Machine Translation of Multiword Expressions: Double Dutch or Crystal Clear?

Rozane Rebechi[1], Nathalia Marcon[2], Guilherme Faller[3] and Aline Villavicencio[4]

[1 2 3] Federal University of Rio Grande do Sul, Porto Alegre RS 90850-050, Brazil
[4] University of Sheffield, Sheffield, UK
rozane.rebechi@ufrgs.br

Abstract. Due to their varying degrees of opacity, contextual dependency, and nuanced meanings, multiword expressions (MWEs) can present considerable challenges for Machine Translation (MT). The objective of this study is twofold. Firstly, it evaluates the effectiveness of MT in rendering both non-compositional (NC) and partly compositional (PC) nominal MWEs from English to Portuguese, vis-à-vis human translation (HT). Then it examines MT-generated translations that deviate from HT and verifies their adequacy. HT for compounds was conducted by supervised trainee translators seeking conventional equivalents, corroborated within the Brazilian subcorpus of the Corpus do Português. Contextualized MWEs extracted from the Corpus of Contemporary American English were subjected to MT by Google Translate and ChatGPT 3.5. When assessing NC MWEs, the findings suggest that, in terms of raw frequency, ChatGPT's choices more frequently align with HT's than Google's. Additionally, ChatGPT's mismatches exceed those of Google in terms of adequacy. As for PC MWEs, both tools produced comparable results, with a slight advantage for Google. Statistically, no significant difference was identified for NC or PC MWEs, either in terms of matches or adequacy between both MT tools. A qualitative investigation of MT mismatches from HT was carried out and demonstrated that both Google Translate and ChatGPT can sometimes recognize the idiomaticity of MWEs and propose suitable translations for PC and NC compounds. Other times, however, they fail in keeping metaphorical and conventional equivalents, frequently resorting to more literal translations. In doing so, they are also not always capable of discerning that some of their suggestions are ludicrous within the context given.

Keywords: Human translation, Machine Translation, Multiword Expressions, Conventionality.

1 Introduction

Machine translation (MT) has been applied to different genres, in different language pairs, with different levels of adequacy. For the English-Portuguese language pair, [1] examines the effects of using MT for the translation and post-editing of literary texts into the Brazilian translators' foreign language and concludes that the effort in the post-editing process is reduced, while [2] compare English-Portuguese MT of legal

Proceedings of New Trends in Translation and Technology, pages 116–137
3–6 July, 2024, Varna, Bulgaria.
https://doi.org/10.26615/issn.2815-4711.2024_011

arbitration clauses to assess adequacy and fluency and conclude that the results tend to favor fluency over adequacy. Nevertheless, the authors point that, despite some inconsistencies, MT terminological accuracy is considered satisfactory in this genre.

The objective of this study is to assess, both quantitatively and qualitatively, the adequacy of MT, generated by ChatGPT 3.5 and Google Translate, in rendering non-compositional (NC) and partly compositional (PC) multiword expressions (MWEs) from English to Brazilian Portuguese. We focus on MWEs as they vary in their degree of compositionality, since not always the meaning of an expression can be straightforwardly composed by the meanings of its component words. As non-compositional NCs may not be easily linked to their individual components, translating them literally may lead to information loss or even incorrect translations. Our hypotheses are that NC MWEs may lead to more translation errors. Additionally, this study aims to examine the adequacy of MT-generated translations that diverge from those generated by humans.

In this research, MWEs are understood as combinations of words which function as a single semantic unit [3]. They may exhibit idiosyncratic syntactic or semantic behavior, and their interpretation depends heavily on context. For example, unless it is contextualized, the phrase "kick the bucket" can be understood literally as "to strike out a cylindrical container with a foot" or idiomatically as "to die." Identifying and properly handling MWEs is essential in many areas of natural language processing (NLP), including MT, as their presence can significantly impact the accuracy and fluency of language processing tasks. Hence, any NLP system aiming to effectively address phrasal semantics must possess the capability to discern between fairly compositional and fully idiomatic compounds [4].

This research draws upon a list of PC and NC nominal compounds in English, specifically those consisting of combinations such as noun + noun and adjective + noun, with a noun serving as the head[1], to evaluate the adequacy of two different systems for automatically translating texts. Our objective is to assess the suitability of Google Translate, a Neural Machine Translation (NMT) method, and ChatGPT 3.5, an Artificial Intelligence Chatbot, for translating MWEs by comparing their choices with human translators'. To achieve this objective, trainee translators translated the compounds according to predetermined meanings and selected sample sentences in which these compounds are contextualized. The sentences were then automatically translated.

A quantitative analysis was conducted using an index match function, followed by a chi-squared test of independence to determine the statistical significance of the disparities observed in the raw numbers derived from the matching index. Next, a manual analysis of the mismatches was performed, aimed at identifying adequate or

[1] For further elaboration on the initial compilation of this list, refer to [4].

inadequate translations for non-matching cases, offering insights into the discrepancies, with consideration given to conventionality [5].

2 Literature Review

2.1 Multiword Expressions

For this study, MWEs are defined as typical combinations of words which form a single unit. They can be categorized as compositional, partly compositional, or non-compositional based on the degree of transparency of their constituent words (cf. [4]). One method of differentiating MWEs from other types of word combinations is by their typically rigid structure. Unlike ordinary phrases, MWEs often exhibit a fixed sequence of words, where the constituent elements are not readily interchangeable with synonyms. For example, the compositional compound "strong coffee" is easily comprehensible from the combination of its words, although it may not readily accept substitutions. The occurrence of "powerful coffee" would certainly evoke the naivety of a non-proficient speaker (cf. [6]) or a deliberate break from conventionality for humorous effect, for example.

Partly compositional compounds contain one transparent word and one opaque word. "Noble gas" is indeed a type of gas, yet it does not possess any nobility title. Non-compositional compounds are formed by word combinations whose meanings are not easily inferred, as exemplified above by "kick the bucket" in its idiomatic sense of dying.

