International Multidisciplinary Research Journal
ISSN: 2424-7073
Available Online at https://imrjournal.info/
Volume:1V; November 2026; Page No.1-11
DOI: https://doi.org/10.47722/imrj.2001.83

IMR JOURNAL CANADA

Research Article

AUDIO DESCRIPTION AND ARTIFICIAL INTELLIGENCE: NEW PERSPECTIVES ON INTERSEMIOTIC TRANSLATION

Dr. María J. Valero Gisbert

Associate Professor, University of Parma, Italy

ARTICLE INFO

Article History:

Received 15.07.2026
Accepted 15.10.2026
Published 25.11.2026

Keywords:

audio description,
AI,
Intersemiotic Translation,
Accessibility.

ABSTRACT

Technological advances are transforming the field of translation, raising fundamental questions about the evolving role of human translators and the potential impact of artificial intelligence (AI). This presentation examines the translation of images into verbal language, with particular attention to audio description as a form of intersemiotic translation. The discussion begins by exploring the need to translate visual content and identifying its target audiences. These include, first, people with visual impairments who rely on verbal mediation to access visual information and, second, individuals from different sociocultural backgrounds who may not share the cultural knowledge embedded in visual representations. Within this framework, audio description is conceptualized as an interpersonal and intercultural bridge that facilitates understanding. The presentation proposes the translation of a visual text as a case study and investigates the potential contribution of AI to this process. The methodological approach addresses key theoretical questions concerning the nature of images, the motivations for translating them, and the characteristics and needs of intended recipients. Particular emphasis is placed on AI's ability to generate image translations, model audience profiles, and adapt outputs according to explicit instructions designed for specific user groups.These issues are examined through a series of practical experiments involving AI-assisted image translation. The analysis evaluates audio descriptions and their communicative effectiveness. The presentation concludes with a set of preliminary findings that contribute to current debates on accessibility, intersemiotic translation, and AI supporting human translation.

Copyright© The Author(s) 2025. Open Access under the Creative Commons Attribution 4.0 International License (CC BY 4.0). This license permits use, sharing, adaptation, distribution, and reproduction in any medium or format, giving proper credit to the original author, including a link to the license, and indicating if any changes were made. To view a copy of the licence, please visit https://creativecommons.org/licenses/by/4.0/

1. Introduction

The advancement of technology compels us to confront new challenges in the field of translation. We are witnessing an unprecedented development that calls into question the role of the human translator or at least challenges the extent of their necessity. We frequently encounter information about various platforms and websites that promote machine translation and emphasize the 'ease' with which communication can take place in any language.

One of the current limitations of neural machine translation (NMT) systems is their inability to adequately render the connotations embedded in conversational texts, such as colloquial dialogue in audiovisual products (Luong & Manning, 2015; Valero Gisbert, 2024). In the past, they often sought to "fill in" the meaning of phrases that appeared incomprehensible to them, precisely because of the absence of extratextual information, world knowledge, lived experience, and interpersonal context. This situation has been evolving with the emergence and rapid advancement of artificial intelligence, particularly with the latest generation of generative Large Language Models, such as ChatGPT-4, Claude 3.5, and Gemini 2, among others.

These technological developments now enable the generation of images from textual descriptions, raising important questions about the type of images produced, the sociocultural realities in which they are grounded, and the models they ultimately reflect (issues that are currently the focus of our work). More importantly, and this constitutes the central proposal of this research, AI also makes possible to produce descriptions derived from visual input.

In this presentation, we focus on this objective, namely, reflecting on AI-generated suggestions based on an audio descriptive text, as well as on the resulting text produced through the generation of descriptions from images via the training of a chatbot we have called *ArtVoice* (Dini et al., forthcoming), designed with specific target audiences in mind. The thesis we advance is that AI can become a key tool, provided that it receives appropriate instructions, for the translation of static, two-dimensional images (as opposed to moving images, such as video).

