https://doi.org/10.5007/2175-7968.2021.e75483

BRIDGING THE GAP BETWEEN MACHINE TRANSLATION OUTPUT AND IMAGES IN MULTIMODAL DOCUMENTS

Thiago Blanch Pires[1]
Augusto Velloso dos Santos Espindola[1]
[1]Universidade de Brasília, Brasília, Distrito Federal, Brasil

Abstract: The aim of this article is to report on recent findings concerning the use of Google Translate outputs in multimodal contexts. Development and evaluation of machine translation often focus on verbal mode, but accounts by the area on the exploration of text-image relations in multimodal documents translated automatically are rare. Thus, this work seeks to describe just *what* are such relations and *how* to describe them. To do so this investigation explores the problem through an interdisciplinary interface, involving Machine Translation and Multimodality to analyze some examples from the *Wikihow* website; and then it reports on recent investigation on suitable tools and methods to properly annotate these issues from within a long-term purpose to assemble a *corpus*. Finally, this article provides a discussion on the findings, including some limitations and perspectives for future research.
Keywords: Multimodality; Machine Translation; Machine Translation Output Classification; Intersemiotic Texture; Intersemiotic Mismatches

APROXIMANDO RESULTADOS DE TRADUÇÃO AUTOMÁTICA E IMAGENS EM DOCUMENTOS MULTIMODAIS

Resumo: O objetivo deste artigo é relatar os recentes achados sobre o uso de resultados do Google Tradutor em contextos multimodais. O desenvolvimento e a avaliação da tradução automática geralmente se concentram no modo verbal, mas são raros os relatos da área sobre a exploração das

relações texto-imagem em documentos multimodais traduzidos automaticamente. Assim, este trabalho busca caracterizar o que são tais relações e como descrevê-las. Para tal, esta investigação examina o problema através de uma interface interdisciplinar envolvendo tradução automática e multimodalidade para analisar alguns exemplos do site *Wikihow*; em seguida, este trabalho descreve estudos recentes sobre ferramentas e métodos adequados para a anotação destas questões com o propósito de construir um *corpus* a longo prazo. Finalmente, este artigo fornece uma discussão sobre os achados, incluindo algumas limitações e perspectivas para pesquisas futuras.
doi

Palavras-chave: Multimodalidade; Tradução Automática; Classificação de Resultado de Tradução Automática; Textura Intersemiótica; Incompatibilidades Intersemióticas

1. Introduction

Since the popularization of computers in the 1980s and the widespread use of the internet that started in the 1990s (Hutchins, *Machine Translation: a concise*), there has been a shift both in the way the population uses technology and the way they read (Saçak, 14). On the technology side, one may notice how far we have gone, from the rudimentary use of machine translation, since the Weaver memorandum in 1949 (Hutchins & Somers, 5-6; Hutchins), to (arguably) optimal Neural Machine Translation (NMT) network output, as is the current practice (Melby). Likewise, on the reading side, it is undeniable that it has been changing faster, becoming more hyperlinked and multimodal than it used to be (Mills & Unsworth, 1). Therefore, reading has been mediated by templates and cognitively sophisticated algorithms, a scenario that affects the contemporary reader of the digital era.

In such globalized informational contexts, readers have been increasingly demanding more automatic translation (Quah), for a wider variety of documents containing illustration, videos, infographics, emoticons, and photographs, all working in cohesive orchestration to build a coherent "multimodal document," such as webpages, manuals, and news articles (Bateman, *Multimodality*).

In the past few years, studies conducted within a Machine Translation and Multimodality interface have started to grow. In general, they adopt an engineering perspective, testing the validity of multimodality to improve the precision of machine translation. That is especially carried out by training machine translation systems with visual representations and speech syntheses[1] (Caglayan *et al.*; Caglayan; Calixto & Liu; Heo *et al.*; Hirasawa *et al.*), along with evaluation methods for automatic machine translation, to measure precision.

As the engagement of Translation Studies scholars with machine translation (Baker & Saldanha, 305) is recent, the potential of text-image relations in multimodal texts that were automatically translated is still much left uncharted within the field. What, then, would be the relations between text and image purposefully made in the source multimodal document, when translated automatically? That is precisely the problem that motivates the following research questions:

- *What* text-image relations emerge from multimodal documents translated automatically?
- *How* to describe such text-image relations within the context of machine translation output and multimodality?

The previous questions lead to the following objectives.

1.1 Objectives

This paper aims at answering the research questions precisely by, firstly, providing an explanation of text-image relations in the context of multimodality, especially through the intersemiotic

[1] This perspective is usually called, in the area of Natural Language Processing, as Multimodal Neural Machine Translation (MNMT). They are systems which use "images related to source language sentences as inputs to improve translation quality" (Takushima *et al.*).

texture approach, developed by Liu & O'Halloran, and Machine Translation (MT), more precisely the classification approaches for errors in Machine Translation output (Vilar *et al*.; Kameyama *et al*.) to analyze examples taken from the *Wikihow* website. Secondly, this article aims at reporting on recent investigations of suitable tools and methods to properly tag and annotate those relations to facilitate manual manipulation of a large quantity of data.