Although many MWEs may have idiomatic and literal senses, depending on their contexts, others are primarily used in their idiomatic form. As an illustration, since insects cannot speak, the occurrence of "spelling bee" as a compositional MWE would only be feasible for intentional disruption of expectations.

2.2 Neural Machine Translation

Machine Translation (MT), the process of automatically translating text through computer software, has evolved from a pursuit of early digital computer enthusiasts to becoming one of the most prominent and accessible applications of AI. It now stands as an integral component of numerous translation production workflows [7]. MT for accessing information ranks among the most prevalent applications of NLP technology, and Google Translate alone handles the translation of hundreds of billions of words daily across more than 100 languages [8].

[9] explain that NMT systems represent source sentences as inputs and interpret outputs as target sentences. Although these systems, containing thousands of neurons and millions of weights, are trained using examples from parallel corpora, by repeatedly modifying the strength of connections between neurons according to anticipated inputs

and outputs, they gradually acquire the ability to generate the desired translations. [7, p. 104] explain that "NMT can usually produce fluent-sounding output that takes account of context (usually only within a sentence rather than a whole document), but that output is not always consistent, and might change depending on the sentence to be translated, the training data, and the words produced so far."

As words can have different meanings depending on context, attention mechanisms are essential for computing contextual word embeddings, allowing the neural network to adjust word representations based on their meanings within specific sentences. This enables the network to differentiate between various senses of words.

2.3 Generative Artificial Intelligence

According to [10, p. 1], "Large Language Models (LLMs) have emerged as cutting-edge artificial intelligence systems that can process and generate text with coherent communication, and generalize to multiple tasks". In other words, LLMs are designed to process and generate human-like text based on large amounts of training data. These models, built on deep learning architectures, have millions or billions of parameters, enabling them to capture even complex patterns in language. LLMs are trained on diverse text sources, including books, articles, websites, and other written materials, to learn the nuances of language structure, syntax, semantics, and context.

ChatGPT is a conversational AI developed by OpenAI, based on the GPT (Generative Pre-trained Transformer) architecture. According to its own definition, ChatGPT can engage in a wide range of conversations, answer questions, provide explanations, and offer assistance across various topics and domains. Released in 2020, GPT-3 [11] is one of the largest neural networks in the field of natural language generation. Its architecture is not strictly classified as Statistical Machine Translation (SMT) or Neural Machine Translation (NMT). Actually, ChatGPT is not explicitly designed for the task of translation, although it can generate translations by processing and understanding input in different languages, especially with multilingual versions trained on various language data.

[12] evaluated ChatGPT for machine translation across various aspects, including translation prompts, multilingual translation, and translation robustness. Findings indicate that ChatGPT performs competitively with commercial translation products, like Google Translate and DeepL Translate, for high-resource European languages, but falls behind for low-resource or distant languages. The authors' conclusion highlights that while GPT-3 demonstrates good performance with spoken language translation tasks, it encounters difficulties when handling biomedical abstracts and Reddit comments. Conversely, the GPT-4 engine displays improved performance, comparable to commercial products, particularly evident in distant language pairs, like Chinese-English.

2.4 Conventionality

When engaging with standard texts, readers typically anticipate encountering familiar features, unless they are immersed in highly creative content. [13] claims that constructing coherent text solely through randomly selected linguistic elements, a concept termed open-choice, is implausible. Instead, users predominantly rely on the idiom principle, drawing upon a repository of semi-preconstructed phrases that represent singular choices. This reliance suggests that language users possess an extensive repertoire of pre-existing phrases, facilitating efficient communication through the selection of appropriate expressions.

[14] further delineates conventionality across three linguistic levels: syntactic, semantic, and pragmatic. These conventions are grounded in usage rather than conceptual logic and depend on the recognition of linguistic units stored in memory. As demonstrated earlier, when encountering the MWE "spelling bee" in everyday texts, a proficient reader will immediately associate it with the competition involving the oral recitation of letters in a word, rather than interpreting it as a literal reference to a buzzing insect engaged in conversation. If the insect in the expression is replaced with "drone," for example, which typically refers to a male bee, the reader may become confused as it deviates from the established convention.

Even the arrangement of words can impact comprehension. For instance, the English MWE "big fish," as in "However the big fish like Urdangarin (royal in-law) or high profile politicians almost never end up in jail [...]", signifying an influential individual, could be literally translated into Portuguese as *peixe grande*. However, if the position of the adjective and noun is switched – *grande peixe* –, the compound would deviate from the idiomatic expression, since, in this order, it is used to refer to an aquatic vertebrate of large proportions.

It has been over three decades since [15] as cited in [16, p. 177] highlighted that the lexicon poses challenges not only in terminology, lexicography, and translation but also in NLP and, consequently, in MT. The researcher concluded that the vocabulary of a language often proves to be a bottleneck in the design of large-scale natural language systems due to the vast number of words.

Corpora, within the realm of corpus linguistics, serve as invaluable tools for translators in identifying peculiarities in source texts and devising solutions for the target language. They quantitatively retrieve data on conventionality patterns for further manual analysis [5]. Therefore, we turned to English and Portuguese corpora to verify whether the proposed equivalents, both by humans and MT, are conventionally used in authentic texts.

3 Methodology

NMT is certainly responsible for recent significant advancements in the quality of automatic translation. Citing [17], [2] remind us that Microsoft envisioned a translated text resembling human output after the establishment of NMT, while Google suggested that this technology could bridge the gap between human and previous machine translation methods ([18], as cited in [2]).

[19] claims that NMT has proven to be quite competitive by presenting considerably fewer errors compared to its predecessors, whereas [20] conclude that MT has achieved better performance in the adequacy than in the fluency parameter. Due to space limitations, this study will primarily concentrate on the adequacy of MT choices, considered as the appropriate transfer of meaning from the source text to the target text, for equivalents that did not match human choices, albeit a brief discussion on conventionality is provided.