To this end, we will first outline the relevance and necessity of this line of research, followed by the theoretical framework that underpins it. We will then present an empirical study and, finally, offer a set of concluding remarks and considerations.

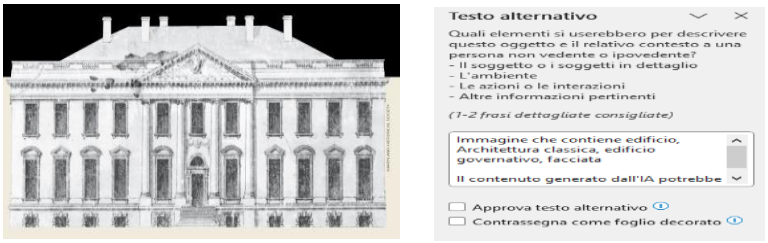
2.The need for Image translation

At the outset, it is necessary to establish several key considerations concerning the translation of visual content into language. These include the motivations underlying image translation, its communicative and accessibility-related functions, the audiences for whom it is intended, and the outputs generated by AI systems that convert visual stimuli into verbal representations.

These questions provide the conceptual foundation for our discussion. Within this context, we situate our analysis in the field of accessible intersemiotic translation, understood as the process of transforming visual information into verbal forms that can be accessed and interpreted by diverse audiences.

At this stage, we can already observe how contemporary AI systems generate alternative textual representations of even relatively simple photographs, often tailoring their outputs to specific user groups and communicative purposes. These examples, as we will see below, illustrate both the opportunities and the challenges associated with AI-mediated image translation and constitute the starting point for our investigation.

Despite the still rudimentary nature of such proposals, their most immediate audience consists of people who are blind or have low vision. This, in fact, provides the answer to the first question: why translate images? The primary purpose is to enable these communities to access iconic content, in a broad sense, the material and immaterial cultural heritage of a society. We are therefore referring to audio description, a form of accessible intersemiotic translation.



Source: Printed version of the White House brochure / Word Technology, Office Accessibility Service

The Word tool prompts us to reflect on a specific target audience through the following question: "What elements would be used to describe this object and its context to a person who is blind or has low vision?" This prompt is followed by four aspects that should be considered: the subject/object in question, the setting, the actions or interactions, and, finally, any other relevant information. In addition, a suggestion about the extension appears in parentheses, namely, "1–2 detailed sentences are recommended".

A text box then displays the proposed description; in this case: "image that may contain building, classical architecture, government building, façade", which clearly indicates that the description has been generated automatically. The user is then asked either to approve the description or to mark the image as decorative (i.e., purely aesthetic). In the latter case, the tool provides the following notice: "For content marked as decorative, descriptions will not be displayed by screen-reading utilities." In other words, no descriptive text will be included.

3. Accessible Intersemiotic Translation

As mentioned above, we are dealing with a type of translation that, in Jakobsonian (1959) terms, corresponds to intersemiotic translation, whereby a text, understood in its broadest sense, is translated into a different semiotic code. In our proposal, we begin with two-dimensional iconic content, such as photographs, reproductions, or pictorial works, and translate it through the use of a verbal code.

For the purposes of the present analysis, intersemiotic translation and visual literacy are the two core concepts. The additional notions discussed below are retained as supporting concepts insofar as they help explain how visual information is identified, selected and reformulated verbally.

These types of works are typically displayed in museum settings, and some guidelines (ITC 2000, RNIB and VocalEyes 2003, AENOR 2005, Gonant and Morisset 2008, ADC 2013, Giansante 2015, Remael et al. 2015), have therefore been developed to support this practice. Audio description in museums has evolved from the audio description of audiovisual products, an area in which significant advances have been made. Today, audio description can be accessed on platforms such as Netflix, Amazon Prime, and Disney+, as well as through applications like MovieReading[1], although the quantity remains limited and overall quality could still be improved. Alongside other, more widely known forms of audiovisual translation, such as subtitling and dubbing, audio description represents a modality with considerable potential, one that requires specialized professionals depending on the context of application (museums, parks, audiovisual materials, etc.).

