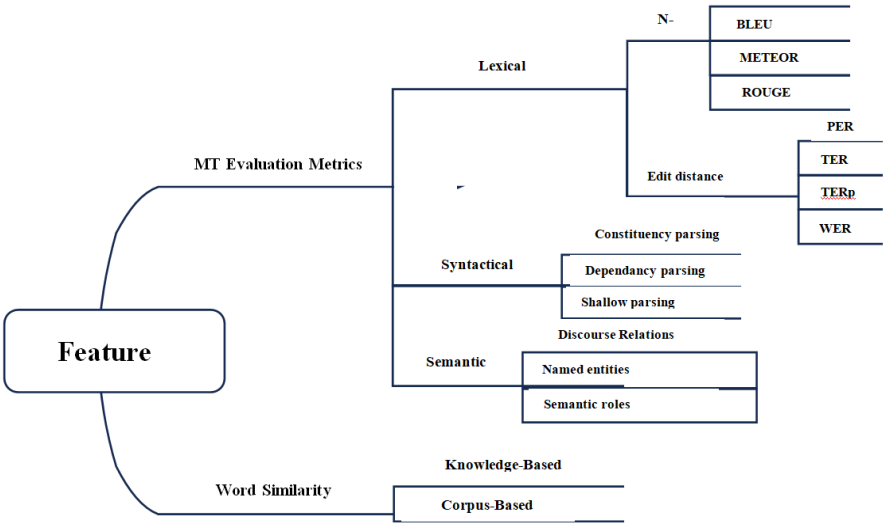
Researchers explored various avenues to address its shortcomings following BLEU's introduction. METEOR was developed to incorporate stemming and synonymy matching, aiming

to capture the meaning of semantic relationships between words, Banerjee et al. (2005). Later iterations of METEOR further refined its alignment strategies and incorporated more sophisticated matching techniques, Banerjee et al. (2005). According to Snover et al. (2006), other metrics, such as Translation Edit Rate (TER) focused on edit distance to measure the number of changes required to transform an MT into a reference translation. More recently, research has explored metrics based on syntactic analysis Owczarzak et al. (2010), semantic role labeling, Lo et al. (2012), and even neural network-based approaches, Zhao et al. (2019). This ongoing evolution of MTE metrics highlights the continuous pursuit of more accurate and comprehensive ways to assess the quality of MT.

The trajectory of MTE research reveals a clear progression from simpler, computationally efficient metrics to more complex, linguistically informed approaches. BLEU metric, as groundbreaking in their time, primarily focused on surface-level features like n-gram overlap, often failing to capture the nuances of meaning and fluency. This limitation spurred the development of more sophisticated metrics like METEOR, which incorporated stemming and synonymy matching to address some of these shortcomings. As discussed above, these advancements reflect a growing understanding of the multifaceted nature of translation quality and the need for metrics that go beyond simple lexical matching. The continued development of metrics based on syntactic analysis, semantic role labeling, and Neural Networks (NNs) demonstrates the ongoing pursuit of more accurate and comprehensive evaluation methods.

As Zaki (17) explains that the development of transformer models represents a pivotal milestone in the evolution of MT, significantly enhancing both translation quality and efficiency. This progress is not solely focused on improving correlation with human judgments; it also involves developing metrics that offer more diagnostic information about the strengths and weaknesses of different MT systems. While BLEU and METEOR provide a general quality assessment, they often lack the granularity to identify specific errors or areas needing improvement. Consequently, recent research aims to develop metrics capable of pinpointing specific linguistic phenomena that challenge MT systems, such as idioms, metaphors, or complex syntactic structures. The schema below explains MTE metrics thus:

Schema: Machine Translation Evaluation Metrics



Source: ACL Anthology, 2025.

The schema above categorises translation evaluation metrics. It distinguishes between “MTE Metrics”, which directly assesses the quality of the machine-translated text, and “Word Similarity Evaluation Metrics”, which focuses on comparing the similarity between words or phrases. MTE Metrics are further classified by the type of knowledge they use (knowledge-based or corpus-based) and the linguistic level they analyse (semantic, syntactical, or lexical). Examples of these metrics include BLEU, METEOR and TER. Word similarity evaluation metrics are categorised by the specific linguistic feature they evaluate, such as N-grams or Named Entities. This structure provides a framework for understanding the different approaches to evaluating MT.

