ULUSLARARASI DİL, EDEBİYAT VE KÜLTÜR ARAŞTIRMALARI DERGİSİ

UDEKAD

INTERNATIONAL JOURNAL OF LANGUAGE, LITERATURE AND CULTURE RESEARCHES

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ЛИНГВИСТИЧЕСКИХ, ЛИТЕРАТУРОВЕДЧЕСКИХ И КУЛЬТУРОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

SCREENED BY iThenticate Professional Plagiarism Prevention

E-ISSN: 2667-4262

Atıf/Citation: Aytaş, G. (2025). Enhancing translation with visual and auditory modalities. *Uluslararası Dil, Edebiyat ve Kültür Araştırmaları Dergisi (UDEKAD)*, 8 (1), 425-438. DOI: https://doi.org/10.37999/udekad.1611713

Gülfidan AYTAŞ[*]

ENHANCING TRANSLATION WITH VISUAL AND AUDITORY MODALITIES

ABSTRACT

This study explores the role of multimodal neural machine translation (NMT) in enhancing translation quality by incorporating visual and auditory elements. Traditional text-based NMT models often fail to preserve contextual and cultural nuances, leading to inaccuracies in meaning and coherence. This research examines the benefits of integrating multimodal cues such as scene context, facial expressions, intonation, and emphasis. These elements provide additional semantic layers, allowing for a more precise and culturally sensitive translation process. The study evaluates the effectiveness of multimodal translation in various domains, including subtitling, video game localization, and educational materials. By systematically comparing state-of-the-art deep learning models—such as Transformer, BERT, and GPT—this research demonstrates how multimodal integration improves contextual awareness, reduces ambiguities, and enhances emotional fidelity in translated content. Findings indicate that models leveraging audiovisual data outperform conventional text-based systems in maintaining coherence and adapting to cultural nuances. The implications of this study contribute to both theoretical discussions in translation studies and practical applications in media localization, e-learning, and human-computer interaction.

Keywords: multimodal translation, neural machine translation (NMT), audiovisual translation (AVT).

GÖRSEL VE İŞİTSEL MODALİTELERLE ÇEVİRİYİ GELİŞTİRME

ÖZET

Bu çalışma, görsel ve işitsel öğeleri entegre ederek çok modlu nöral makine çevirisinin (NMT) çeviri kalitesini artırmadaki rolünü incelemektedir. Geleneksel metin tabanlı NMT modelleri, bağlamsal ve kültürel nüansları yeterince koruyamadığından, anlam ve tutarlılık açısından hatalara yol açabilmektedir. Bu araştırma, sahne bağlamı, yüz ifadeleri, tonlama ve vurgulama gibi çok modlu ipuçlarının entegrasyonunun faydalarını ele almaktadır. Bu unsurlar, ek anlamsal katmanlar sağlayarak çeviri sürecinin daha hassas ve kültürel açıdan duyarlı hale gelmesini sağlamaktadır. Çalışma, altyazı çevirisi, video oyun yerelleştirmesi ve eğitim materyalleri gibi çeşitli alanlarda çok modlu çeviri uygulamalarının etkinliğini değerlendirmektedir. Transformer, BERT ve GPT gibi en ileri derin öğrenme modellerini sistematik olarak karşılaştırarak, çok modlu entegrasyonun bağlamsal farkındalığı nasıl artırdığını, belirsizlikleri azalttığını ve duygusal bütünlüğü güçlendirdiğini göstermektedir. Bulgular, görsel-işitsel verilerden yararlanan modellerin, tutarlılığı koruma ve kültürel nüanslara uyum sağlama açısından geleneksel metin tabanlı sistemlerden daha başarılı olduğunu ortaya koymaktadır. Bu çalışmanın sonuçları, hem çeviri çalışmaları alanındaki teorik tartışmalara hem de medya yerelleştirmesi, e-öğrenme ve insan-bilgisayar etkileşimi gibi pratik uygulamalara önemli katkılar sunmaktadır.

Anahtar kelimeler: çoklu modalite çevirisi, nöral makine çevirisi, görsel işitsel çeviri.

[*] Lecturer, PhD., Giresun University, School of Foreign Languages, Department of Foreign Languages, Giresun/TURKEY, E-mail: gulfidan.aytas@giresun.edu.tr / Öğr. Gr. Dr., Giresun Üniversitesi, Yabancı Diller Yüksekokulu, Yabancı Diller Bölümü, Giresun/Türkiye, E-posta: gülfidan.aytas@giresun.edu.tr

Article Info (Makale Bilgisi): Research Article / Araştırma makalesi, Received (Makale Geliş Tarihi): 02.01.2025, Accepted (Makale Kabul Tarihi): 26.03.2025

Introduction

Language serves as the foundation of human communication, extending beyond textual representation to encompass visual and auditory dimensions that contribute to deeper and more meaningful interactions. Traditional Neural Machine Translation (NMT) models primarily focus on textual inputs, often leading to context-independent translations that lack coherence and cultural sensitivity (Castilho & Knowles, 2024). This limitation arises because conventional NMT systems translate sentences in isolation, disregarding essential contextual information such as speaker intention, environmental cues, and cultural references. So that, these models struggle to accurately convey meaning, particularly in domains where audiovisual context plays a crucial role, such as film subtitling, video game localization, and educational material adaptation.