[1] https://www.moviereading.com/

4. The Role of Images in Communication

Artificial Intelligence is transforming the ways in which we communicate, while images increasingly occupy a central and dominant role. This context prompts us to reflect on the competencies required, if any, to "read" the visual universe in which we are immersed. We are referring to visual literacy (VL), the term most widely used in English-language scholarship.

With regard to terminology, the concept was originally coined by J. Debes (1969). Since then, the range of terms used to refer to this competence has expanded considerably, Kedra (2018: 4) lists several words to refer to visual communication as:

> *"Visual literacy, digital visual literacy (Spalter & van Dam, 2008), image literacy (Emanuel, Baker & Challons-Lipton, 2014, 2016), media literacy, multimodal literacy (Serafini, 2014), multiliteracies, information design (Pettersson, 2018), visual competency (Müller, 2008) and visual competencies (Griffin, 2008). Many of these terms overlap, but some are discussed as opposite to each other (Chauvin, 2003; Griffin, 2008). (Kedra, 2018, p. 4)"*

Despite its growing relevance, this author argues that its presence within academic curricula remains limited.

Castillejo and Gutiérrez Manjón (2023: 38) further contend that the development of this competence is essential and raise the issue of interpretation, namely, "*hasta qué punto se interpreta adecuadamente aquello que vemos, ya que esta competencia no solo afecta a los profesionales de la información sino que también compete a todas aquellas personas que consumen habitualmente imágenes*[2]."

Indeed, Kedra (2018: 9) argues that merely seeing is not enough to understand what we see, and that specific competencies must be developed. As she explained, "*Vision is a natural act for those who are able to see, influencing socialization, learning, and the perception of the world. However, effective visual communication requires a trained gaze, since visual literacy is not acquired solely through exposure to images."* In other words, the act of seeing does not automatically result in knowledge or comprehension.

5. Competencies for Visual Translation

Further considerations concern the competencies required for the translation of images, the extent to which these competencies are comparable to those associated with linguistic literacy, and the possibility of developing the skills necessary to interpret and "read" images as meaningful communicative texts. As Marzal (2020: 6) explains: "*Visual literacy aims at the development of a visual metalanguage, with its iconic semiotics to support a visual semantics and thereby an appropriate "alphabet" for iconic reading, with its own grammar for visual discourse."*, and later emphasizes that one of the primary objectives is to develop skills and competencies to "*analyze the contextual, cultural, ethical, aesthetic, intellectual, and technical factors involved in the use and production of audiovisual materials."*

Finally, we highlight a key specification that defines visual literacy as the ability to understand images that convey culture or any other form of content embedded within sociocultural knowledge or that contributes to the identity of a given community. This aspect

[2] Our translation: "*to what extent what we see is adequately interpreted, since this competence concerns not only information professionals but also all individuals who regularly consume visual content.*"

becomes particularly relevant in the context of foreign language learning, although it does not constitute the focus of the present paper.

The VL competence can be learned through visual grammar (Kress and van Leeuwen, 1996). As these authors teach us, images, like verbal language, have a structure. Understanding the elements that compose it helps us interpret them accurately. Therefore, image analysis involves examining its compositional structure, framing and perspective, lighting, spatial relationships among elements, symbolic features and their cultural significance, as well as the visual style through which the image is represented.