3. COMPARISON BETWEEN BLEU AND METEOR METRICS

This research compares two widely used MTE metrics: BLEU and METEOR. While both are common in the field, they differ significantly in their approach and the aspects of translation quality they prioritise. As Zaki (16) notes that the objective of such comparisons is to explore the intricacies of these metrics and analyse their respective strengths, weaknesses, and limitations within the context of translation technology. A key difference lies in lexical matching. BLEU relies on n-gram overlap, counting matching word sequences. METEOR, however, employs a more sophisticated alignment process, considering exact, stemmed, and synonym matches using resources like WordNet (a large lexical database of English words created by George Miller and his team at Princeton University in the 1980s). As Triveni et al. (2012) confirm, METEOR incorporates features like stem and synonym matching. This allows METEOR to capture semantic similarity and recognise paraphrases, a crucial aspect often missed by BLEU. Another key difference is their sensitivity to word order. While BLEU implicitly considers word order through n-grams, METEOR explicitly incorporates it into its scoring function, penalising deviations from the reference translation's word order. This makes METEOR more sensitive to grammatical errors and fluency issues.

Despite these differences, BLEU and METEOR share some similarities as well. Both are automatic metrics, designed to provide objective and efficient evaluation of MT output. Both rely on reference translations, comparing the candidate translation (candidate translations are generated by the system and then evaluated to determine the best translation from the source text into a target language) against one or more human-generated translations. Both also calculate scores based on a combination of precision and recall, though they use different formulas and weighting schemes. Furthermore, both metrics, despite their advancements over simpler word-matching techniques, still struggle to fully capture the nuances of human language understanding and could sometimes be gamed by systems that optimise for the metric scores rather than overall translation quality. They are, however, valuable tools when used thoughtfully and in conjunction with other evaluation methods.

The choice between BLEU and METEOR depends on the specific evaluation context and the priorities of the research. BLEU's simplicity and speed make it suitable for rapid prototyping and large-scale comparisons of different MT systems, especially when computational resources are limited. It is also useful for tracking progress over time and identifying general trends in translation quality. However, when semantic similarity and word order are critical factors, METEOR is generally preferred. Its ability to recognise paraphrases and penalise grammatical errors makes it more suitable for evaluating systems that prioritise meaning preservation and fluency. In practice, researchers often use both BLEU and METEOR, along with other metrics and human evaluation, to obtain a more comprehensive and nuanced assessment of MTQ. Using multiple metrics provides a more robust evaluation and helps to mitigate the limitations of any single metric.

4. RESEARCH METHODOLOGY

This research critically evaluates the BLEU and METEOR metrics in MTE, comparing their scores against human judgments of machine-translated text to identify their respective strengths and weaknesses. The quantitative evaluation utilises a combination of datasets: the researchers' datasets (Hausa-French-English). The research assesses the robustness of GT software focusing on the standard metric versions. Specifically, the research analyses the quality assessment of five proverbs translated from Hausa to both English and French by using these two metrics software. The research evaluates the robustness of these metrics applying a comparative, scientific, and technical approach grounded in the theory of meaning.

5. DATA PRESENTATION AND ANALYSIS

The research presents data from five (5) selected Hausa proverbs as stated above, translated to English using free translation by considering the cultural and contextual nuances of the language. These translations are compared with those generated by the online GT software, which offers both Hausa-French and Hausa-English language pair translations. The research considers the meaning and philosophy behind each proverb relating it to proverbs where necessary. These proverbs are then presented below with their meanings, philosophies and interpretations based on researchers' findings:

Proverb 1

Original (Hausa proverb)	Google Translate (generated in French)	Google Translate (generated in English)	Human Translation
Haihuwar guzuma, uwa kwance dìya kwance.	Un accouchement par le siège, la mère allongée, la fille allongée.	A breech birth, mother lying down, daughter lying down.	Like mother, like daughter.