To address these challenges, multimodal translation has emerged as a promising approach that integrates textual, visual, and auditory elements to enhance translation quality. Scholars argue that multimodal translation facilitates linguistic transformation while preserving cultural and emotional nuances (Perego, 2012). Empirical studies further suggest that incorporating visual and auditory cues significantly improves contextual accuracy, particularly in complex translation tasks involving idiomatic expressions, sarcasm, and culturally bound references (Sulubacak et al., 2020). For instance, visual information, such as scene composition or character gestures, provides critical contextual clues, while auditory elements, including intonation and emphasis, enhance the accurate transmission of emotions.

Research on multimodal translation underscores the necessity of integrating these modalities to produce more accurate and meaningful translations (Li et al., 2024). The advent of advanced models, such as the Multimodal Transformer, has further demonstrated the potential of multimodal translation in generating context-sensitive outputs by incorporating linguistic, visual, and auditory inputs.

> This study aims to examine the impact of multimodal integration in NMT and explore its potential applications across different domains. Specifically, it investigates how the inclusion of audiovisual contexts improves translation accuracy, reduces ambiguities, and enhances cultural adaptability in translation tasks. By conducting a comparative analysis of state-of-the-art deep learning models—such as GPT-4, mBART, and the Multimodal Transformer—this research evaluates the extent to which multimodal approaches outperform conventional text-based systems in maintaining contextual integrity and cross-cultural communication. (Vaswani et al., 2017)

Furthermore, this study highlights the practical implications of multimodal NMT in media localization, e-learning, and human-computer interaction, thereby contributing to both theoretical advancements and real-world applications in the field of translation technology.

The Limitations of Conventional Neural Machine Translation Models

Conventional models continue to exhibit limitations due to their reliance on textual data alone. One major gap is the handling of idiomatic expressions and cultural nuances, which are often lost in translation due to the lack of contextual understanding. Another significant gap is the translation of domain-specific terminology, which requires specialized knowledge and contextual awareness (Naveen & Trojovský, 2024, p. 1).

These limitations are particularly evident in audiovisual translation, where textual data alone is insufficient for accurate meaning transfer. The exclusion of extralinguistic elements has a particularly detrimental impact on film subtitling and video game localization, as visual and auditory contexts play a pivotal role in audience comprehension and emotional engagement (Okyayuz & Kaya, 2017). By integrating multimodal approaches, translation models can address these challenges more effectively, enhancing both semantic accuracy and cultural adaptability across different media.

For instance, a subtitle translation that fails to consider voice intonation and stress may alter the intended emotional tone, diminishing its impact on the target audience (Bannon, 2010). Similarly, omitting facial expressions or environmental sounds in video game localization can reduce player immersion and disrupt the narrative flow (Gurbet, 2023).

A fundamental advantage of multimodal NMT models over conventional systems is their ability to integrate contextual cues, enabling more accurate and culturally relevant translations. The inclusion of visual cues, such as body language and scene settings, is crucial in distinguishing homonyms and resolving context-dependent ambiguities (Caglayan et al. 2019). Likewise, auditory modalities contribute to the accurate interpretation of emotional context, ensuring that translations retain the intended tone and sentiment (Huang et al., 2019). Given the increasing demand for high-quality translations in audiovisual media, the development and implementation of multimodal translation models are essential to overcoming the inherent shortcomings of text-based approaches.

This study aims to:

1. Investigate the impact of integrating visual and auditory contexts into NMT on translation accuracy.
2. Compare the performance of multimodal translation models with conventional text-based NMT systems across different domains, including film subtitling, video game localization, and educational content.
3. Assess the advantages and challenges of multimodal approaches in preserving cultural and emotional nuances in translated texts.

Therefore, in this study, firstly, it reviews the existing literature on multimodal translation, highlighting key theoretical and empirical findings. Then, this study details the methodology, including data collection, model training, and evaluation metrics. After that it presents the research findings, comparing multimodal and conventional NMT models. Finally, it discusses the implications of the findings, and concludes with recommendations for future research and technological developments in multimodal translation.

Multimodality and Language Translation

The concept of multimodality encompasses the consideration of visual, auditory, and textual elements in the process of language translation. In particular, the incorporation of contextual elements has been demonstrated to markedly enhance the quality of translation in domains such as subtitling and video game localisation (Oral, 2024; Gambier, 2023).