A few how-to instruction articles provided by *Wikihow* in English and in their corresponding *Google translation* outputs into Portuguese, are used as object of analysis. Part of this first analysis describe and compares *Evernote, Nimbus Capture,* and *UAM image tools* as most effective non-specialist tools for annotating and tagging text-image.

In section 2, the theoretical framework informs the proposed concept of "intersemiotic mismatches" (Pires, *Ampliando*; Pires, *Multimodality*) "plainCitation" (Pires, Ampliando olhares sobre a tradução automática online : um estudo exploratório de categorias de erros de máquina de tradução gerados em documentos multimodais; Pires, "Multimodality and Evaluation of Machine Translation", as well as some instances of its occurrences. Subsequently, three available tools are examined for tagging and annotating these specific text-image relations.

2. Machine Translation output classification, intersemiotic texture, mismatches, and tools for annotation

The concepts of Machine Translation output classification, intersemiotic texture as cohesive devices, and the phenomena of intersemiotic mismatches are dealt with in the subsequent subsections. The last subsection presents analysis of tools for annotating intersemiotic mismatches generated by errors in machine translation outputs.

2.1 Machine translation output classification

The manual Machine Translation output classification elaborates linguistic categories to classify errors in Machine Translation output. For instance, Vilar *et al*. present a framework to classify MT errors, encompassing five major categories, namely "missing words," "word order," "incorrect words," "unknown words," and "punctuation." This framework is illustrated as follows:

Figure 1: Classification for errors in Machine Translation based on Vilar *et. al*.

Source: the authors (adapted from Pires, *Ampliando,* 88).

As indicated by Vilar *et. al*., "missing words" refer to cases in which a word is missing from sentences produced by Machine Translation. Its subcategories, "content words" and "filler words,"

are separately expected to communicate the meaning of the sentence and to frame the sentence in terms of its grammar (698); however, the meaning is kept unchanged. The subsequent class is identified by the reordering of words and syntactic blocks of words. The contrast between the two levels depends on the exclusive movement of words or in movements of blocks of words while producing the sentences. In terms of local or long range, the differentiation is not absolute, yet it depends on the need to reorder words in a local context (inside a syntactic block) or to reorder the words in another block (698).

Vilar *et. al*.'s "incorrect words" can be distinguished by an MT system that is unable to locate a fitting counterpart for a word. Its first subcategory represents changes in the meaning of the sentence, which in turn may lead the system to process an incorrect disambiguation or a wrong lexical decision (698). The other subcategory of incorrect words is "incorrect forms", which occurs when the MT does not generate the appropriate word form, though the translation of its basic form is correct. The fourth category, "unknown words," can be recognized by words or stems unknown to the system and unseen types of known stems. The last category, "punctuation," is considered a minor issue for machine interpretation assessment (698).

Similarly to Vilar *et. al*, Kameyama *et al*. (194) have also developed categories to catalogue errors in Machine Translation output. The latter authors focus on aspects of error classification in Machine Translation, which extrapolates on the grammatical categorization (in one specific work, it expands on Dorr, 1990's concept of "translation divergences"). Although it was published fifteen years earlier than Vilar *et. al.'s* work, Kameyama *et. al*.'s distinction relies on introducing the concept of "translation mismatches" (within the field of Computational Linguistics) to identify situations in which the grammar of a given language does not set a required distinction between itself and the grammar of another language[2] (194).

[2] For instance, we note countable nouns in English, the definite character and

According to Kameyama *et. al*, there are two important effects of Machine Translation when it comes to relevant mismatches between two languages in relation to their contextual information (194). Firstly, human translation may be forced to draw upon information not expressed in its source segment in such a way that it is only inferable from its context; secondly, sometimes a translation may need to explicit information, when it had been implicit in the source segment (194).

Thus, the joint study of Kameyama *et. al*. on "translation mismatches" regarding context to infer meanings from distinct grammar structures, along with Vilar *et. al's* typology of Machine Translation output errors, presents a relevant standpoint for manually classifying MT types of error when analyzing the phenomenon of intersemiotic mismatches automatically generated by machine translation in multimodal documents.

However, before approaching that phenomenon, the following subsection explores Liu and O'Halloran's text-image cohesive devices.

2.2 Intersemiotic texture

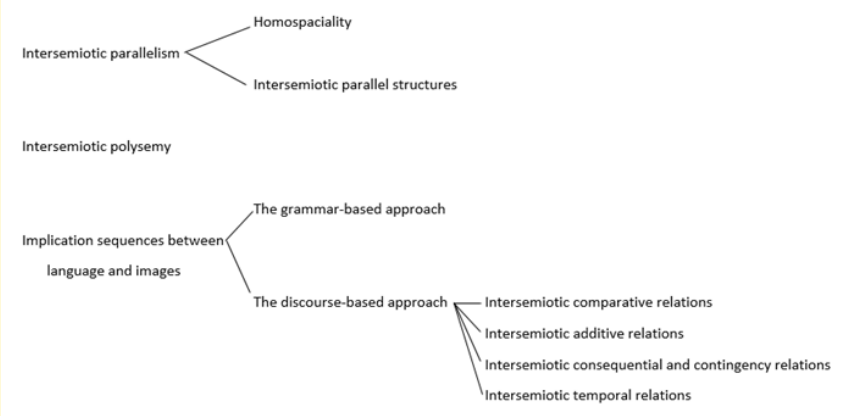
Text-image relationships may take several forms and combinations based on the questions and answers informing the object of a given study. According to Bateman (*Text and images*), different definitions for different text-image relationships might be influenced by different categorizations and systems; moreover, each answer will contextualize the criteria for categorization (44).