This concept of the proverb centers on a perceived weakness, arguing that the strong, by failing to adequately protect or provide for themselves, inadvertently create a similar vulnerability in the weak. This philosophy is starkly illustrated by the scenario of Nigeria a country, often hailed as the "Giant of Africa", yet struggling to effectively care for its own citizens, demonstrating a failure that permeates all levels of society.

The Hausa proverb, "*haihuwar guzuma, uwa kwance dìya kwance*", exemplifies the difference between literal and figurative translations. GT generates it as "un accouchement par le siège, la mère allongée, la fille allongée" in French and "a breech birth, mother lying down, daughter lying down" in English. These are translations made by machine, while accurately capturing the individual words but fail to convey the proverb's deeper meaning. The core phrase, "*haihuwar guzuma*", is literally interpreted as "breech birth" by the machine instead of "old birth". However, the human translation, "like mother, like daughter", offers a figurative interpretation that resonates with the proverb's essence. The phrase "*uwa kwance, dìya kwance*" emphasises the mirroring of weakness, as a breech birth is a challenging delivery. Therefore, "like mother, like daughter" effectively captures the proverb's implication that a daughter follows her mother's path weakness. The researcher concludes that the human translation is superior. While the MT is accurate in its literal rendering, it misses the cultural and proverbial significance. A good translation, as demonstrated by the human version, conveys the meaning and impact of the original text, not just its individual words. In this case, "like mother, like daughter" provides a concise and culturally relevant equivalent that accurately reflects the implied meaning of the Hausa proverb.

Proverb 2

Original (Hausa proverb)	Google Translate (generated in French)	Google Translate (generated in English)	Human Translation
Inda shanuwar gaba ta sha ruwa, nan ta baya ka sha.	Là où les vaches devant boivent de l'eau, les vaches derrière la boivent.	Where the front cows drink water, the back cows drink it.	Follow in the footsteps of those before you.

This Hausa proverb delves into the interconnectedness of power, character, and attitude, particularly within the Hausa cultural context. It posits that children often mirror their parents' behaviours, a consequence of the parents consistently adhering to religious teachings and setting a virtuous example. This philosophy echoes the sentiment of the related Hausa proverb, "*kamar kumbo, kamar kayan ta*", which further emphasises the inherent link between a source and its reflection, highlighting the powerful influence of parental conduct on the younger generation.

The proverb, "*inda shanuwar gaba ta sha ruwa, nan ta baya ka sha*", reveals a clear disparity between machine and human translations. GT, in both French and English, provides literal translations focusing on cows drinking water: "là où les vaches devant boivent de l'eau, les vaches derrière la boivent" and "where the front cows drink water, the back cows drink it". The French translation has defective syntax. The machine misrepresents "*shanuwar gaba*" in Hausa with "les vaches devant" in French and "front cows" in English. These translations, while accurate in their word-for-word rendering, fail to capture the proverb's deeper meaning.

In contrast, the human translation, "follow in the footsteps of those before you", encapsulates the proverb's wisdom. This translation highlights the metaphor of cows following each other to the water, symbolising the process of learning from predecessors and benefiting from their experience. The researcher concludes that "follow in the footsteps of those before you" is the most accurate translation. While the MT is technically correct in its literal interpretation, it lacks the cultural understanding necessary to convey the proverb's intended message. The human translation, by capturing the proverb's essence, demonstrates that effective translation goes beyond mere word substitution; it requires a deep understanding of cultural context and meaning.

Proverb 3

Original (Hausa proverb)	Google Translate (generated in French)	Google Translate (generated in English)	Human Translation
Mai kaza a aljihu, bai jimirin as.	Avec un poulet dans sa poche, il ne supportait pas la chaleur.	With a chicken in his pocket, he couldn't stand the heat.	A guilty conscience fears exposure.

The core message of this Hausa proverb revolves around the themes of guilt, fear, and the consequences of failure. At its heart, it conveys that a life lived with fidelity and honesty eliminates the path to wrongdoing. Furthermore, it asserts that the security provided by a strong, supportive figure, like a father, dispels the fear of loss. This idea is reinforced by the analogous proverb, "*mai uwa a bakin murhu bai rashin abinci*", which translates to "one who has a mother by the hearth will not lack food", emphasising the reassurance and provision that comes from having a dependable protector.