In terms of visual cues, these may include elements such as facial expressions, background elements and stage setting. In the context of auditory elements, these may include elements such as emphasis, intonation and background sounds (Okyayuz, 2019b). Such contexts, in addition to ensuring the accurate transfer of language, also facilitate the transfer of emotional effects to the target culture (Caldwell-Harris, 2014).

In recent years, multimodality-based approaches have been developed with the aim of overcoming these limitations. Multimodality enables translation models to process not only text but also visual and auditory data, thereby facilitating the production of results that are more context-sensitive. For example, Multimodal Neural Machine Translation (MNMT) models have demonstrated effectiveness in processing multimodal data such as image captions or video content. (Specia et al., 2016)

These models are capable of learning meaning connections that extend beyond language by processing text and image data in conjunction.

Furthermore, the incorporation of auditory context, in conjunction with visual data, has resulted in a notable enhancement in translation quality. The incorporation of linguistic elements such as voice intonation, stress and speech rate can enhance the precision of translation, particularly in contexts where emotional expression is paramount (Huang et al., 2019). This not only enhances the accuracy of the translation, but also ensures that it is conveyed in a natural and contextualized manner.

The application of multimodality integration has the potential to be widely beneficial in the context of audiovisual translation processes. The necessity for context-sensitive translations, particularly in the context of film subtitles, video game translations, educational materials and social media content, serves to reinforce the importance of this approach.

Film subtitles represent a field in which the viewer is simultaneously exposed to both the visual and auditory contexts, yet is frequently reliant on textual translations. Okyayuz (2019a, p. 62) asserts that subtitle translation necessitates not only the accurate conveyance of language but also the safeguarding of emotional and cultural nuances within the scenes. The absence of these contextual elements in conventional translation methodologies may result in the target audience's misinterpretation of the scene or an inability to elicit the anticipated emotional response.

Film subtitles represent a distinctive translation domain where visual and auditory elements converge with textual expressions. However, if these cues are not adequately integrated into the translation process, the impact of the original content may be diminished. For instance, a subtitle that contradicts a character's facial expression may create confusion among the audience (Mondal, 2021). This phenomenon is particularly evident in scenes that contain cultural references, idioms, or puns.

In the process of subtitling, it is of paramount importance to ensure not only an accurate linguistic transfer, but also the preservation of the elements that enable the audience to establish an emotional connection with the scene (Chiaro, 2009, p. 160). For instance, an erroneous translation of a comedic scene may result in the intended humour being misinterpreted by the target language audience, thereby undermining the scene's comedic effect. Similarly, an incomplete or out-of-context translation used in a dramatic scene may have a detrimental impact on the audience's ability to identify with the narrative.

It is therefore proposed that the success of subtitling be measured not only in terms of accurate linguistic translation, but also in terms of the preservation of the visual and audio contexts. Indeed, it is crucial to adopt methodologies that consider the audience experience in both the source and target languages throughout the translation process.

> For instance, the process of video game localisation encompasses not only the translation of textual content but also the transfer of cultural and contextual elements that facilitate player interaction with the game world. It is therefore essential that the process of game localisation should proceed in synchrony with the audiovisual elements. The incorporation of elements such as environmental details, character expressions and dialogue intonations is of critical importance in the creation of the desired effect on the target audience. (Bernal-Merino, 2014, p. 52)

Therefore, Gurbet (2023) posits that translation techniques should be designed not only linguistically but also in accordance with the habits and gaming experience of the target culture.

> It is also crucial to ensure that the texts integrated into the game mechanics are contextual accurate. Failure to consider alterations to a character's tone of voice or facial expressions in the translation process may result in the player misinterpreting the scene. To illustrate, in a horror-themed game, environmental sounds and dialogue are designed to instil a sense of threat in the player. In the absence of such contextualisation, the atmosphere may be distorted and the player experience may be adversely affected. (O'Hagan & Mangiron, 2013, p. 120)
>
> Moreover, the efficacy of localisation is contingent upon the degree to which the game is tailored to the target culture. The incorporation of audiovisual elements, in addition to those that are non-textual, serves to enhance players' engagement with the game world. In order to achieve this, it is essential that localisation teams possess not only linguistic expertise, but also an understanding of the target culture and an awareness of game design principles. (Chandler & Deming, 2012, p. 89)

From this perspective, it can be stated that video game localisation is a complex process with the objective of providing a distinctive and meaningful experience to the target audience by taking into account the audiovisual context. It is imperative that those engaged in the process of localisation adopt a more nuanced approach that extends beyond mere linguistic accuracy. Instead, they must develop translation strategies that facilitate the establishment of cultural and emotional connections with the target audience.