In that sense, approaches that inform potential frameworks for analysis of cohesion as a semiotic resource are notably valid in attempts to contribute to define text-image relationships. Among them, Liu & O'Halloran's work, "Intersemiotic Texture: Analyzing Cohesive Devices between language and image," is a promising one. It constitutes a model that can explain cohesive mechanisms

number of those nouns in relation to their determiners in a given automatic translation into Japanese (Kameyama *et al*.)

between language and image (Bateman, *Text* 171). Liu & O'Halloran present a preliminary attempt to categorize intersemiotic texture in a multisemiotic text, as displayed in Figure 2 as follows:

Figure 2: Intersemiotic texture categories proposed by Liu and O'Halloran



Source: the authors (adapted from Pires, *Multimodality and Evaluation,* 90)

Figure 2 is a representation of Liu & O'Halloran's categories of intersemiotic texture, that is, of cohesive devices between language and image (372-4). The authors take the concepts of Halliday's language as a semiotic resource as point of departure (Halliday, 1978), as well as Martin's (1992) and Hasan's (1985) ideas of texture and cohesion within discourse. These ideas are employed by Liu & O'Halloran as the basis for text-image semantic relation analysis within discourse, thus expanding on Royce's framework for "intersemiotic complementarity" (1998; 2007).

Liu & O'Halloran's intersemiotic texture is composed of three main categories of cohesive devices, namely "intersemiotic parallelism," "intersemiotic polysemy," and "implication sequences between language and images." The first type of cohesive device, "intersemiotic parallelism", is "a relation that interconnects both

language and images when the two semiotic components share a similar form" (372). Language and image may share a similar form, as the page where the picture of a bonfire has the wording "hot" above it in the form of a smoke (372), or by means of the configuration between both modes, as in a photograph of a woman being bitten by a dog, followed by its caption "Israeli army dog attacks Palestinian woman" (373). The former example is entitled "homospatiality," frequently found in comics and advertising campaigns, and the latter is named "intersemiotic parallel structures," possibly represented within the context of advertising campaigns, too, but also newspaper articles.

The second type, perhaps a more sophisticated cohesive device described by Liu & O'Halloran, is the "intersemiotic polysemy" (375). Unlike "intersemiotic parallelism," "intersemiotic polysemy" refers to language and image sharing *multiple meanings* in a multisemiotic text. Therefore, polysemy points at similarities, resulting in "co-contextualization relations" and experiential convergence between both semiotic components (375). This is very much the case of advertising campaigns (but not limited to them), in which the meaning of the linguistic and visual components is organized to converge and multiply several meanings with the intent of appealing to the brand's target audience (375-7).

The third major category of intersemiotic cohesive devices is perhaps the most developed one, referring to two approaches: grammar-based and discourse-based. The former refers, in fact, to terms of its limitations, to propose an expansion of text-image relations within the scope of multimodal discourse (378). Indeed, the authors expand the discourse-based intersemiotic approach into four subcategories, namely: i) intersemiotic comparative relations; ii) intersemiotic additive relations; iii) intersemiotic consequential relations; and iv) intersemiotic temporal relations, as "Implication sequences between language and images" (378-384).

According to Liu & O'Halloran, comparative relations are "a kind of resource for organizing logical meaning with respect to similarity between language and images in multimodal discourse"

(379). That similarity, however, may vary in the way the linguistic message (for example, the message of the caption for a photograph in a news article) reformulates its corresponding visual elements (the photograph). That relationship may be a reformulation in terms of *generality*, such as the linguistic component specifying more general and relevant information in the visual component; or it may be in terms of *abstraction*, such as a caption that renders a more abstract message in relation to a similar but more concrete message expressed in the visual.

Unlike intersemiotic comparatives, in intersemiotic additives, the linguistic component or the visual component adds "new" information to the other, and thus both messages are connected to each other (379). That is, in *Intersemiotic Comparatives* "language and images have different semiotic reformulations of more or less identical experiences, [however] in Intersemiotic Additives verbal and visual parts convey related but different messages" (380).

The third category, intersemiotic consequential relations, occurs when "one semiotic message is seen as enabling or determining the other rather than simply preceding it" (Liu & O'Halloran, 380). This category is subdivided into two others, namely intersemiotic consequence and intersemiotic contingency. The former represents causal relations between verbal and image messages, "where the effect has been ensured" (380); the latter, however, represents a text-image logic with a "potential to determine a possibility while there is no ensured effect" in the combination of intersemiotic messages (382). Thus, contingency relations refer to a meaning of "purpose," rather than the meaning of "cause" in consequential relations.

The fourth discourse-based implication sequence is intersemiotic temporal relations. Liu & O'Halloran especially address multimodal documents with procedurals, such as manuals (383). According to the authors, that genre is representative of these relations because one may observe "different procedural steps represented both verbally and visually" (383). Therefore, different prominent levels of diverging modalities may take place in these cases, resulting in a variety of intersemiotic configurations.