6. Towards Audio Description: From Image to Text

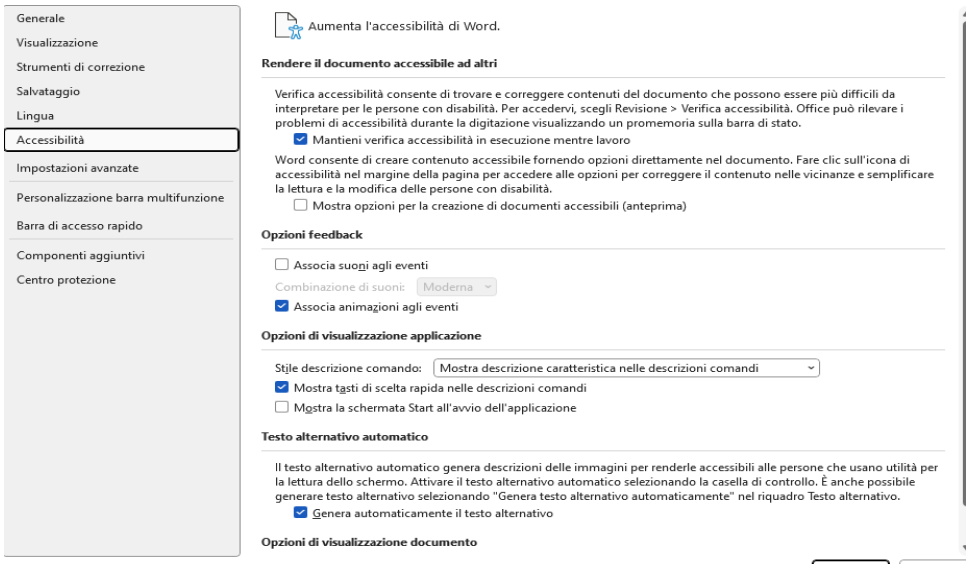
Before moving on to the empirical research and examining how AI can serve as a tool for the translator, it is important to provide some notes on the preparation**,** required to approach the translation of an image.

As we have explained, semiotic analysis, which needs to be clarified here, since we have so far only discussed visual literacy, is of fundamental importance for uncovering the meaning conveyed by the elements that compose an image, whether two-dimensional or three-dimensional, static or moving. Without a semiotic study of the compositional elements, could be not possible to fully understand their meaning, a task that is also achieved through the interactive relationships established among the various components.

For the purposes of this presentation, we skip the intermediate phases, such as documentation, creation of a glossary of terms, and common verbs aimed at a non-specialist audience, as recommended by the guidelines, and proceed directly to presenting our audio description, produced based on the semiotic analysis of the work.

7. AI in Image Translation

Our research explores the ways in which artificial intelligence can assist describer-translators in the translation of images. The output generated by the AI system in the *Accessibility* section of Word has been presented previously:



Source: Word Technology, Office Accessibility Service

Although current AI systems remain relatively limited in their interpretative capabilities, they frequently identify and report visual elements in an unpredictable manner. For example, in an image showing a pyramid on a volcano, the AI focused exclusively on the natural environment while failing to mention the archaeological structure. However, these shortcomings are not insurmountable. Through targeted training and optimization, AI can be guided to identify and prioritize salient information, such as the archaeological significance of the site in this case.

8. Empirical Research

From this point onward, we present the study, which involved submitting our descriptions to AI systems. Initially, the experiment focused on two artworks: an engraving depicting the White House (previously shown) and a portrait of Abraham Lincoln (omitted from the present discussion due to space limitations). The analysis presented below therefore concentrates on one image and three textual products: the output generated by ArtVoice, the output generated by Unidescription AI, and the final human-validated description produced after comparison with the authors' description.

As noted above, the study was conducted using two AI models. The first, named *ArtVoice*, is a chatbot trained according to specific guidelines for the audio description of museum artworks. Although still experimental, the system continues to be refined through the incorporation of additional inputs and training data aimed at improving its accuracy and reliability. ArtVoice was developed within an interdisciplinary research project involving linguists, translation scholars, and experts in computer science and digital technologies. While further development and the integration of additional disciplinary perspectives remain necessary, the tool has already demonstrated practical usefulness. Moreover, it is capable of generating descriptions in multiple languages.

The second tool emerged from participation in a large-scale project dedicated to the description of natural park guides and hosted on the platform *Unidescription*[3]. This initiative is promoted by the Department of Communication at the University of Hawai'i and seeks to explore the application of AI technologies to accessible description practices.