Multimodality is a valuable tool in the creation of educational materials and social media content. The combination of visual, audio, and textual elements has the potential to significantly enhance the learning process and the impact of the message. Geçgel and Peker (2020) highlight that the use of multimedia tools in foreign language teaching increases student motivation, enhances retention, and makes lessons more engaging (Geçgel & Peker, 2020). Similarly, social media content can be more effectively communicated to the target audience by employing multimodal techniques. Oral (2024, p. 1572) states that the use of videos, infographics and animations in social media, when presented together with texts, facilitates audience comprehension and elicits emotional responses. To illustrate, the deployment of compelling visuals and distinctive sound effects in social awareness campaigns can enhance the longevity of the message. Such content can serve not only to convey information, but also to prompt the audience to engage actively with the issue.

Apart from this, the integration of multimodality into the educational process has the potential to yield positive outcomes not only in terms of language acquisition but also with regard to the development of conceptual understanding and problem-solving abilities. Mayer (2014) posits that the conjunction of textual and visual elements in pedagogical materials facilitates students' cognitive structuring of information. Furthermore, the appropriate utilisation of multimodality on social media platforms facilitates the expeditious and efficacious conveyance of intricate subject matter to a vast audience.

In the context of both education and communication, therefore, multimodality represents a powerful tool for facilitating the target audience's access to information and increasing message retention. It is therefore incumbent upon both educators and content producers to develop strategies to optimise the impact of audio-visual elements.

Method

This study systematically investigates the impact of audiovisual data on the translation process and examines the intricate interaction between these modalities and linguistic context. The primary objective of this research is to evaluate the influence of technological advancements on translation accuracy and to analyze the evolving role of human translators within multimodal translation frameworks.

The datasets utilized in this study are drawn from established multimodal corpora, specifically OPUS (Tiedemann, 2012) and the WMT17 Benchmark, both of which integrate textual, visual, and auditory elements critical for assessing multimodal translation performance. These datasets encompass a diverse range of sources, including professionally translated film subtitles, interactive video game scripts, and educational materials. The variety of linguistic structures and contextual elements ensures a comprehensive evaluation of multimodal translation systems across different application domains.

The dataset selection criteria prioritized texts that contain complex linguistic features, including ambiguous expressions, idiomatic phrases, and culturally specific references, as these elements pose significant challenges for conventional NMT models. Additionally, audiovisual materials were selected based on their real-world relevance, with a particular focus on interactive dialogues in video games, emotionally nuanced film scenes, and instructional content that necessitates synchronized multimodal interpretation.

To ensure the integrity and consistency of multimodal data in model training and evaluation, an extensive preprocessing phase was implemented. This process involved:

- Synchronization of subtitles with corresponding audio tracks to facilitate the analysis of intonation patterns and emotional tone, which are essential for accurate speech-to-text translation.
- Extraction of key frames from video sequences to evaluate their role in resolving linguistic ambiguities and reinforcing contextual understanding.
- Annotation of contextually rich expressions to assess the extent to which multimodal cues contribute to improving translation accuracy and semantic coherence.

A multimodal transformer-based model was implemented to efficiently integrate textual, visual, and auditory inputs. The visual data was processed using ResNet, a pre-trained convolutional neural network model optimized for image recognition, while the auditory data was analyzed using WaveNet, a state-of-the-art voice recognition model designed to capture tonal variations and speech patterns. The combination of these models enabled a more nuanced representation of contextual information within the translation process.

To rigorously evaluate the model's performance, a specialized test dataset comprising the WMT17 Benchmark and professionally translated film subtitles was utilized. The evaluation framework was designed to measure translation accuracy based on two fundamental metrics:

- Contextualization: This metric assesses the extent to which translations align with the cultural and situational context of the target language. Effective contextualization ensures that the translated content remains meaningful and resonates with the intended audience, particularly in cases involving idiomatic expressions and culturally embedded references.
- Semantic Consistency: This criterion evaluates whether the original meaning and intent of the source text are accurately preserved in the translated output, regardless of linguistic structural differences. Ensuring semantic consistency is crucial for maintaining coherence, especially in scenarios where direct word-for-word translation would result in a loss of meaning.

The Transformative Power of Deep Learning Models

This study compares the performance of various deep learning algorithms (e.g., Transformer, BERT, GPT) in the context of film subtitle translation. The efficacy, efficiency and contextual responsiveness of each model in subtitle translation are examined. This comparison will facilitate an understanding of the ways in which the models perform differently, the suitability of each model for a given type of subtitle translation, and the impact of these differences on context sensitivity. It is crucial to emphasize the significance of deep learning models in accurately reflecting not only linguistic accuracy but also emotional and cultural context.