Within the scope of multisemiotic documents in [Machine] Translation relations, this intersemiotic texture approach presents a potential to inform new configurations of intersemiotic meanings, based on a single error of machine-translated output. These situations will be explained in the following subsection.

2.3 Intersemiotic mismatches in webpages translated automatically

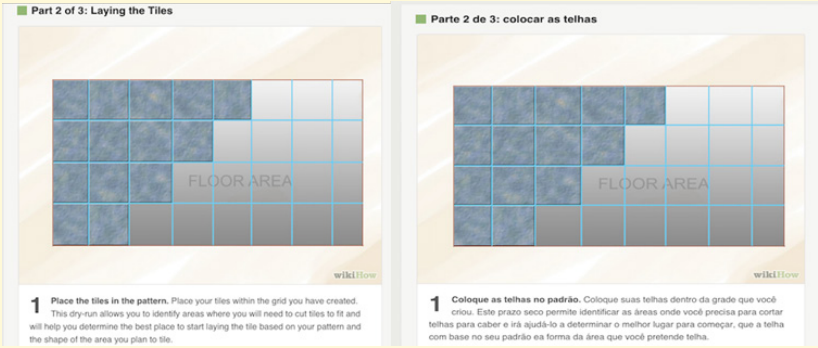
One may observe that web pages are often rather "rigid" regarding variation of multimodal discourse, that is, variation between the modes of expression (such as the menu bar, home page, or main title, for example) that create a unified meaning. Such "rigidness" is often due to templates; by these elements cannot, by any means, be used creatively to spark more variation among its modes (such as webpages about art, for instance), but its default usage does not enable much variation. To acknowledge that rigidness is also to acknowledge that images on a webpage do not change when a text is automatically translated. MT reproduces statistical and probabilistic relationships in its database. Thus, if part of a webpage that semantically links text and image is automatically translated, therefore, part of the multimodal meaning of the source text "changes".

As a result, an image on a web page will remain the same image after Machine Translation, still occupying the same position on the page. The text also takes the same position on the webpage, though with certain "mismatches" (Kameyama *et. al.*) based on the results generated by MT. In other words, there may be a shift in meaning with machine translation output that may lead to shifts in meaning in relation to other linguistic components and/or visual components, creating a new configuration of the unified meaning in the automatically translated multimodal document.

Basic units of linguistic components, such as the lexical semantic units of an image's captions, may present divergences, which may change their intersemiotic relationship, as shown in the following

example, taken from part of a *Wikihow* article in English and its *Google Translate* version into Portuguese:

Figure 3: Sample 1: Intersemiotic mismatch on a *Wikihow* article



Source: adapted from the *Wikihow* website.

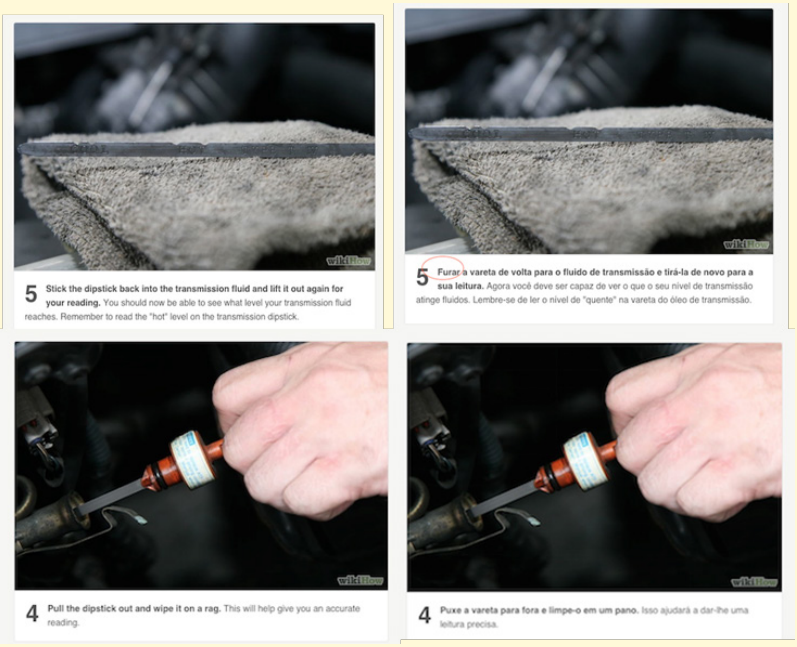
The example shown in Figure 3 illustrates a phenomenon involving lexical semantic intersemiotic mismatches in automatically translated texts. The image was taken from a *Wikihow* article originally in English that explains the procedure to lay marble floors. As previously explained, the image and layout were maintained in the translation. For the output of the translated caption, the *Google Translate* add-on was used on the *Google Chrome* browser. The text is automatically replaced by its *Google translation* into Portuguese, maintaining the originally designed position on the webpage. Both parts were then screenshot and set side by side for further analysis.

Figure 3 shows that the lexical item "tile" displays the translation "telhas". The contextual use of the word "tiles" eliminates other possible meanings for the lexical item. That context could be expressed by the visual or by the linguistic component of the article; however, that is not visually represented in Portuguese. "Telhas" corresponds to "roof tiles" into English, though the image depicts a floor, reinforced by the words "floor area". In that sense, the

corresponding visual meaning of "tiles" represents a different meaning from "telhas".