Based on a compilation of nominal compounds [21] categorized as compositional, non-compositional (NC), and partly compositional (PC) – where the head word is a noun [4] –, we specifically selected 101 NC and 83 PC compounds, each with established meanings. For example, a potentially polysemous compound such as "double Dutch" – which may refer to a jump rope game or incomprehensible talk – was kept as the latter meaning.

Compounds were then manually translated according to their predetermined meanings. Whenever possible, one to four equivalents were provided, validated by their frequency in the Brazilian subcorpus of the Corpus do Português: Web/Dialects (CP) [22], encompassing 713,047,728 words across diverse genres, to ensure their conventional usage. Essentially, after an exhaustive search for equivalents through various resources – excluding MT – and subsequent deliberations within the group, both the trainee translators and the research coordinator checked their suggestions in the CP to verify their recurrence. This approach was intended to confirm that the translation choices for the compounds were not merely regional variations, for example. Next, the participants utilized the Corpus of Contemporary American English (COCA) [23] to find sentences contextualizing the compounds, with enough context to enable disambiguation.

The selected sentences were automatically translated using Google Translate and ChatGPT 3.5. It is worth noting that Google Translate does not offer specific variants for languages spoken in different countries. Therefore, we simply selected English as the source language and Portuguese as the target language. For the ChatGPT translation, we provided a prompt based on one of the possibilities outlined by [12]: "Please provide the Brazilian Portuguese translation for this sentence."

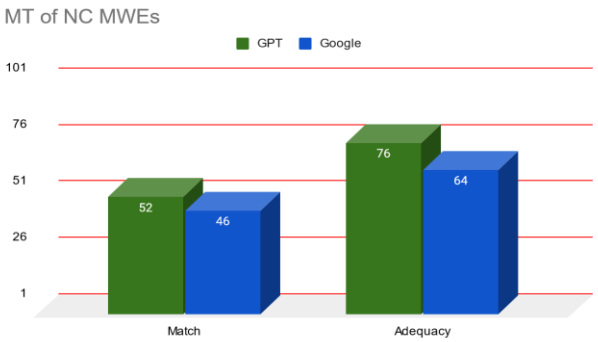
Both translations – Google's and ChatGPT's – were pasted into designated cells of a Google Drive spreadsheet. Subsequently, the translations of the compounds were manually inserted into their respective cells. During this process, conjugated verbs and

plural or feminine nouns and adjectives were lemmatized to align with the equivalents provided by HT. Exceptions were made for translations primarily used in another format. For instance, an equivalent for "cloud nine" in the context of great happiness is "*nas nuvens*" [literally, "in the clouds"]. In this case, the noun in the expression was retained in its conventional form as it appears in the idiomatic expression.

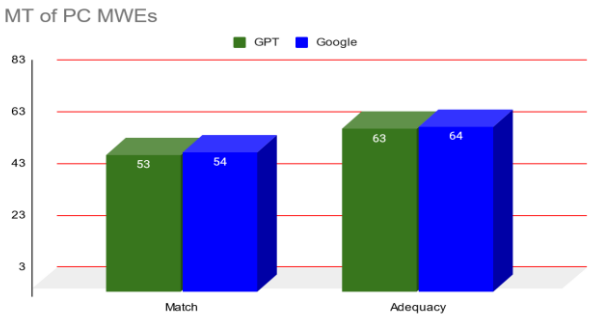
4 Analysis

4.1 Quantitative Analysis

An index match function was elaborated to compare the MT and HT of the MWEs. Out of the total 101 NC compounds, GPT's translations matched 52 human-generated equivalents (51%), while Google's yielded 46 matches (45%). For the 83 PC compounds, Chat GPT's translation was considered a match 53 times (63%) while Google's was 54 (65%). Nevertheless, we noticed that certain MT equivalents, while not identical to HT, were still suitable. Therefore, we conducted a manual analysis of the MT choices to assess their adequacy within the given contexts, which will be detailed in the next subsection (4.2). Graphs 1 and 2 below quantitatively summarize the results:



Graph 1. Comparison between translations of NC MWEs by ChatGPT and Google Translate.



Graph 2. Comparison between translations of PC MWEs by ChatGPT and Google Translate.

Next, we investigated whether the results are statistically significant, meaning we could assert that ChatGPT's performance was significantly superior to Google's when dealing with NC MWEs. To ascertain the statistical significance of the observed variances between the MT methods (ChatGPT and Google Translate) and human translation, we conducted a chi-squared test of independence. Initially, contingency tables for match rates and adequacy rates were established:

Table 1. Match rates for NCs.

	GPT Matches	Google Matches	Total
HT Matches	52	46	98
No HT Matches	49	55	104
Total	**101**	**101**	**202**

Table 2. Adequacy rates for NCs.

	GPT Adequate	Google Adequate	Total
HT Adequate	76	64	140
No HT Adequate	25	37	62
Total	**101**	**101**	**202**

For the Match Rates, considering the Observed and Expected Values, the following chi-squared test was set up:

$$\chi^2 = \Sigma ((\text{Observed} - \text{Expected})^2 / \text{Expected}) \quad (1)$$

$$\chi^2 = ((52 - 49)^2 / 49) + ((46 - 49)^2 / 49) + ((49 - 52)^2 / 52) + ((55 - 52)^2 / 52)$$

$$\chi^2 = 0.1837 + 0.1837 + 0.1731 + 0.1731$$
$$\chi^2 \approx 0.7136$$
Degrees of Freedom (df) = (number of rows - 1) * (number of columns - 1) = (2 - 1) * (2 - 1) = 1

Using a significance level of $\alpha = 0.05$, the critical value of χ^2 for df = 1 is approximately 3.841. Since our calculated χ^2 value (0.7136) is lower than the critical value, we fail to reject the null hypothesis. Thus, the difference in match rates between GPT and Google Translate for the NC compounds is not statistically significant.