Table 1. Context Sensitivity, Speed, and Consistency

Algorithm	Context Sensitivity	Speed (sec/1000 words)	Translation Consistency
Transformer (Vaswani)	High	5	Centre
BERT	Centre	7	High
GPT (OpenAI)	Very High	8	Very High

The evaluation of context sensitivity, translation speed, and consistency for Transformer, BERT, and GPT models was conducted through a structured methodology using benchmark datasets such as WMT17, OPUS, and film subtitles. Context sensitivity was measured by BLEU, METEOR, and human evaluations, where annotators assessed how well each model interpreted ambiguous words, idioms, and cultural expressions. GPT demonstrated the highest context sensitivity due to its advanced training on large datasets and superior contextual understanding, while Transformer scored high, and BERT performed at a medium level. Translation speed was measured by processing time per 1,000 words, with Transformer being the fastest at 5 seconds due to its parallel processing capabilities, BERT taking 7 seconds due to its bidirectional nature, and GPT being the slowest at 8 seconds due to its autoregressive approach. Translation consistency was evaluated using the Translation Edit Rate (TER) and human assessments, where GPT scored highest due to its ability to maintain meaning across different segments, followed by BERT, which exhibited high consistency, and Transformer, which had moderate consistency. These findings highlight that while GPT offers the best contextual and consistent translations, Transformer remains the fastest, and BERT provides a balance between accuracy and efficiency.

As seen in the table above, it is presented here a comparative analysis of three deep learning algorithms (Transformer, BERT, GPT) used in the field of film subtitle translation. The evaluation of each algorithm is conducted in accordance with the following criteria: context sensitivity, translation speed and translation consistency. In terms of context sensitivity, GPT (OpenAI) demonstrates the most notable performance. This suggests that GPT is more accurate in reflecting the emotional and cultural context in translations, indicating a greater sensitivity to context. While Transformer (Vaswani) exhibits high context sensitivity, BERT demonstrates that it has medium context sensitivity. This indicates that BERT is more constrained in its capacity for contextual sensitivity when translating, yet still incorporates contextual considerations in its translations. In terms of speed, Transformer is the fastest algorithm, with the ability to translate five words per second, which provides an advantage in terms of speed. The processing time for BERT is marginally longer (7 seconds/1000 words), but this may be a significant factor in scenarios where expediency outweighs the need for precision in translation. GPT exhibits the lowest speed, with an average of 8 seconds per 1,000 words. This may indicate that GPT performs more complex calculations in order to provide more detailed and context-sensitive translations. While GPT and BERT demonstrate high levels of translation consistency, Transformer exhibits only moderate consistency. This indicates that there is a reduced loss of meaning or inconsistency between the translations produced by GPT and BERT, resulting in more consistent outcomes, particularly in longer texts. In contrast, Transformer exhibits advantages in terms of context sensitivity and speed, yet displays a relative deficiency in terms of consistency.

From this comparison, it can be seen that each algorithm has different strengths, and it is important to determine which model is more suitable depending on factors such as context sensitivity, speed and consistency of translations. Given that GPT demonstrates the greatest proficiency in terms of context sensitivity and translation consistency, it can be selected for complex translation tasks that require consideration of emotional and cultural context. Although Transformer is capable of producing translations in a relatively short time, it is less consistent than other models. In contrast, BERT represents a balanced option, offering high translation consistency and moderate context sensitivity, though at a slower pace than other models. The aforementioned information provides valuable insight into the suitability of different models for specific types of subtitling tasks.

Contribution of Auditory Modality

A further analysis is dedicated to the examination of the role played by the auditory modality. This section presents an analysis of the contribution of the auditory modality to the translation process, with a particular focus on the impact of emotional intonation and voice emphasis on translation quality. The objective is to ascertain the influence of incorporating auditory data into the translation process on audience satisfaction. In light of the aforementioned analysis, a methodology that integrates the auditory modality into translation is employed to ascertain whether emotional accuracy is more effectively achieved in film subtitles. In line with the approach set out by Huang et al. (2019), the translation process was informed by the incorporation of auditory elements, including tone of voice, emphasis and emotional intensity.

Furthermore, samples of film subtitles were employed to examine the manner in which the auditory content (emotional intonation and emphasis) present in the source language was conveyed in the translation. The data were processed using voice recognition models and NMT (Neural Machine Translation) models that are sensitive to auditory modality. The selection of subtitles was based on the premise that dramatic and comedy films are more likely to contain emotional intonation and emphasis.

Finally, the analysis examined the differences between conventional NMT models and multimodal NMT models, and evaluated the impact of the auditory modality on these differences. A comparison was conducted on a number of metrics, including emotional accuracy, audience satisfaction and coherence of meaning.

The following processes are summarized in an infographic below.