Another example that shows a shift in text-image relationships when automatically translated is the *Wikihow* article titled "How to add transmission fluid". However, in the following example, one may find differences in relation to other parts of the document.

Figure 4: Sample 2: Intersemiotic mismatch with temporal relation.



Source: adapted from the *Wikihow* website.

Figure 4 displays parts from a sequence of steps to guide the reader about how to add transmission fluid to a vehicle. The methods for analyzing the text-image relationships in Figure 4 are the same as for Figure 3, except for one factor: in Figure 4, the intersemiotic mismatch concerns a text-image narrative, instead of exposing a direct impact of a single caption and its respective illustration.

The first sentences in each of the two steps are in bold letters, clearly claiming more relevance to the procedures they describe. One can observe that the first clause contains an intersemiotic additive (Liu & O'Halloran) for the action "pull the dipstick/puxar a vareta"; thus, it adds meaning to a hand that "holds" the dipstick in the photograph. The second clause, "wipe it on a rag", is not related to the photograph itself in step 4, but, rather, and partially, it relates to the consequence shown by the photograph in step 5, that is, the cloth with oil stains below the dipstick. Up until this point, there have been no errors in Machine Translation output.

However, in step 5, the caption starts with the following description: "Stick the dipstick back into the transmission fluid and lift it out again for your reading". Both clauses add two pieces of information to the picture in the previous step, forming a sequence of "sticking back" and then "lifting it out". Therefore, this linguistic component has a temporal relation to the previous photograph, just as to the previous caption (step 4) does with the photograph in step 5. As a result, they imply an intersemiotic temporal relation (Liu & O'Halloran).

Regarding their *Google translations* into Portuguese, however, the caption for step 5 has some issues in temporal relations. One could say that a possible translation into Portuguese for the sentence "Stick the dipstick back into the transmission fluid and lift it out again for your reading" would be something like "espete a vareta de volta no fluido de transmissão". However, what one observes in the *Google Translate* output is an intersemiotic incompatibility that starts with an incorrect disambiguation (Vilar *et al*.) in "furar"[3] ("to pierce"), next to the word "vareta" ("dipstick"), forming the idea of "piercing the dipstick" in Portuguese. Therefore, the temporal intersemiotic relationship once established between the action of "sticking the stick back" in caption 5 with the hand holding the dipstick in the previous step generated an **intersemiotic mismatch in temporal relations**.

[3] Appropriate translation possibilities into Portuguese for the verb "stick" in this context would be "colocar" or "espetar," for example.

Both samples are part of a range of possibilities to explore and categorize text-image semantic mismatches generated by machine translation output. While undergoing scrutiny on the web could be very time-consuming, the analyst should still go through several results to test the translated text from a machine-translated output against its visual components. Then, there should be some sort of "error" to be observed, along with the image or part of it to find any new emerging configurations in the same text-image units from the source document.

That is why appropriate tools and methods are required to facilitate tagging these relationships. The following subsection analyzes some user-friendly tools, which could be used for undertaking the study of intersemiotic mismatches.

2.4 Tools for intersemiotic mismatch analysis

The range of tools aimed at analyzing text-image translations has still been scarcely explored (Pires, *Ampliando* 18). Therefore, this study produced an analysis of tools to verify whether their use could make the process of annotation and tagging for static multimodal documents in corpora construction less time-consuming. This article analyzed three tools, namely *Evernote*, *Nimbus Capture*, and *UAM ImageTool*. The choice of tools considered the partial or full gratuity of the functions used to annotate and tag selected multimodal documents.

Evernote

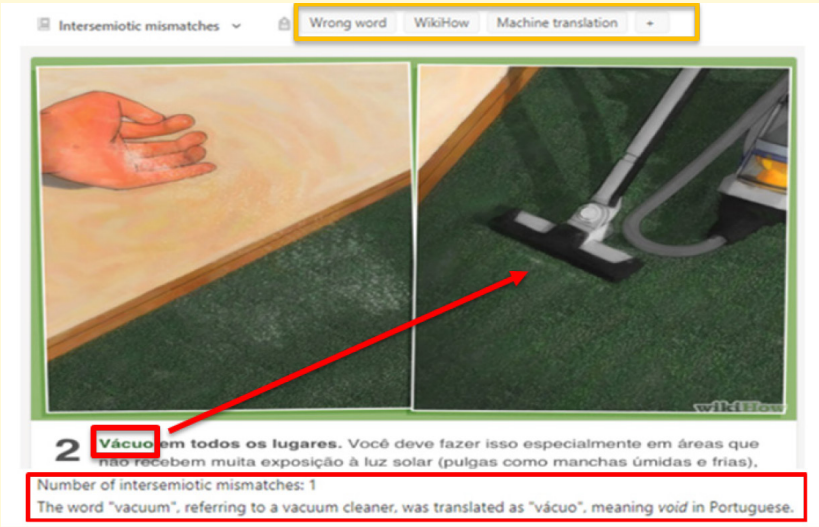
Evernote is a tool that allows saving, editing, and sharing files. It can be adapted to tag and annotate static multimodal documents from web pages. The tool has a simple and intuitive interface, which means the user does not need much information in order to use it. It enables a less time-consuming process as the investigator can about *Evernote*'s functions while using it.