For the Adequacy Rates, considering the Observed and Expected Values, the following chi-squared test was set up:

$$\chi^2 = \Sigma ((\text{Observed} - \text{Expected})^2 / \text{Expected}) \quad (2)$$

$$\chi^2 = ((76 - 70)^2 / 70) + ((64 - 70)^2 / 70) + ((25 - 31)^2 / 31) + ((37 - 31)^2 / 31)$$
$$\chi^2 \approx 0.52 + 0.52 + 1.16 + 1.16$$
$$\chi^2 \approx 3.36$$
Degrees of Freedom (df) = (number of rows - 1) * (number of columns - 1) = (2 - 1) * (2 - 1) = 1

Using a significance level of $\alpha = 0.05$, the critical value of χ^2 for df = 1 is approximately 3.841. Since our calculated χ^2 value (3.36) is lower than the critical value, we fail to reject the null hypothesis. Thus, the difference in adequacy rates between GPT and Google Translate is not statistically significant.

In summary, based on the outcomes of the chi-squared tests, we ascertain that the disparity in adequacy rates and the variance in match rates are not statistically significant between GPT and Google Translate. Regarding PC compounds, MT methods demonstrated nearly identical performance in matches and adequacy compared to human translation (Graph 2). Consequently, statistical tests were deemed unnecessary in this context.

4.2 Qualitative Analysis

Following the quantitative assessment of matches between translations generated by MT and HT, we conducted manual analyses on each translated compound to determine if any of them, despite not being a direct match, could serve as a viable option for the MWE under scrutiny. In instances where such translations were deemed plausible, we validated them as adequate translations[2]. Example 1 illustrates both translations of the sentence containing the MWE “bad hat” (our emphasis):

[2] Appendices A and B list the NC and PC MWEs, respectively.

Example 1

Source Text	We'd thought he was a gentleman, but he turned out to be a very **bad hat** indeed.
Google Translate	Pensávamos que ele era um cavalheiro, mas ele acabou se revelando um **péssimo chapéu** [lit. very bad hat].
ChatGPT	Pensávamos que ele era um cavalheiro, mas ele acabou sendo uma **pessoa** muito **mal-intencionada** [lit. lit. person with bad intentions] mesmo.

Although the only HT option was *mau caráter*, and hence, no match was attributed, through manual analysis we concluded that ChatGPT's rendering *pessoa mal-intencionada* fits perfectly well in the context expressed in the source text, so it was considered adequate. On the other hand, the literal translation offered by Google Translate is insufficient, as it would be interpreted in Brazilian Portuguese simply as a poorly made hat, disregarding the context in which the term is used in contrast with *cavalheiro* [lit. gentleman].

Furthermore, we identified MTs that would be appropriate had the word order been adjusted conventionally. For instance, the MWE "big fish," as discussed earlier, could indeed be translated directly into Portuguese with the noun-adjective order – *peixe grande* –, as done by Google, whereas the reversed order, as suggested by ChatGPT, would only convey its literal meaning. Conversely, the phrase "heavy cross" could conventionally be translated in the noun-adjective order – *cruz pesada* – as proposed by ChatGPT, or in the adjective-noun order – *pesada cruz* –, Google's choice, as evidenced in authentic texts to emphasize the weight of the burden.

Other peculiarities regarding the quality of the machine translated MWEs caught our attention during the manual validation of equivalents. One notable observation was that both tools appeared to favor literal translations[3] more frequently than other alternatives, when this strategy was appropriate. For instance, when translating the sentence containing the MWE "hot potato," both Google Translate and ChatGPT provided the literal equivalent *batata quente* to refer to a delicate matter. While that translation is valid and conventional, it is worth noting that this particular MWE can also be translated using other non-literal equivalents, such as *abacaxi* [lit. pineapple] and *pepino* [lit. cucumber], both of which are equally suitable and conventional.

We also observed that both machine translators were occasionally able to provide alternative equivalents when a literal translation was not suitable. As an illustration, the

[3] In this paper, we consider literal translation as it was outlined by [24, p. 91]: "maximally close to the SL form, but nevertheless grammatical".

MWE “baby blues” does not have a suitable Brazilian Portuguese equivalent that communicates the exact circumstance conveyed in English, i.e. occasional feelings of mild sadness that some mothers experience in the first few days after having a baby. Although *depressão pós-parto* is indiscriminately used in Brazilian Portuguese to describe a condition like “baby blues,” it actually refers to a much more severe state – “postpartum depression” –, entailing profound feelings of sadness, anxiety, and exhaustion, often hindering mothers from properly caring for themselves and their newborns. Both Google Translate and chatGPT kept that MWE in its English form, which was an adequate way of solving the problem by resorting to a loan [24].

ChatGPT adopted a similar procedure to deal with the MWE “sugar daddy” and even included quotation marks to feature the figurative sense of it, whereas Google Translate resorted to omission, when challenged with the task of translating “silver screen.”

While there were instances, as demonstrated, where both MT systems effectively addressed challenges, we observed occasional failures in evaluating the feasibility of equivalents within a specific context, disregarding contextual hints. This occasionally resulted in amusing or even nonsensical outcomes, as exemplified by the following literal translation:

Example 2

Source-text	Rachel, our stage actress, shows up in the interrogation room looking like a **flower child** from the 60s, and she plays that role while she’s being bombarded by very difficult questions from some very serious homicide detectives.
Google Translation	Rachel, nossa atriz de teatro, aparece na sala de interrogatório parecendo uma **criança das flores** [lit. child of the flowers] dos anos 60, e ela desempenha esse papel enquanto é bombardeada por perguntas muito difíceis de alguns detetives de homicídios muito sérios.
ChatGPT Translation	Rachel, nossa atriz de teatro, aparece na sala de interrogatórios parecendo uma **filha das flores** [lit. daughter of the flowers] dos anos 60, e desempenha esse papel enquanto é bombardeada por perguntas muito difíceis de alguns detetives de homicídios muito sérios.