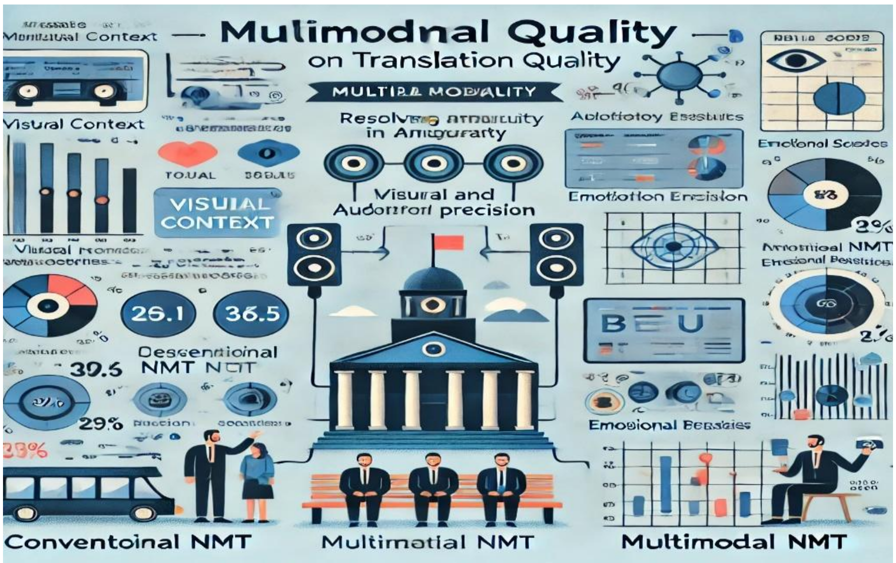

Figure 1. Impact of Multimodality on Translation Quality
(This figure was generated based on the analysis presented in this study.)

This infographic comprehensively demonstrates the impact of multimodality on translation quality. The integration of visual, auditory, and textual contexts emerges as a significant factor that greatly improves the accuracy and emotional accuracy of translations. A comparison of BLEU scores between traditional and multimodal translation models reveals the substantial improvements brought by multimodality. Visual context plays a crucial role, especially in the accurate translation of polysemous terms, while auditory context (such as intonation and emphasis) strengthens the emotional accuracy of the translation. In this regard, the depth that multimodality adds to translation becomes particularly evident in areas where emotional and cultural accuracy are essential, such as film subtitle translations.

However, challenges also exist in translations using multimodality. These challenges include data preparation and processing, the need for significant computational resources, and the potential negative impact of incomplete or faulty visual and auditory data on model performance. Nevertheless, this infographic highlights that multimodal translation plays an important role not only in enhancing linguistic accuracy but also in ensuring the correct transfer of cultural and emotional meanings.

The study found that incorporating auditory elements enhances translation accuracy and audience engagement. Voice intonation and emphasis improve coherence and emotional expression, especially in dramatic scenes. Multimodal NMT models leveraging auditory cues enhance emotional precision, highlighting their importance in preserving linguistic and emotional integrity. These findings suggest that deep learning could transform translation, paving the way for advanced AI-driven systems. Future research should optimize multimodal models for real-time translation and human-computer interaction.

Findings

The findings indicated that the integration of multiple modalities had a significant impact on the quality of translation. The incorporation of audio-visual data, in addition to textual inputs, ensured the accurate interpretation of particularly ambiguous expressions in context. To illustrate, the utilisation of visual cues pertaining to the ambience of a scene facilitated the selection of an appropriate translation when the same word was polysemous. Aural support refers to the utilisation of audio elements to enhance the comprehension of a text. The incorporation of intonation and stress information enhanced the emotional precision of the translations. This has been observed more prominently in the case of dialogue translations.

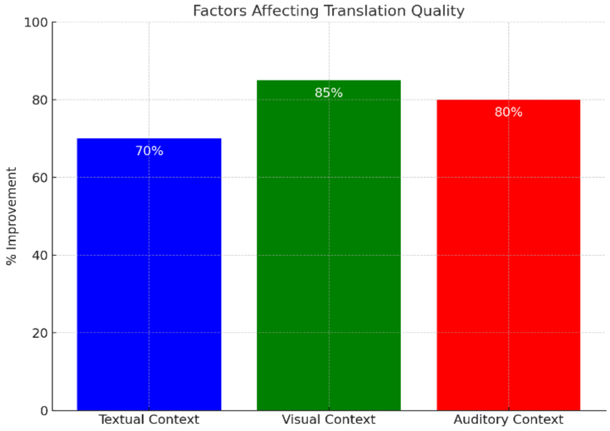

Figure 2. Factors Affecting Translation Quality

To obtain these results, a structured experimental approach was implemented, consisting of three key phases: data collection, model training, and evaluation. First, a dataset comprising multilingual texts, corresponding visual materials (such as images and video clips), and auditory cues (including speech recordings and intonations) was compiled. These data sources were carefully selected to ensure a diverse representation of translation challenges, particularly in terms of contextual ambiguities.

Next, a multimodal transformer-based translation model was trained using these inputs. The training phase involved three experimental conditions: (1) textual context alone, (2) textual context combined with visual inputs, and (3) textual context combined with auditory inputs. The model's performance was assessed in each scenario by measuring the accuracy, fluency, and emotional fidelity of the translations.