The software also provides an editing mode for static multimodal documents, such as the analyzed text-image compositions. In this

mode, it is possible to add text and geometric shapes to documents; both forms have been used by Pires (*Ampliando; Multimodality*) to annotate multimodal documents. A special feature in *Evernote* is the option to insert tags into saved documents. It is possible to annotate and tag the files, which helps locate, group, or semi-automatically quantify them. Another positive feature of the tool is that technical errors, such as screen freezes or sudden shutdowns, do not commonly occur.

Figure 5 highlights *Evernote*'s functions of annotation and tagging for multimodal documents extracted from web pages. Examples of tags, inserted manually by the investigators, are highlighted in yellow. The annotation system, highlighted in red, consists of the insertion of additional information in a notepad format, that is, below the collected document, similar to a caption, in addition to geometric forms to emphasize a verbal mismatch and its relationship to the visual mode presented.

Figure 5: *Evernote*'s annotation and tagging resources



Source: authors (adapted from *Evernote* and *Wikihow*).

Despite the fact that the program allows the user to choose which and how many tags they want to use to mark a document, there is no division of projects. That is, regardless of how many folders or projects the user creates, the organization of tags is alphabetical, in a single shortcut. That limitation is a negative aspect if the researcher uses the tool to carry out different investigations, as their information would, then, be stored in the same folder, merging contents that they did not intend to merge at first (Espindola & Pires 259).

Moreover, *Evernote* does not have a system that allows automatic quantification in the tagging system, which implies a manual or semi-automatic process to extract statistical data or to check for patterns. The tagging system is individual, meaning it is not possible to reuse annotation and tagging schemes automatically in other files.

Nimbus Capture

Nimbus Capture is a screen capture tool. More than that: the tool also offers annotations in static multimodal documents and tagging schemes, as described in the following paragraphs.

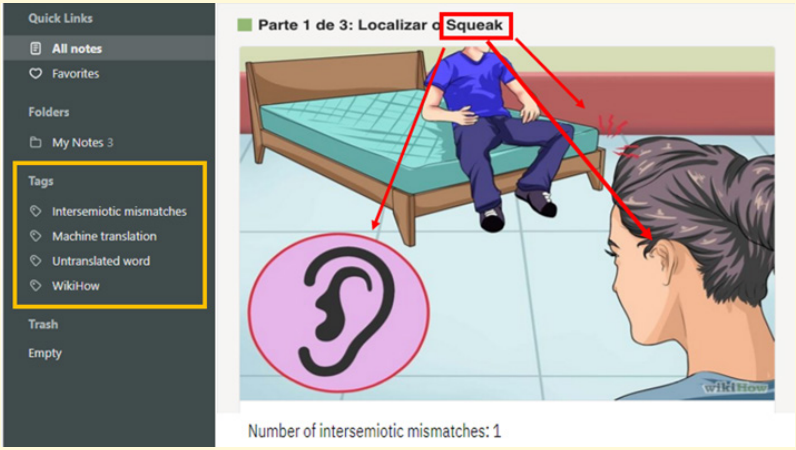
The interface of this tool is user-friendly, with simple and intuitive commands. This characteristic allows users to quickly learn how to use its features, reducing the period of learning curve for users. Besides, similarly to the *Evernote* tool, malfunctions or frozen screen errors are not common.

Nimbus Capture also offers options for image editions. This function allows the annotation of text-image documents, such as the addition of shapes and text. The tool enables the annotation of generated documents through markings, as exemplified in Figure 6: the red marking highlights intersemiotic mismatches generated by machine translation and their relationship with the visual mode, indicated by the arrows.

In addition to the edition mode that enables annotation in a text-image multimodal document, *Nimbus Capture* offers a tagging system. It is possible to add, delete, or modify tags at any time, as highlighted in yellow in Figure 6. To optimize the process, its

interface records previously registered tags for later use. Tags also make it easier to search and group documents, as the system allows data grouping by tagging them into various categories.

Figure 6 – *Nimbus Capture* tagging and annotation systems



Source: authors (adapted from *Nimbus Capture* and *Wikihow*).

This study adapted its functions to annotate and tag the multimodal text-image documents. The tool lacks certain analysis functions common to specialized tools, such as statistical functions, which can make it easier to observe the existence of patterns.

UAM ImageTool

UAM ImageTool (O'Donnell) is a specialized image annotation tool for corpus analysis. The program allows tagging and annotation of images, in addition to offering the possibility to quantify the types of analysed categories (Pires, *Ampliando* 106).