The following section outlines both the process and the results of producing an audio description of a two-dimensional engraving depicting the façade of the White House. The study is exploratory rather than a large-scale performance evaluation. The same image and the same task request were used for the two AI systems, and one output from each system is reported. The exact historical wording of the prompt was not archived in the manuscript; what can be established from the procedure is that both systems were asked to generate an audio description of the same engraving. The three descriptions are compared using five criteria derived from the aims already discussed in the paper: (1) accuracy of visually verifiable information; (2) completeness and selection of salient information; (3) clarity of wording and terminology; (4) visual hierarchy and coherent sequencing; and (5) suitability for a non-specialist blind or low-vision audience. Evaluation was qualitative and author-led; no independent double coding was

[3] https://unidescription.org/

carried out and, consequently, inter-rater reliability was not calculated. These constraints are acknowledged as part of the pilot nature of the study.

8.1 ArtVoice - AD

After preparing our own description, we conducted an experiment in which we asked the platforms introduced above to generate descriptions of the same image. The purpose was to compare the AI-generated outputs with our own description and evaluate their respective approaches to image translation.

On the one hand, we prompted the AI to generate a description of the engraving. The resulting output was structured into a general introductory section, followed by the identification and description of the architectural elements depicted in the image, as well as a discussion of the artistic technique employed.

Table 1: AD via ArtVoice

The image shows an architectural drawing of a classical, symmetrical building with a rectangular structure and a hipped roof. The façade is divided into three main sections, with a central entrance emphasized by a portico with columns and a decorated triangular pediment. **Highlighted architectural elements:** **Symmetrical façade:** The building features a balanced layout, with an equal number of windows on both sides of the central entrance. **Rectangular windows:** Arranged on two levels, with decorative frames. Those on the lower floor have arched tops. **Main entrance:** Located at the center, it is emphasized by a doorway with a semicircular arch and a small balcony above it. **Classical columns:** These support the triangular pediment, following a style inspired by Greco-Roman architecture. **Decorated pediment:** It features a relief or sculpture at its center, although the details are not entirely clear. **Roof with chimneys:** Several chimneys can be seen emerging from the roof, indicating the presence of multiple heating sources within the building. **Technique and style of the drawing:** The image appears to be an architectural sketch or an old illustration. The texture and use of shading suggest that it was created using ink or engraving techniques in black and white. The background is dark in the upper part, which highlights the building, while the lower part is lighter in tone. This design evokes palaces or residences in the neoclassical style, with a clear influence from European architecture of the 18th and 19th centuries.

As can be observed, the AI-generated output is occasionally only approximate, may not accurately capture or describe forms, and at times provides interpretative accounts rather than objective descriptions.

8.2Unidescription AI

The identical prompt was then submitted to the second AI tool to examine the consistency and differences between the responses generated by the two systems.

Table 2: AD via Unidescription

This image is a detailed architectural drawing of a large, stately building. The building is symmetrical, with a central entrance flanked by columns and an ornate triangular pediment above. The structure has two main floors with evenly spaced windows on each level. The roof is pitched with several small chimneys visible. The style is grand and classical, reflecting a historic or governmental architecture. The overall look is elegant and formal.

As an initial assessment, the description remains relatively general and includes technical terms (such as *pediment*) that may not be readily understood by a general audience.

It should be emphasized that this output represents only a preliminary stage in our experimentation with UniDescription. The platform forms part of a considerably more complex ecosystem of tools that supports increasingly sophisticated and personalized approaches to AD, including feedback from experts and non-experts, as well as blind users with different levels of experience. A detailed examination of this system falls beyond the scope of the present study. What is relevant here is the progression enabled by AI: from relatively simple initial descriptions to increasingly refined, user-oriented forms of AD currently under development.

For the purposes of the present study, we remain at this preliminary stage. Drawing on the description developed by the authors and comparing it with the AI-generated outputs, a revised description was then produced following the three-part structure used by UniDescription: an Introductory Description (corresponding to the object label), a Synthetic Description, and a Detailed Description. The resulting version is presented below.