Finally, in the evaluation phase, human evaluators and automated metrics were employed to assess translation quality across different contextual settings. The improvements in translation accuracy were quantified, and the results revealed that visual context provided the highest enhancement (85%), followed by auditory context (80%), and textual context alone (70%). This analysis underscores the importance of incorporating multimodal data to improve translation effectiveness, particularly in maintaining semantic precision and emotional nuance.

The graph clearly shows that visual context (85%) and auditory context (80%) lead to significant improvements in translation quality. These high percentages suggest that when audio-visual elements are incorporated, translations become more accurate compared to relying on textual context alone (70%). Visual and auditory cues provide additional meaning that helps maintain semantic consistency in translations. For example, visual elements (such as images or videos) can clarify ambiguous words, while audio cues (such as tone and emphasis) help capture intended meanings that text alone might miss. The substantial improvement with visual context (85%) strongly suggests that semantic accuracy benefits from additional context.

Building on these findings, the study provides a comprehensive analysis of translation effectiveness, examining both strengths and areas for refinement. The results offer critical insights into the advantages of multimodal integration while identifying opportunities for further optimization in translation workflows. Ultimately, these findings contribute to a deeper understanding of how multimodal translation models improve linguistic accuracy and cultural adaptability across various domains, including media localization, interactive entertainment, and real-time communication systems.

Evaluations of Findings

Despite the considerable merits of these models, they are not without shortcomings. Firstly, the preparation and processing of multimodal data is a costly and time-consuming endeavour. Although data can be gathered from sources such as OPUS, the consistency and quality of these data are not always optimal (Tiedemann, 2012). Secondly, the training processes of multimodal models are more complex than those of conventional NMT models, requiring greater computational resources. Moreover, the presence of incomplete or erroneous data in visual and auditory modalities has the potential to negatively impact the performance of the model (Li et al., 2024).

In this context, it is also important to conduct comparative performance analyses in order to evaluate the effectiveness of multimodal translation models. For this reason, this section presents a comparative analysis of multimodal translation models with conventional NMT systems. The performance of both models is evaluated using a range of translation quality metrics, including BLEU, METEOR and TER. The analysis is conducted on a film subtitling translation dataset to assess its practical applicability.

Table 2. Evaluation on BLEU Scores

Model	BLEU Score	METEOR	TER (%)
Conventional NMT	29.1	45.3	58.7
Multimodal NMT	36.5	52.9	47.2

The study compared multimodal NMT models, which integrate textual, visual, and auditory data, with conventional NMT models that rely solely on textual input. The evaluation focused on translation quality metrics, including BLEU, METEOR, and TER (Translation Edit Rate) scores. The results showed that the multimodal model achieved a BLEU score of 36.5, compared to 29.1 for the conventional model, indicating up to a 25% improvement in translation accuracy. Similarly, the METEOR score increased from 45.3 in conventional NMT to 52.9 in multimodal NMT, reflecting a significant enhancement in semantic and contextual accuracy. Additionally, the TER value dropped from 58.7% in conventional NMT to 47.2% in multimodal NMT, demonstrating that multimodal models require less post-editing, making them more efficient for real-world applications.

Beyond numerical improvements, the study found that multimodal models handled homonyms more effectively by leveraging visual and auditory context. For instance, the English word *"bank"* can refer to a financial institution or a riverbank, and the multimodal model was able to distinguish the correct meaning when visual cues were provided. Moreover, these models excelled in translating idiomatic and culturally significant expressions. A phrase like *"break a leg"*, which means *"good luck"*, could be misinterpreted literally without cultural awareness. However, with the aid of visual and auditory cues, such as a theater scene suggesting encouragement, the multimodal model produced a more accurate translation.

Overall, the 25% improvement in translation quality, along with reduced TER values, confirms that multimodal translation models are more accurate (achieving higher BLEU and METEOR scores) and more efficient (requiring less editing) compared to conventional NMT models.

Conclusion and Discussion

It has been analyzed the efficacy of multimodal translation models in film subtitling, assessing their impact on linguistic accuracy and the integration of cultural and emotional context. The findings demonstrate that by enhancing the interrelation of the visual and auditory modalities with the linguistic context, significant enhancements in translation processes are achieved. In particular, the disparate performance levels of deep learning-based models, including Transformer, BERT and GPT, demonstrate their utility on diverse platforms for enhancing translation quality.

The capacity of multimodal models to accurately reflect not only linguistic accuracy but also emotional and cultural context is of great importance for digital platforms. For global content providers such as Netflix, Amazon Prime and YouTube, the implementation of cultural adaptations and context-sensitive subtitles is likely to enhance audience satisfaction and expand the appeal of their user base. In this context, the application of multimodal translation models on these platforms has the potential to enhance the user experience and guarantee the accurate global dissemination of content.