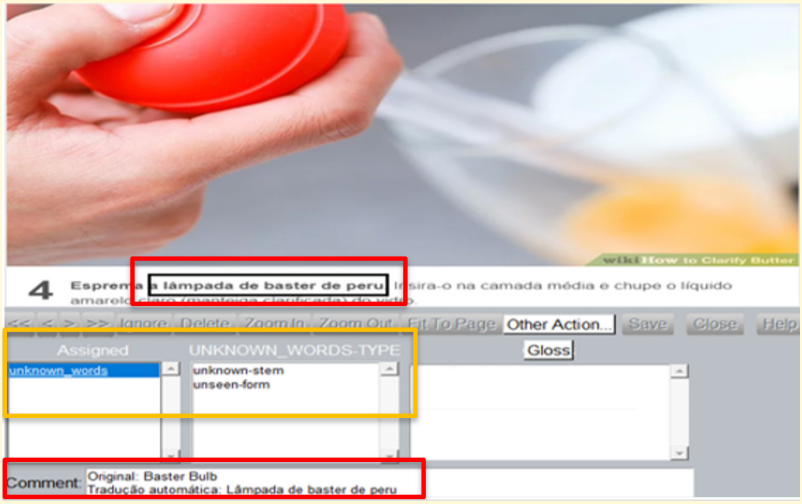
The program provides "cross-classification and under-specification of resources" (O'Donnell 15), which enables correlation between documents, whether in the same folder or not. With that, any change to an image's tagging scheme will automatically update

all *corpus* files that use the same tags (Espindola & Pires 260). Thus, it is possible to reuse the same tagging scheme in multiple images, which makes the time spent on that process shorter when compared to the individual systems of *Nimbus Capture* and *Evernote*.

The addition of an annotation scheme appears directly on the document's visual mode, with the insertion of squares, or as written comments in a space similar to a note, titled "comment" section. Differently from *Nimbus Capture* and *Evernote*, the image annotation in *UAM ImageTool* (O'Donnell) is available only in black, with a single geometric form; that implies only one option to annotate in the visual mode of the multimodal document.

In Figure 7, the red markings show examples of annotation; the yellow mark, an example of tagging. It is worth mentioning that one may share only the tagging scheme amongst files, whereas the annotation is manual and individual. One of the traits that sets it from others is that it has support for a series of corpora statistical analyses, enabling the investigation of patterns (O'Donnell 16).

Figure 7 – *UAM ImageTool*'s annotation and tagging tools



Source: the authors (adapted from *UAM ImageTool* and *Wikihow*).

In comparison to the other tools analyzed in this study, the *UAM ImageTool* (O'Donnell) has the advantage of being a specialized software. However, the time taken to learn how to use that program is longer due to the complexity of the tool's functions, when compared to *Nimbus Capture*'s or *Evernote*'s tools.

The analysis shows that the combination of specialized and non-specialized tools has a greater potential to make the annotation and tagging processes less time-consuming than their individual use. That finding considers semi-automatic tagging and automatic statistical functions offered by the *UAM ImageTool* and the ease of annotation of the *Evernote*.

3. Final remarks

This study provided an explanation about text-image relations within the context of Multimodality, specifically under the intersemiotic texture approach developed by Liu & O'Halloran, and Machine Translation (MT), precisely the classification approaches for errors in Machine Translation output (Vilar *et al.*; Kameyama *et al.*) to analyze some examples from the *Wikihow* website. The analysis shows one intersemiotic ambiguity generated from a lexical semantic error [Figure 3] and an intersemiotic mismatch in temporal relation, generated from an incorrect disambiguation [Figure 4]. Moreover, it reported on recent investigations of suitable tools and methods to properly tag and annotate these relationships to facilitate manual manipulation of a large quantity of data.

By employing Liu & O'Halloran's intersemiotic texture approach, its systemic-functional portion has not been taken into consideration for this analysis. This is due to the fact that the focus here is exploratory, but it constitutes a relevant qualitative step for the analysis of intersemiotic mismatches.

In relation to the annotation tools, it is worth noting that non-specialized programs, that is, tools adapted to corpora study have significant limitations for analysis. The limitations of each tool

corroborate Duncan's studies, in which the findings suggest that no tool is free from disadvantages concerning certain analytical objectives (Duncan, 1020).

An important concern regards image rights for further *corpus* compilation. Certain news articles, photographs, and other visual elements constitute third parties. Therefore, finding and contacting the authors may pose a challenge to studies.

The investigation of the phenomenon described as "intersemiotic mismatch" generated by Machine Translation outputs is a complex and promising area, with much to be done yet. Many of Liu & O'Halloran and Vilar *et. al*. categories can be explored, along with a wider variety of text genres, thus contributing to the interface of Multimodality and Translation Technology.

References

Baker, Mona; Saldanha, Gabriela (Orgs.). *Routledge encyclopedia of translation studies*. 3rd ed. London: Routledge, 2019.

Bateman, J.A. *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. London: Routledge, 2014. Avaible to: <https://books.google.de/books?id=JvPAngEACAAJ>.

Bateman, J. A. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. London: Palgrave MacMillan, 2008.

Caglayan, Ozan, *et al*. "Does Multimodality Help Human and Machine for Translation and Image Captioning?" *Proceedings of the First Conference on Machine Translation: vol. 2, Shared Task Papers*, (2016): 627-33. DOI:10.18653/v1/W16-2358. Avaible to: <*arXiv.org*>.

Caglayan, Ozan, *et al*. *Multimodal Machine Translation*. Université du Maine, 2019.

Calixto, Iacer; Liu, Qun. "An error analysis for image-based multi-modal neural machine translation." *Machine Translation*, vol. 33, n. 1, (2019): 155-77. *PubMed Central*. DOI:10.1007/s10590-019-09226-9.

Dorr, Bonnie. "Machine Translation Divergences: A Formal Description and Proposed Solution." *Computational Linguistics*, vol. 20, no. 4, (1994): 597-634.