From the previous example, we observe that both MT systems disregarded the context “looking like a [...] from the 60s” and failed to render a suitable counterpart in Portuguese – *hippie, riponga, bicho-grilo,* etc.

In other cases, the translations were not inadequate or literal, but the MWEs ended up losing their figurative essence or deviating from conventionality, as in the following translations of “lip service”:

Example 3

Source-text	There was a lot of **lip service** about doing whatever was necessary to make the environment accessible and doing whatever was necessary to make accommodations' but there was a little bit of a disconnect between the words and the actions.
Google Translation	Houve muita **conversa** [lit. chat] sobre fazer o que fosse necessário para tornar o ambiente acessível e fazer o que fosse necessário para fazer as acomodações, mas havia uma certa desconexão entre as palavras e as ações.
ChatGPT Translation	Houve muitos **discursos vazios** [lit. empty discourses] sobre "fazer o que for necessário para tornar o ambiente acessível" e "fazer o que for necessário para fazer acomodações", mas houve um pouco de desconexão entre as palavras e as ações.

Google's translation is appropriate and conventional, although it fails in capturing the idiomaticity conveyed by the English compound. On the other hand, ChatGPT's selection could not be confirmed as a common pattern, based on a search in the Brazilian corpus. For the MWE above, HT suggested three different equivalents that would maintain the idiomaticity of the English MWE, all of which are highly prevalent in authentic Brazilian texts: *conversa fiada*, *papo furado*, and *conversa mole*.

Another translation problem observed particularly with Google Translate was the rendering of equivalents typically from European Portuguese. This may be due to the tool's inability to allow users to choose from different language variants. From our list, two MWEs – "cellular phone" and "zebra crossing" – were translated by terms used in Portugal – *telemóvel* and *passadeira*, respectively, instead of the conventional Brazilian terms *telefone celular* and *faixa de pedestre*, adequately suggested by ChatGPT.

Despite the limitations demonstrated, MT systems were occasionally able to provide appropriate and idiomatic translation solutions for the MWEs which had not been proposed by humans. For example, both systems offered the translation choice *golpe duplo* for the MWE "double whammy," not only preserving the idiomatic nature of the compound but also selecting an expression that retains the notion of something negative, which was not captured by the human's choice *dose dupla* [lit. double dose].

5 Conclusions and Future Works

In this study, we compared Machine Translation (MT) with human translations (HT) of English non-compositional (NC) and partly compositional (PC) multiword expressions

(MWEs) into Portuguese. HT served as our gold standard, considering idiomaticity and conventionality parameters for suggested equivalents, all of which were validated in corpora.

Both Google Translate and ChatGPT 3.5 demonstrated capabilities in handling MWEs, offering diverse translation solutions and often distinguishing between different senses according to context. However, MT systems tended to offer less idiomatic options than human translators, who can devise more creative and diverse equivalents.

Our hypothesis that MT performs better with PC than with NC non-compositional MWEs has been validated. This finding highlights the significant challenges both systems encounter when attempting to recognize and accurately translate idiomatic expressions.

ChatGPT 3.5 had a greater number of matches in terms of raw frequency and adequacy for NC MWEs, whereas Google Translate exhibited a slight advantage in handling PC MWEs. No statistically significant difference was identified between the two MT systems.

A qualitative investigation into MT mismatches revealed mixed performance in recognizing the idiomaticity of MWEs and proposing suitable translations. Both ChatGPT and Google Translate occasionally failed to maintain metaphorical characteristics and present conventional equivalents. Google Translate also provided inadequate translations for the Brazilian context.

Our research has limitations, including a restricted number of translators to establish gold standard equivalents, and a small sample size of expressions. It remains uncertain whether the provided context was sufficient in all cases, especially when an additional layer of complexity, such as disambiguation, may be needed. Future analyses with longer contexts could provide further insights into MT of MWEs.

Moving forward, additional research and development efforts are necessary to tackle the complexities inherent in translating MWEs. The next step involves training translators in rendering sentences containing MWEs within a limited timeframe. Additionally, we plan to evaluate the performance of both human and machine translators in rendering compounds typically used idiomatically when they appear in their literal senses, thus deviating from their more common usage, such as “spelling bee” in the sentence “Frank was spelling bee so his five-year-old son could write it.”