The Hausa proverb, "*Mai kaza a aljihu, bai jimirin as*", highlights the distinction between literal and figurative translation. GT renders it as "avec un poulet dans sa poche, il ne supportait pas la chaleur" in French and "with a chicken in his pocket, he could not stand the heat" in English. While these translations offer a literal depiction, creating a somewhat comical image, they fail to

capture the proverb's deeper meaning. In contrast, the human translation, "a guilty conscience fears exposure", effectively conveys the proverb's essence. The phrase "*kaza a aljihu*" or "chicken in the pocket" symbolises something hidden, a concealed wrongdoing. The "heat" represents the pressure and discomfort of potential discovery, suggesting that a guilty person constantly fears their actions being revealed. This discomfort is not physical heat, but the metaphorical heat of exposure.

Consequently, the research concludes that "a guilty conscience fears exposure" is the superior translation. While the MT accurately reflects the individual words, it lacks the figurative understanding necessary to convey the proverb's true meaning. The human translation, by capturing the underlying message of guilt and fear, demonstrates the crucial role of cultural context in accurate proverb translation.

Proverb 4

Original (Hausa proverb)	**Google Translate (generated in French)**	**Google Translate (generated in English)**	**Human Translation**
Rashin sani ya sa kaza ta kwan bisa gwallo da yunwa.	L'ignorance pousse une poule à pondre des œufs sur un lit d'herbe et à souffrir de la faim.	Ignorance causes a hen to lay eggs on a bed of grass and hunger.	Lack of knowledge leads to poor decisions.

This Hausa proverb centers on the critical concept of ignorance. It shares a common thread with similar expressions like "*Jahilci rigar ƙaya*" literally "ignorance is a thorny garment" and "*rashin sani ya fi dare duhu*" literally "lack of knowledge is darker than night", all highlighting the detrimental nature of being uninformed. The underlying philosophy underscores the paramount importance of knowledge in navigating life, effectively illustrating the profound problems and limitations that arise from ignorance or a lack of understanding.

The proverb, "*rashin sani ya sa kaza ta kwan bisa gwallo da yunwa*", vividly illustrates the challenges of proverb translation. The researcher's analysis reveals a stark contrast between machine and human interpretations. GT, in both French and English, generates the proverb with a literal focus on "a hen laying eggs on a bed of grass" and "suffering hunger". Specifically, the English version reads, "Ignorance causes a hen to lay eggs on a bed of grass and hunger". This literal approach, however, misses the proverb's deeper meaning, primarily due to the misinterpretation of "*gwallo*" which is "food" as "bed of grass". Considering the cultural meaning of the prover, the word "gwallo" means "*tsaba*" which is food for the hen but due to ignorance it ends up with hunger without eating it.

In contrast, the human translation, "lack of knowledge leads to poor decisions", captures the proverb's essence from the original and target language. This translation highlights that "*rashin sani*", meaning lack of knowledge or ignorance, is the root cause of poor choices, symbolised by the "hen's hunger" or "*da yunwa*". While the MT accurately reflect the individual words, it fails to grasp the proverb's figurative meaning. Ultimately, the researcher concludes that "lack of knowledge leads to poor decisions" is the most accurate translation. It effectively conveys the proverb's underlying message about the importance of knowledge and understanding in making sound decisions, demonstrating the superior ability of human translators to interpret and convey the cultural wisdom embedded within the Hausa language.

Proverb 5

Original (Hausa proverb)	Google Translate (generated in French)	Google Translate (generated in English)	Human Translation
Sussuka ɗaka, shiƙa ɗaka.	De rien, de rien.	You are welcome, you are welcome.	Home arrangement.

This Hausa proverb addresses the negative aspects of human nature, specifically highlighting individualism, greed, and pretense. It conveys a philosophy rooted in the desire to withhold resources or opportunities from others, revealing a self-centered mindset that prioritises personal gain over shared prosperity.