Table 3: final AD

DESCRIBING: A black-and-white horizontal drawing reproduction of a design of the front of the White House. **SYNOPSIS:** The white stone building has a rectangular shape. The upper background is completely black. There are faint white areas, especially near the lower edges. A central section of the facade slightly protrudes and includes pediment columns that support a triangular section at the top (pediment). The design is symmetrical, with four windows on each side of the entrance. **IN-DEPTH DESCRIPTION:** The main entrance is framed by four pillars/columns. Above the entrance, there is a triangular pediment. A dentil pattern is along the edge of the pediment (a small block used as a repeating ornament in the cornices). In the centre of the pediment, a relief or sculptural element is present**.** The main entrance is marked by an arched doorway, flanked by two additional curved-molding windows at ground level. Above the entrance, there are three tall, narrow windows aligned vertically with the doorway. Flanking the central entrance on the ground floor, there are eight rectangular windows, each adorned with alternating triangular and curved molding above them. On the upper level, there are also eight windows that align with those below. The roof is pitched and extends across the width of the building, featuring six vertical chimneys.

As can be observed, although the term *pediment* has been retained, it is accompanied by an explanatory definition within the description. This strategy aims to facilitate the reader's construction of a mental image of the building's architectural elements while simultaneously introducing relevant specialist terminology that may enhance their understanding of future descriptions. In this way, the description remains accessible to non-specialist audiences while preserving the accuracy of architectural vocabulary.

The comparison also clarifies the complementary contribution of human and AI work in this pilot study. The AI systems rapidly identify candidate visual features and generate alternative organizations of the information, but human intervention is required to verify details against the image, distinguish observation from interpretation, establish a meaningful hierarchy, explain specialist terminology, and adapt the final sequence to the intended audience. On this basis, the study proposes a four-stage collaboration model: AI-generated

draft; human verification of visual accuracy; human selection, hierarchy, and audience adaptation; and, in a subsequent research phase, validation with blind and low-vision users.

From this point onwards, the experimentation continued by introducing the Unidescription platform, which has implemented a tool in the form of "guide dogs". This tool first evaluates the description provided by the user and subsequently generates structured feedback divided into two main categories: strengths and areas for improvement, including aspects such as terminology and clarity, narrative flow, and visual hierarchy.

The platform also enables the generation of alternative versions of the description, each adopting a different focus and tailored to specific communicative purposes, such as *Enhancing Clarity and Consistency, Focusing on Visual Detailing, or Structuring for Simplicity and Accessibility.*

This aspect will not be explored in depth here, as it requires further analysis beyond the scope of this study. Such an investigation would need to examine both the linguistic choices involved and the mental images generated by the verbal representations used to convey visual information.

Nevertheless, it is evident that the greater the precision and quality of the initial description, the more relevant and effective the suggestions provided by the tool become. These annotations are valuable for several reasons. Firstly, they encourage critical reflection on the descriptive process by helping to identify significant elements that may have been overlooked and suggesting possible improvements. Secondly, they support decisions regarding the sequencing of information, vocabulary selection, and other strategies aimed at progressively and coherently constructing a mental image for the audience. Most importantly, these annotations provide a foundation for future reception studies designed to evaluate and validate the descriptive choices made.

9. Conclusion

As a preliminary conclusion, although these tools remain subject to certain limitations, they are increasingly emerging as valuable, and potentially indispensable, resources for human translators engaged in the translation of visual content. At the current stage of the research, we continue to refine the informational and instructional components of the process through iterative adjustments, while also incorporating theoretical insights from related disciplines that may contribute to and strengthen the achievement of our research objectives.

These findings should be read as those of a pilot study rather than as evidence of general AI performance. Its main contribution is to show how specialized instructions and iterative human guidance can substantially shape AI-generated descriptions. The small sample, the experimental nature of the tools and the absence of a formal reception study limit broader claims. Future work will extend the dataset, document the interaction process more systematically and test the resulting descriptions with blind and low-vision audiences.