Acknowledgments. This study was conducted during the first author’s post-doctoral position at the University of Sheffield, funded by Capes (the Brazilian Coordination for the Improvement of Higher Education Personnel) (grant number 88887.838833/2023-00).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Rosa, G. N. The use of machine translation to support literary translation into an L2: a study in the language pair Brazilian Portuguese-English. Dissertation. European Masters in Technology for Translation and Interpreting. Ghent (2023)
2. Borges, T. M., Pimentel, J. M. M.: Avaliação humana da tradução automática de combinações lexicais especializadas: o caso do Google Translate e do DeepL. Belas Infiéis, Brasília, 9(4), pp. 21-43 (2020)
3. Ramisch, C. A generic and open framework for multiword expressions treatment: from acquisition to applications. Thesis (PhD). Universidade Federal do Rio Grande do Sul. Porto Alegre (2012)
4. Ramisch, C., Cordeiro, S., Zilio, L., Idiart, M., Villavicencio, A., Wilkens, R.: How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality. In: Erk, K., Smith, N. A. (eds.) Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 156–161, Association for Computational Linguistics, Berlin (2016)
5. Stewart, D. Conventionality, Creativity and Translated Text: The Implications of Electronic Corpora in Translation. In M. Olohan (Ed.), Intercultural Faultlines, pp. 73–91. St. Jerome Pub (2000)
6. Fillmore, C. Innocence: A Second Idealization for Linguistics. Berkeley: Linguistics Society (1979)
7. Rothwell, A., Moorkens, J., Fernández-Parra, M., Drugan, J., Austermuehl, F. Translation tools and technologies. London/New York: Routledge (2023)
8. Jurafsky, D., Martin, J. H. Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Unpublished manuscript. (2023, February 3), https://web.stanford.edu/~jurafsky/slp3/ed3bookfeb3_2024.pdf, last accessed 2024/03/20
9. Pérez-Ortiz, J. A., Forcada, M. L.; Sánchez-Martínez, F. How neural machine translation works. In: Kenny, D. (ed.): Machine translation for everyone: Empowering users in the age of artificial intelligence, Berlin: Language Science Press. pp. 141–164 (2022)
10. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A. A Comprehensive Overview of Large Language Models. arXiv:2307.06435v8 [cs.CL] (20 Feb 2024), https://arxiv.org/pdf/2307.06435.pdf, last accessed 2024/04/02
11. Brown, T. B. et al. Language Models are Few-Shot Learners. Computer Science > Computation and Language arXiv:2005.14165v4 (2020), https://arxiv.org/abs/2005.14165, last accessed 2024/03/14
12. Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., Tu, Z.: Is ChatGPT a Good Translator? Yes With GPT-4 As The Engine. arXiv:2301.08745v4 (2023), https://arxiv.org/pdf/2301.08745.pdf, last accessed 2024/03/10.
13. Sinclair, J. Corpus, Concordance, Collocation. Oxford: Oxford University Press (1991).
14. Tagnin, S. E. O. O jeito que a gente diz: combinações consagradas em inglês e português. São Paulo: Disal (2013)
15. Levin, B. Building a Lexicon: The Contribution of Linguistics. International Journal of Lexicography, 4(3), pp. 205-226 (1991)
16. Pérez Hernández, C., Ortiz, A. M., Faber, P. Lexicografía Computacional y Lexicografía de Corpus. Revista Española de Lingüística Aplicada: Panorama de la investigación en

lingüística informática, v. 1, p. 175-214 (1999), https://www.researchgate.net/publication/28106208_Lexicografia_computacional_y_lexicografia_de_corpus, last accessed 2024/03/20.
17. Hassan, H. et al. Achieving human parity on automatic Chinese to English news translation. Microsoft Report, arXiv:1803.05567 (2018), https://www.microsoft.com/en-us/research/publication/achieving-human-parity-on-automatic-chinese-to-english-news-translation/, last accessed 2024/02/20
18. Wu, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv, pp. 1-23 (2016), https://arxiv.org/pdf/1609.08144.pdf, last accessed 2024/01/25
19. Teixeira, M. O jogo da avaliação: um estudo prático sobre tradução automática. 2018. Dissertation thesis. Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro (2018), https://www.maxwell.vrac.puc-rio.br/41711/41711.PDF, last accessed 2024/02/15
20. Koehn, P.; Knowles, R. Six challenges for neural machine translation. In: Luong, T., Birch, A., Neubig, G., Finch, A. (eds.) Proceedings of the First Workshop on Neural Machine Translation. Vancouver (Canada), 2017. pp. 28-39. https://aclanthology.org/W17-3204.pdf, last accessed 2024/01/25.
21. Cordeiro, S.; Villavicencio, A.; Idiart, M.; Ramisch, C. Unsupervised Compositionality Prediction of Nominal Compounds. Computational Linguistics (2019), https://aclanthology.org/J19-1001.pdf, last accessed 2024/04/14
22. Davies, Mark. Corpus do Português: Web/Dialects, (2016-) http://www.corpusdoportugues.org/web-dial/, last accessed 2024/01/09.
23. Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA), https://www.english-corpora.org/coca/, last accessed 204/01/09.
24. Chesterman, A. Memes of Translation: The Spread of Ideas in Translation Theory. Amsterdam-Philadelphia: John Benjamins (1997)

Appendix A. List of non-compositional (NC) multiword expressions (MWE) in English, with their up to four human "gold standard" translations (HT) and machine translations (MT) by Google and ChatGPT.

MWE	HT 1	HT 2	HT 3	HT 4	Google	ChatGPT
agony aunt	conselheira sentimental				tia agonizante	conselheira sentimental
backroom boy	coadjuvante	figurante	come quieto		garoto de bastidor	garoto de bastidor
bad apple	maçã podre	laranja podre			maçã podre	maçã podre
bad hat	mau caráter				péssimo chapéu	pessoa mal-intencionada
banana republic	república das bananas				república das bananas	república das bananas
basket case	caso perdido	caso sem solução			caso perdido	estado deplorável
big cheese	chefão	manda-chuva			grande queijo	figurão
big fish	peixe grande	figurão			peixe grande	grande peixe
big picture	todo o contexto	panorama geral	todo		quadro geral	panorama geral
big wig	bambambã	vip	poderoso		peruca grande	importante
black box	caixa preta				caixa preta	mistério
blame game	jogo de empurra				jogo da culpa	jogar a culpa
blind alley	beco sem saída	encruzilhada			beco sem saída	beco sem saída
blood bath	banho de sangue	derramamento de sangue			banho de sangue	banho de sangue
blue blood	sangue azul				sangue azul	sangue azul
blue print	projeto	planta			projeto	esboço
box office	bilheteria				bilheteria	bilheteria
brain drain	fuga de cérebros	fuga de talentos			fuga de cérebros	fuga de cérebros
brain teaser	quebra-cabeça	desafio	charada		quebra-cabeça	quebra-cabeça
brass ring	oportunidade de ouro	chance de ouro			símbolo	anel de latão
brick wall	pedra no caminho	obstáculo			parede de tijolos	parede de tijolos
busy bee	pessoa que faz mil coisas ao mesmo tempo	atarefado			abelha muito ocupada	workaholic