The proverb, "*Sussuka ɗaka, shiƙa ɗaka*" presents a unique challenge in translation for the machine. GT generates it as "De rien, de rien" in French and "You are welcome, you are welcome" in English. However, these literal translations diverge significantly from the human translation. Notably, the Hausa word "*ɗaka*" (room), is misinterpreted by machines as "*daka*" (relating to pounding or grinding). This proves that the machine did not understand the context and meaning of the word. This is coupled with the ambiguous "*Sussuka"* (threshing) and *"shiƙa"* (winnowing) that make direct translation difficult without contextual understanding. Consequently, the machine's translation "You are welcome" interpretation is demonstrably inaccurate, revealing a misinterpretation of the individual words. In contrast, the human translation, of "Home arrangement", arises from a deep understanding of the specific context in which the proverb was used. This discrepancy underscores the proverb's inherent complexity and the machine's struggle to grasp its true meaning. Finally, the researcher concludes that "Home arrangement" is the most accurate translation, given the specific context. This example highlights the critical importance of cultural and contextual awareness in translation, capabilities that remain essential for accurately interpreting the nuances of the Hausa language.

From the data presented and interpretations above, human translation consistently outperforms MT in accurately conveying the nuanced meanings of Hausa proverbs. While GT provides literal translations that often capture the individual words, it frequently fails to grasp the cultural and figurative essence of the proverbs. In contrast, human translators demonstrate a superior ability to interpret and translate proverbs using free translation by considering the cultural context and understanding, idiomatic expressions, and implied meanings and philosophies behind each proverb.

For instance, "*haihuwar guzuma, uwa kwance dîya kwance*" is correctly translated as "like mother, like daughter" by humans, while machines offer literal descriptions of "a breech birth". Similarly, "*Mai kaza a aljihu, bai jimirin as*" is accurately rendered as "a guilty conscience fears exposure" by human translators, capturing the proverb's metaphorical meaning, which is lost in the machine's literal rendition. In all five examples above, the human translations effectively convey the intended wisdom and cultural understanding, proving that the ability to recognise and interpret cultural nuances is crucial for accurate translation, a capability that currently surpasses the limitations of MT. Therefore, human translation is demonstrably the best method for translating culturally rich and context-dependent expressions like proverbs.

6. RESEARCH IMPLICATIONS

This research highlights the critical need for culturally sensitive translation of Hausa proverbs, demonstrating the limitations of MT and emphasising the importance of human expertise. It underscores the need for enhanced Hausa language resources and suggests avenues for improving

MT by incorporating cultural context. The findings have implications for language preservation, cross-cultural communication, and the development of more effective translation tools.

7. CONCLUSION

This research examined five Hausa proverbs, comparing human translations, which considered cultural and contextual nuances, with those generated by GT's Hausa-French and Hausa-English MT tools. The analysis revealed a consistent pattern while GT provided literal translations of the individual words, it frequently failed to capture the deeper, figurative meaning embedded within the proverbs. For example, "*Haihuwar guzuma, uwa kwance dìya kwance"*, literally translated as "A breech birth, mother lying down, daughter lying down", was more accurately conveyed by the human translation, "Like mother, like daughter". This highlights the crucial role of cultural understanding in translation, as the human translators were able to discern the implied comparison and convey the proverb's wisdom. Similarly, the proverb about the hen laying eggs on grass and suffering hunger was literally translated, but the human translation, "Lack of knowledge leads to poor decisions", effectively captured the proverb's message about the consequences of ignorance. In the case of "*Sussuka ɗaka, shika ɗaka*," the literal translations were completely inaccurate, whereas the human translation, "Home arrangement", while potentially context-dependent, demonstrated an attempt to interpret the phrase beyond its individual words.

The findings of this research underscore the limitations of MT tools when dealing with culturally rich and context-dependent language like proverbs. While GT could be a useful tool for basic vocabulary and sentence structure, it often struggles with figurative language, idiomatic expressions, and cultural nuances. Proverbs, by their very nature, are concise expressions of cultural wisdom and often rely on metaphor, analogy, and shared cultural knowledge. Therefore, human translators, equipped with cultural sensitivity and an understanding of the proverb's context, are essential for accurately conveying the meaning and impact of these expressions. This research reinforces the importance of human involvement in translation, particularly when dealing with texts that are deeply embedded in culture and require interpretation beyond the literal level.