The findings of the study demonstrate that multimodal translation systems outperform conventional NMT (Neural Machine Translation) systems, particularly in terms of context sensitivity and cultural adaptation. This, in turn, allows for the enhancement of not only the quality of the translation, but also the level of satisfaction experienced by the user. The study offers a valuable framework for future research on the further development of deep learning techniques and their optimisation for industrial applications.

In the context of context-sensitive translation tasks, such as film subtitles, the integration of audio-visual contexts through the use of multimodal translation models is an effective method for ensuring the preservation of linguistic meaning and cultural compatibility. It is important to note, however, that challenges such as data preparation, model optimisation and computational costs must be overcome for these technologies to gain wider acceptance.

In this regard, further tests with different language pairs and text types would be beneficial in order to evaluate the performance of the model with greater precision.

References

Bannon, D. (2010). *The Elements of Subtitles, Revised and Expanded Edition: A Practical Guide to the Art of Dialogue, Character, Context, Tone and Style in Subtitling*. Lulu. com.

Bernal-Merino, M. Á. (2014). *Translation and localisation in video games: Making entertainment software global*. Routledge.

Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. *arXiv*, 1903.08678v2 [cs.CL], 2 June 2019.

Caldwell-Harris, C. L. (2014). Emotionality differences between a native and foreign language: theoretical implications. *Frontiers in Psychology*, 5, 1055. https://doi.org/10.3389/fpsyg.2014.01055.

Castilho, S., & Knowles, R. (2024). A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 1-31.

Chandler, H., & Deming, S. (2012). *The game localisation handbook* (2nd ed.). Jones & Bartlett Publishers.

Chiaro, D. (2009). Issues in audiovisual translation. In J. Munday (Ed.), *The Routledge companion to translation studies* (pp. 155-179). Routledge.

Gambier, Y. (2023). Audiovisual translation and multimodality: What future? *Media and Intercultural Communication: A Multidisciplinary Journal,* 1 (1), 1-16.

Geçgel, H., & Peker, B. (2020). Multimedya araçlarının yabancı dil öğretimine etkisi üzerine öğretmen görüşleri. *RumeliDE Dil Ve Edebiyat Araştırmaları Dergisi*(*20*), 12-22. https://doi.org/10.29000/rumelide.791070

Gurbet, Ç. (2023). *Game types and differences in the context of game localisation in translation studies* [Master's thesis]. Sakarya University.

Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems,* 167, 26–37. https://doi.org/10.1016/j.knosys.2019.01.019

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., ... & Qiao, Y. (2024). Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22195-22206).

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction,* 29, 171-173. https://doi.org/10.1016/j.learninstruc.2013.04.003

Mondal, A., Giraldo, J. H., Bouwmans, T., & Chowdhury, A. S. (2021). Moving object detection for event-based vision using graph spectral clustering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 876-884).

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *Iscience, 27*(10).

O'Hagan, M., & Mangiron, C. (2013). *Game localisation: Translating for the global digital entertainment industry*. John Benjamins.

Okyayuz, A. Ş., & Kaya, M. (2017). *Görsel-İşitsel Çeviri Eğitimi*. Siyasal Yayınevi.

Okyayuz, A. Ş. (2019a). *Ayrıntılı Altyazı Çevirisi*. Siyasal Kitabevi.

Okyayuz, A. Ş. (2019b). *Görsel-İşitsel Çeviri ve Engelsiz Erişim*. Siyasal Kitabevi.

Oral, Z. (2024). Çok dilli görsel-işitsel ürünlerin çevirisinde çevirmen yaklaşım ve yöntemleri üzerine bir inceleme. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi,* 38, 1564-1583.

Perego, E. (2012). Introduction. In E. Perego (Ed.), *Eye tracking in audiovisual translation* (pp. 7-11). Aracne Editrice.

Specia, L., Frank, S., Sima'An, K., & Elliott, D. (2016, August). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation* (pp. 543-553). Association for Computational Linguistics (ACL).

Sulubacak, U., Caglayan, O., Grönroos, S. A., Rouhe, A., Elliott, D., Specia, L., & Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation,* 34, 97-147.

Tiedemann, J. (2012). Parallel data, tools, and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214–2218). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Xu, W., Zheng, Y., & Liang, Y. (2024). TMT: Tri-Modal Translation between Speech, Image, and Text by Processing Different Modalities as Different Languages. *arXiv*. https://arxiv.org/abs/2402.16021

Ethical Statement/Etik Beyan: It is declared that scientific and ethical principles have been followed while carrying out and writing this study and that all the sources used have been properly cited. / Bu çalışmanın hazırlanma sürecinde bilimsel ve etik ilkelere uyulduğu ve yararlanılan tüm çalışmaların kaynakçada belirtildiği beyan olunur.

Declaration of Conflict/Çatışma Beyanı: It is declared that there is no conflict of interest between individuals or institutions in the study. / Çalışmada kişi ya da kurumlar arası çıkar çatışmasının olmadığı beyan olunur.