cash cow	galinha dos ovos de ouro	mina de ouro			vaca leiteira	mina de dinheiro
close call	por um triz	por pouco	susto		por pouco	por um triz
closed book	mistério	enigma	livro fechado		livro fechado	livro fechado
cloud nine	nas nuvens				nas nuvens	nas nuvens
couch potato	sedentário	preguiçoso			preguiçoso	preguiçoso
crash course	curso intensivo	intensivão			curso intensivo	curso intensivo
crocodile tear	lágrima de crocodilo				lágrima de crocodilo	lágrima de crocodilo
cutting edge	vanguarda				de ponta	de ponta
damp squib	balde de água fria	ducha de água fria			final úmido	desanimador
dark horse	azarão				azarão	azarão
dead end	beco sem saída	encruzilhada	dilema		beco sem saída	beco sem saída
diamond wedding	bodas de diamante				casamento de diamante	casamento de diamante
double cross	jogo duplo				traição	traição
double dutch	grego				holandês duplo	grego
double whammy	dose dupla				golpe duplo	golpe duplo
dream ticket	par perfeito	dupla perfeita			bilhete dos sonhos	dream ticket
eager beaver	pau pra toda obra	pé de boi			castor	ávido por progresso
elbow grease	esforço físico	muque	força		esforço	esforço
elbow room	margem de manobra	espaço			espaço	espaço
eye candy	colírio para os olhos	agradável aos olhos	rosto bonito	rostinho bonito	colírio para os olhos	rosto bonito
face value	ao pé da letra	literalmente			ao pé da letra	ao pé da letra
fashion plate	fashionista	consultor de moda			marca de moda	fashionista
flower child	hippie	riponga	bicho-grilo		criança das flores	filha das flores
foot soldier	peão	soldado	soldado raso		soldado de infantaria	soldado raso
front man	líder	porta-voz	vocalista	representante	vocalista	líder
glass ceiling	desigualdade	teto de vidro			teto de vidro	teto de vidro
goose egg	galo	galo na cabeça			ovo de ganso	galo
grandfather clock	relógio cuco	cuco			relógio de pêndulo	relógio de avô
grass root	de base				popular	de base

graveyard shift	turno da noite	turno noturno			turno da noite	turno da noite
gravy train	mina de ouro	dinheiro fácil	mamata		trem da alegria	trem de dinheiro
grey matter	massa cinzenta				massa cinzenta	massa cinzenta
guilt trip	consciência pesada	sentimento de culpa			sensação de culpa	viagem de culpa
guinea pig	cobaia	rato de laboratório			cobaia	cobaia
half wit	lesado	miolo mole			idiota	tolo
hard shoulder	acostamento				acostamento	acostamento
head hunter	recrutador	headhunter	head hunter		caçador de cabeças	head hunter
heavy cross	fardo	fardo pesado	calvário	cruz pesada	pesada cruz	cruz pesada
hot potato	batata quente	pepino	abacaxi		batata quente	batata quente
inner circle	círculo interno	círculo íntimo			público interno	círculo interno
ivory tower	lugar privilegiado	torre de marfim			torre de marfim	torre de marfim
kangaroo court	julgamento sumário	tribunal arbitrário	tribunal armado		tribunal canguru	tribunal arbitrário
lip service	conversa fiada	papo furado	conversa mole	conversa pra boi dormir	conversa	discurso vazio
low profile	discreto	low profile			discreto	discreto
melting pot	caldeirão	mistureba			caldeirão	caldeirão
monkey business	trabalho sujo	palhaçada	presepada	sujeira	confusão	coisa
mother tongue	língua materna	língua nativa	idioma materno		língua materna	língua materna
nest egg	poupança	pé-de-meia	economias	patrimônio	pé-de-meia	bom dinheiro
night owl	coruja	vampiro	notívago	da noite	pessoa noturna	coruja
nut case	maluco	lelé da cuca			maluco	lunático
old flame	antiga paixão				antigo namorado	antigo amor
old hat	fora de moda	demodê	ultrapassado		ultrapassado	lugar comum
old timer	veterano	velha guarda			veterano	mais velho
pipe dream	sonho distante				sonho	devaneio
poison pill	poison pill	pílula de veneno			veneno absoluto	pílula venenosa
rat race	corrida desenfreada	corrida maluca	corrida insana	corrida contra o tempo	corrida desenfreada	corrida dos ratos
rat run	atalho	caminho alternativo			corrida de ratos	rota de fuga
rock bottom	fundo do poço				fundo do poço	fundo do poço

rocket science	bicho de sete cabeças				ciência de foguete	ciência de foguete
sacred cow	vaca sagrada	relíquia			vaca sagrada	vaca sagrada
shelf life	vida de prateleira	tempo de prateleira			vida útil	validade
shrinking violet	bicho do mato				violeta encolhida	violeta encolhida
silver bullet	bala de prata	solução milagrosa			solução mágica	solução milagrosa
silver lining	boa notícia	lado bom			lado positivo	lado positivo
silver spoon	berço de ouro				colher de prata	berço de ouro
sitting duck	presa fácil	alvo fácil			alvo fácil	alvo fácil
smoke screen	cortina de fumaça				cortina de fumaça	cortina de fumaça
smoking gun	prova cabal	prova definitiva			arma fumegante	arma fumegante
smoking jacket	paletó	blazer			smoking	casaco
snail mail	correio	correio comum			correio tradicional	de caracol
snake oil	poção mágica	cura milagrosa			óleo de cobra	óleo de cobra
spinning jenny	tear				fiação jenny	máquina de fiar
sugar daddy	sugar daddy	velho da lancha	"sugar daddy"		sugar daddy	"sugar daddy"
swan song	gran finale				canto do cisne	cisne song
tennis elbow	cotovelo de tenista	tendinite			cotovelo de tenista	cotovelo de tenista
think tank	laboratório de ideias	centro de estudos	think tank		grupo de reflexão	think tank
top dog	cara	líder	chefão	alfa	líder	primeira
wet blanket	estraga-prazeres				cobertor molhado	desanimador
zebra crossing	faixa de pedestre	faixa de segurança			passadeira	faixa de pedestre

Appendix B. List of partly compositional (PC) multiword expressions (MWE) in English, with their up to four human "gold standard" translations (HT) and machine translations (MT) by Google and ChatGPT.