8. RECOMMENDATIONS

Based on the findings of this research, the following recommendations are made:

- When translating proverbs, idioms, or other culturally nuanced texts, human translators with deep cultural understanding should be prioritised. MT tools could be a starting point, but they should not be relied upon as the sole source for accurate and meaningful translations. Human translators are better equipped to discern the figurative language, cultural context, and implied meanings that are essential for conveying the true essence of such texts.
- Developers of MT tools should focus on incorporating cultural knowledge and context into their algorithms. This could involve training models on larger datasets of culturally annotated texts, developing methods for recognising and interpreting figurative language, and incorporating feedback from human translators with cultural expertise.
- Collaboration between human translators and MT developers is crucial for improving the accuracy and cultural sensitivity of MT tools. Human translators could provide valuable insights into the nuances of language and culture, while MT developers could use this feedback to refine their models.

- The research highlights the need for more resources dedicated to the Hausa language and culture, including comprehensive dictionaries, annotated texts, and cultural guides.
- Users of MT tools should be educated about their limitations, particularly when dealing with culturally rich texts. It should be emphasised that MT-generated translations should be reviewed and, if necessary, revised by human translators to ensure accuracy and cultural appropriateness.
- This research focused on a limited number of Hausa proverbs. Further research is needed to explore a wider range of proverbs and other culturally bound expressions in Hausa and other languages.

9. RESEARCH CHALLENGES AND FUTURE DIRECTION

This research, while illuminating the difficulties in translating Hausa proverbs, reveals limitations that point to future research directions. Primarily, the scarcity of annotated Hausa texts and cultural resources hinders effective MT training and evaluation. Additionally, the subjective nature of translation, especially with figurative language, creates variability in human interpretations. Furthermore, the limited proverb sample size restricts the generalisability of findings. Future research should focus on expanding the Hausa proverb corpus, developing cultural resources, investigating inter-translator variability, exploring computational approaches to cultural context, conducting cross-lingual and cross-cultural studies, evaluating MT-assisted translation, and focusing on specific domains within Hausa proverbs. These efforts aim to improve both the human and MT of Hausa proverbs by addressing resource limitations, acknowledging interpretive variability, and exploring computational methods for capturing cultural nuances.

ACKNOWLEDGEMENTS

The authors would like to thank Almighty Allah for his favours and blessings.

REFERENCES

[1] Banerjee, S., et al., METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, (2005).

[2] Callison-Burch, C., et al., BLEU: Still a Good Baseline. Proceedings of the ACL Workshop on Statistical Machine Translation, (2006).

[3] Hutchins, W. J., et al., An Introduction to Machine Translation. Academic Press, (1992).

[4] Lo, C., Wu, Y. et al., Evaluating Machine Translation using Semantic Role Labeling. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (2012).

[5] Owczarzak, K., et al., Towards Syntactic Evaluation of Machine Translation. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), (2010).

[6] Nwanjoku, A. C and Zaki, M. Z., Exploring Translation Technology in the 21st Century: The Fate of the Human Translator. SSR Journal of Arts, Humanities and Social Sciences (SSRJAHSS) (2024) Vol 1, Issues 2 pages 102-107.

[7] Nwanjoku, A. C and Zaki, M. Z., A Reflection on the Practice of Auto-Translation and Self-Translation in the Twenty-First Century. Case Studies Journal (2021) Volume 10, Issue, pages 8 - 16.

[8] Nwanjoku, A. C and Zaki, M. Z., Achieving Correspondence and/or Equivalence in Translation, An Evaluation of the Translation Ekwensi's Burning Grass into French as La Brousse ardente by Françoise Balogun. Quest Journals - Journal of Research in Humanities and Social Science. (2021) Volume 9, Issue 5, pages: 50-55.