Dorr, Bonnie. "Solving thematic divergences in machine translation." *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1990, p. 127-134. DOI:10.3115/981823.981840.

Duncan, Susan. "Multimodal annotation tools". *Body–Language–Communication: an international handbook on multimodality in human interaction*. Berlin: De Gruyter Mouton, 2013, pp. 1015-1022.

Espindola, Augusto; Pires, Thiago Blanch. *"Coleta, etiquetagem e anotação de incompatibilidades intersemióticas geradas por tradução automática". Cultura e Tradução,* vol. 6, no. 1, (2020): 248-264.

Halliday, M. A. K. *Language as social semiotic: the social interpretation of language and meaning*. London: Arnold, 1978.

Hasan, R. "The texture of a text.". *Language, Context and Text: Aspects of language in a socio-semiotic perspective*. Deaking: Deaking University Press, 1985.

Heo, Yoonseok, *et al*. "Multimodal Neural Machine Translation with Weakly Labelled Images." *IEEE Access*, vol. 7, (2019): 54042-53. *IEEE Xplore*. DOI:10.1109/ACCESS.2019.2911656.

Hirasawa, Tosho, *et al*. "Multimodal Machine Translation with Embedding Prediction." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 2019, pp. 86-91. *ACLWeb*. DOI:10.18653/v1/N19-3012.

Hutchins, W. J. *Machine Translation: Past, Present, Future*. John Wiley & Sons, Inc., 1986.´

Hutchins, W. J. "Machine translation: A concise history." *Journal of Translation Studies*, vol. 13, no. 1-2, (2010): 29-70.

Hutchins, William John; Somers, Harold L. *An Introduction to Machine Translation*. Massachusetts: Academic Press, 1992.

Kameyama, Megumi, *et al*. "Resolving Translation Mismatches with Information Flow." *Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL91*, 1991, pp. 193-200.

Liu, Yu; Kay L. O'Halloran. "Intersemiotic Texture: analyzing cohesive devices between language and images." *Social Semiotics*, vol. 19, no. 4, (2009): 367-388. DOI:10.1080/10350330903361059.

Martin, J. R. *English text: system and structure*. Amsterdam: John Benjamins Pub. Co, 1992. Avaible to: <http://bangor.eblib.com/patron/FullRecord.aspx?p=861548>.

Melby, Alan K. "Future of Machine Translation.". *The Routledge Handbook of Translation and Technology*, O'Hagan, Minako (Org.). London: Routledge, 2019, pp. 419–36. *DOI.org (Crossref)*. DOI:10.4324/9781315311258-25.

Mills, Kathy A.; Len Unsworth. "Multimodal Literacy." *Oxford Research Encyclopedia of Education*. Oxford: Oxford University Press, 2017. *DOI.org (Crossref)*. DOI:10.1093/acrefore/9780190264093.013.232.

O'Donnell, Michael. "Demonstration of the UAM CorpusTool for text and image annotation". *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics,* 2008, pp. 13-16.

Pires, Thiago Blanch. *Ampliando olhares sobre a tradução automática online : um estudo exploratório de categorias de erros de máquina de tradução gerados em documentos multimodais*. Universidade de Brasília, 2017, Avaible to: <https://repositorio.unb.br/handle/10482/23727>.

Pires, Thiago Blanch. "Multimodality and Evaluation of Machine Translation: A Proposal for Investigating Intersemiotic Mismatches Generated by the Use of Machine Translation in Multimodal Documents". *Texto Livre: Linguagem e Tecnologia*, vol. 11, no. 1, June (2018): 82-102. Avaible to: <*www.periodicos.letras.ufmg.br*>. DOI:10.17851/1983-3652.11.1.82-102.

Quah, Chiew Kin. *Translation and technology*. London: Palgrave Macmillan, 2006.

Royce, Terry. "Intersemiotic Complementarity: A Framework for Multimodal Discourse Analysis.". *New Directions in the Analysis of Multimodal Discourse*, Royce, Terry, and Bowcher, Wendy (Orgs). London: Routledge, 2007, pp. 63-109.

Royce, Terry. "Synergy on the Page: Exploring intersemiotic complementarity in page-based multimodal text." *JASFL Occasional papers*, vol. 1, no. 1, (1998): 25-49.

Saçak, Begüm. "Media Literacy in a Digital Age: Multimodal Social Semiotics and Reading Media." *Handbook of Research on Media Literacy Research and Applications Across Disciplines*, 2019. DOI:10.4018/978-1-5225-9261-7.ch002.

Takushima, Hiroki, *et al*. *Multimodal Neural Machine Translation Using CNN and Transformer Encoder*. EasyChair Preprints, EasyChair, April 2nd 2019. *DOI.org (Crossref)*. DOI:10.29007/hxhn

Vilar, David, *et al.* "Error Analysis of Statistical Machine Translation Output." *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, European Language Resources Association (ELRA), 2006. *ACLWeb*. Avaible to: <http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf>.

Recebido em: 06/12/2020
Aceito em: 12/03/2021
Publicado em maio de 2021

Thiago Blanch Pires. E-mail: thiagocomaga@gmail.com. ORCID: https://orcid.org/0000-0002-0060-6075.
Augusto Velloso dos Santos Espindola. E-mail: augustovse@gmail.com. ORCID: https://orcid.org/0000-0002-7606-1913.