MWE	**HT1**	**HT2**	**HT3**	**HT4**	**Google**	**ChatGPT**
academy award	Oscar				Oscar	Oscar
acid test	prova de fogo	prova definitiva			teste ácido	teste ácido
ancient	águas passadas	passado			história antiga	história antiga

history						
arcade game	arcade	jogo de arcade	flipe rama	jogo de fliperama	jogo de arcade	jogo de fliperama
armchair critic	metido a crítico				crítico de gabinete	crítico de poltrona
baby blues	depressão pós-parto				baby blues	baby blues
beauty sleep	sono de beleza	sono da beleza			sono de beleza	sono da beleza
benign tumour	tumor benigno				tumor benigno	tumor benigno
best man	padrinho	padrinho de casamento			padrinho	padrinho
black operation	operação secreta	missão secreta			operação negra	operação secreta
bow tie	gravata borboleta				gravata borboleta	gravata borboleta
bull market	mercado em alta	bull market			mercado altista	mercado de alta
car park	estacionamento				estacionamento	estacionamento
carpet bombing	bombardeio intenso				bombardeio massivo	bombardeio de tapete
cellular phone	celular	telefone celular			telemóvel	telefone celular
chain reaction	reação em cadeia	efeito dominó			reação em cadeia	reação em cadeia
cheat sheet	cola				folha de dicas	cola
china clay	caulim	caulinita			argila da China	caulim da China
cocktail dress	traje esporte fino	vestido de festa			vestido de coquetel	vestido de coquetel
con artist	golpista				vigarista	vigarista
contact lens	lente de contato				lente de contato	lente de contato
copy cat	imitador	imitação			imitação	imitação
cotton candy	algodão doce				algodão doce	algodão doce
dirty money	dinheiro sujo				dinheiro sujo	dinheiro sujo
dirty word	palavrão	palavra de baixo calão	nome feio		palavrão	palavra suja
disc jockey	disc jockey	DJ			disc jockey	DJ
dry land	terra firme				terra firme	terra seca
dry wall	drywall				parede de gesso	gesso
dutch courage	coragem líquida				coragem holandesa	coragem
end user	consumidor final	usuário final	consumidor		usuário final	usuário final
eternal rest	descanso eterno	morte			descanso eterno	descanso eterno
fairy tale	conto de fadas				conto de fadas	conto de fadas

fall guy	bode expiatório	boi de piranha	gaiato	bode expiatório	bode expiatório
field work	trabalho de campo			trabalho de campo	trabalho de campo
fine line	linha tênue			linha tênue	linha tênue
firing line	linha de fogo	linha de tiro		linha de tiro	linha de tiro
fish story	história de pescador	lorota	história pra boi dormir	história do peixe	história de peixe
flea market	mercado de pulgas			mercado de pulgas	feira de pulgas
fresh water	água doce	água fresca		água doce	água doce
front runner	favorito			favorito	líder
game plan	plano de jogo	estratégia		plano de jogo	plano de jogo
ghost town	cidade fantasma			cidade fantasma	cidade fantasma
gold mine	mina de ouro			mina de ouro	mina de ouro
half sister	meia-irmã			meia-irmã	meia-irmã
hard drive	HD	disco rígido		disco rígido	disco rígido
head teacher	diretor			diretor	diretor
hen party	despedida de solteira			festa de despedida de solteira	festa de despedida
high life	vida de luxo	vida luxuosa		vida nobre	vida de luxo
home run	home run			home run	home run
honey trap	isca	armadilha		armadilha de mel	armadilha
injury time	acréscimo			acréscimo	tempo extra
inner product	produto interno			produto interno	produto interno
insider trading	uso indevido de informação privilegiada	insider trading		uso de informação privilegiada	negociações internas
jet lag	jet lag			jet lag	jet lag
leap year	ano bissexto			ano bissexto	ano bissexto
life belt	boia			cinto salva-vidas	colete salva-vidas
life vest	colete salva-vidas			colete salva-vidas	colete salva-vidas
loan shark	agiota			agiota	agiota
loose woman	piriguete	piranha		mulher livre	mulher fácil
lotus position	posição de lótus	perna de índio		posição de lótus	posição de lótus
memory lane	nostalgia	saudosismo		passado	estrada da memória

milk tooth	dente de leite			dente de leite	dente de leite
narrow escape	por um triz	por pouco		por um triz	por pouco
noble gas	gás nobre			gás nobre	gás nobre
number crunching	cálculo			análise de números	cálculo
pain killer	analgésico			analgésico	analgésico
pecking order	hierarquia			hierarquia	ordem de precedência
polo shirt	camisa polo			camisa polo	camisa polo
private eye	detetive particular			detetive particular	detetive particular
rice paper	papel de arroz			papel de arroz	papel de arroz
role model	exemplo a ser seguido	modelo a ser seguido		bom modelo	bom modelo
rush hour	horário de pico	hora do rush	horário do rush	hora do rush	hora do rush
search engine	motor de busca	ferramenta de busca		mecanismo de busca	mecanismo de busca
sex bomb	gostosa	boazuda		bomba sexual	bomba de sexo
silver screen	telona	tela de cinema		(omissão)	cinema
small fry	zé mané	zé ninguém	peixe pequeno	peixe pequeno	pequeno
speed trap	radar	radar de velocidade		armadilha de velocidade	armadilha de velocidade
spelling bee	jogo de soletrar	spelling bee		concurso de ortografia	soletrar
stag night	despedida de solteiro			despedida de solteiro	despedida de solteiro
street girl	garota de programa	puta	prostituta	garota de rua	garota
traffic jam	congestionamento	engarrafamento		engarrafamento	congestionamento
web site	site	website		site	site
white noise	ruído branco			ruído branco	ruído branco
word painting	word painting	pintura de palavras		pintura de palavras	pintura de palavras