[9] Papineni, K., et al., BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (2002).
[10] Sajo, M. A., and Zaki, M. Z., A Study of Communication Miscarriages in News Translation. Translation Studies: Theory and Practice International Scientific Journal. Vol 3, Issue 2 (6) (2023) pages 70 – 79.
[11] Snover, M., et al., A Study of Translation Edit Rate (TER) with Human Judges. Proceedings of the 7th conference of the Association for Machine Translation in the Americas, (2006).
[12] Triveni, L. P., Evaluation of Quality Metrics for Machine Translation: A Case Study of English-Hindi Machine Translation. 2012.
[13] Zaki, M. Z., Transforming Worlds: The Intersection of Translation Technology and Transformers in the Digital Age. Computer Science and Engineering: An International Journal. (2024) Vol. 14 Issue 3 pages 1-14.
[14] Zaki, M. Z., Revolutionising Translation Technology: A Comparative Study of Variant Transformer Models – BERT, GPT and T5. Computer Science and Engineering: An International Journal. (2024) Vol. 14 Issue 3 pages 15-27.
[15] Zaki, M. Z. et al., Bridging Linguistic Divides: The Impact of AI-powered Translation Systems on Communication Equity and Inclusion. Saba Publishers: Journal of Translation and Language Studies. (2024) Vol. 5, Issue 2 pages 20-30.
[16] Zaki, M. Z., Explaining Some Fundamentals of Translation Technology. GAS Journal of Arts Humanities and Social Sciences (GASJAHSS). Gas Publishers: (2024) Vol 2 Issue 3 pages 177-185.
[17] Zaki, M. Z., Translator's Vantage points: Modern Translation Technology. Moldova: Lambert Academic Publishing, 2024.
[18] Zaki, M. Z., and Nwanjoku, A. C Understanding Terminologies of CAT Tools and Machine Translation Applications. Case Studies Journal, (2021) Volume 10, Issue 12, pages 30-39.
[19] Zaki, M. Z. et al., et al., Appreciating Online Software Based Machine Translation: Google. Translator. International Journal of Multidisciplinary Academic Research (2021) Vol. 2, Issue (2) pages 1-7.
[20] Zaki, M. Z. et al., An appraisal of Audio-Visual Translation, Implications of Subtitling from Hausa of Sub-Saharan Africa to English. International Journal of Scientific & Engineering Research (IJSER), (2021) Volume 12, Issue 4, Pages 238-244.
[21] Zhao, W., et al., BERT-based Semantic Similarity for Paraphrase Identification. arXiv preprint arXiv:1909.03846. (2019).

AUTHORS

ZAKI, Muhammad Zayyanu is a translator, translation technologist and Senior Lecturer of Translation Studies in the French Department, Faculty of Arts, Usmanu Danfodiyo University, Sokoto - Nigeria. He is a member of Translation bodies that include Nigerian Institute of Translators and Interpreters (NITI) and South East Chapter and International body of Translators Without Boarders (TWB). He majored in French, English, and Hausa Translation and Interpretation. He has attended more than thirty (30) International and National academic conferences and seminars, and has published widely in Hausa, English, and French in reputable journals locally and internationally with impact factors (many of which are Translation Studies and translation technology). He also published a book title "*A Concise HandBook of Modern Translation Technology Terms"* in 2023, "*Notions pertinentes de la traductologie*" in 2024 and "*Translator's Vantage points: Modern Translation Technology*" in 2024. His research areas of interest include Translation Studies, Translation technology, Linguistics, Cultural Shifts (with a passion for French, English, Arabic, and Hausa translations).

Nazir Ibrahim Abbas was born in Sokoto, Nigeria on the 15th December 1977. He hails from Maradun Local Government, Zamfara State. He attended Model Primary School Rabah 1984-1989, Sheikh Abubakar Gummi Memorial College, Sokoto 1990-1995 and Usmanu Danfodiyo University, Sokoto 1997 -2000 for his B.A. Hausa Language. He later obtained his M.A. and Ph.D in Hausa Language between 2009-2012 and 2014-2019 respectively. He was employed as a Graduate Assistant in 2008. He is currently an Associate Professor in Hausa Language, in the Department of Nigerian Languages, Usmanu Danfodiyo University, Sokoto. His area of specialisation and interest is dialectology and Translation Studies. He has attended many International and National academic conferences and seminars, and has published widely in Hausa and English in reputable journals locally and internationally.