References

AENOR. Norma UNE: 153020. (2005). Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías. Madrid: AENOR.

Audio Description Coalition (ADC). (2013). Standards for Audio Description and Code of Professional Conduct for Describers. Retrieved April 15, 2026, from https://audiodescriptionsolutions.com/the-standards

Debes, J. (1969). The loom of Visual Literacy: an overview. Audiovisual Instruction 14(8), 25-27.

Dini, S., Valero Gisbert, M.J., & Ludovico L.A (forthcoming). Customizing AI Language Models for the Production of Artwork Audio Descriptions: The *ArtVoice* Project.

Giansante, L. (2015). Writing verbal description audio tours. Retrieved April 15, 2026, from http://www.artbeyondsight.org/mei/verbal-description-training/writing-verbal-description-for-audio-guides

Gonant F., & Morisset, L. (2008). La charte de l'audiodescription. Paris, Ministère des Affaires Sociales et de la Santé.

Gutiérrez-Manjón, S., & Castillejo-de-Hoces, B. (2023). El futuro de la alfabetización visual: Evaluación de la detección de imágenes generadas por inteligencia artificial. *Hipertext.net: Revista Académica sobre Documentación Digital y Comunicación Interactiva 26*, 37-46. Retrieved April 15, 2026, from https://doi.org/10.31009/hipertext.net.2023.i26.06

Independent Television Commission, 2000. ITC Guidance On Standards for Audio Description. Retrieved April 15, 2026, from http://www.ofcom.org.uk/static/archive/itc/uploads/ITC_Guidance_On_Standards_for_Audio_Description.doc

Jakobson, R. (1959). *On Linguistic Aspects of Translation*. In R. Brower (ed.), On Translation. Cambridge Mass, Harvard University Press, 232-239.

Kędra, J. (2018). What does it mean to be visually literate? Examination of visual literacy definitions in a context of higher education. Journal of Visual Literacy. Retrieved April 15, 2026, from https://jyx.jyu.fi/bitstream/handle/123456789/63132/articlejkvorall.pdf?isAllowed=y&sequence=1

Kress G., & van Leeuwen, T. (1996). *Reading images. The grammar of visual design*. London: Routledge.

Luong, M.T., & Manning, C.D. (2015). Stanford neural machine translation systems forspoken language domains. Proceedings of the International Workshop on Spoken Language Translation, 76-79.

Marzal, M.A. (2020). "A taxonomic proposal for multiliteracies and their competences". *Profesional de la información*, *29*(4), 1-16. Retrieved April 15, 2026, from *https://doi.org/10.3145/epi.2020.jul.35*

Remael, A., Reviers, N., & Vercauteren, G. (Eds). (2015). *Pictures painted in words: ADLAB audio description guidelines*. EUT. Retrieved April 15, 2026, from http://www.openstarts.units.it/dspace/handle/10077/11838.

RNIB, & VocalEyes. (2003). Talking images guide. Museums, galleries and heritage sites: Improving access for blind and partially sighted people. Retrieved April 15, 2026, from http://audiodescription.co.uk/uploads/general/Talking_Images_Guide_-_PDF_File_5.pdf

Valero Gisbert, M.J. (2024). Traducir la oralidad en el texto audiovisual con sistemas de traducción automática neuronal (TA). In L. Vilei et al. (ed.), L' italiano in Connessione: Studi Di Linguistica, Traduzione e Didattica Nell'era Digitale, Berlin/Alemania, Peter Lang, 601-613.

How to cite this article:
Gisbert M. J. V. (2026) 'Audio Description and Artificial Intelligence: New Perspectives on Intersemiotic Translation, *International Multidisciplinary Research Journal*, Volume:1V; November 2026; Page 1-11. DOI: https://doi.org/10.47722/imrj.2001.83