



BILINGUAL EVALUATION BY PHRASE STRUCTURE

AVALIAÇÃO BILÍNGUE POR ESTRUTURA FRASAL

DOI: 10.5281/zenodo.23222254



*Alan Henrique Ferreira Silva*¹

*Alana da Silva Magalhães*²

*Júnio Santos Bulhões*³

*Gabriel Wainer*⁴

*Wesley Pacheco Calixto*⁵

ABSTRACT

This article presents a new metric for evaluating translation between two languages. Bilingual evaluation by phrase structure (BEPHS) uses corresponding sentence structure across two languages in order to calculate a translation score for text candidates. By using an electronic dictionary, the metric searches for correlated terms and words, defining the weight of all parts of the sentence. The tests were carried out using samples of texts translated in English-Portuguese and Portuguese-English pairs. Experiments on these two language pairs show the ability for more accurate scoring compared to other classical translation evaluation methods. The results of the experiments were validated by professional human translators.

Keywords: Machine Translation; BEPHS; Translation Evaluation; Bilingual Translation; Evaluation by Phrase Structure.

RESUMO

Este artigo apresenta uma nova métrica para avaliar a tradução entre dois idiomas. A avaliação bilíngue por estrutura de frase (BEPHS) utiliza a estrutura de frases correspondentes em ambos os idiomas para calcular uma pontuação de tradução para os textos candidatos. Utilizando um dicionário

1 Doutor. E-mail: alanhfs@gmail.com. ORCID: 0000-0001-5798-4061.

2 Doutora. ORCID: 0000-0003-2027-1783.

3 Doutor. ORCID: 0000-0002-7476-7058.

4 Doutor. ORCID: 0000-0003-3366-9184.

5 Doutor. ORCID: 0000-0002-1928-4432.





eletrônico, a métrica busca termos e palavras correlacionados, definindo o peso de todas as partes da sentença. Os testes foram realizados utilizando amostras de textos traduzidos em pares inglês-português e português-inglês. Experimentos com esses dois pares de idiomas demonstram a possibilidade de uma avaliação mais precisa em comparação com outros métodos clássicos de avaliação de traduções. Os resultados dos experimentos foram validados por tradutores humanos profissionais.

Palavras-chave: Tradução Automática; BEPhS; Avaliação de Traduções; Tradução Bilíngue; Avaliação por Estrutura Frasal.

1. INTRODUCTION

In the world of linguistically and culturally diverse societies, strongly connected both physically and electronically, constant challenges arise for nations of different languages to communicate and seek content in a multilateral manner. The establishment of effective communication relies directly on the ability to understand and be understood in a particular language. Electronic devices have been aiding in the comprehension of various languages since the 1970s (Soanes and Stevenson, 2004). Researchers have been developing techniques and mechanisms to enable translation between languages, using machine translation for various purposes such as authorial obfuscation, authorial identification, and, most importantly, content comprehension across languages.

With technological advancements and the increasing use and necessity of the Internet, the first online translation services emerged in the middle of 1988. However, these resources were paid and faced serious difficulties regarding the quality of the final results. In 1997, more sophisticated and refined online services emerged. Innovations and free online translation tools were made available by Google. The provided tool, Google Translate, incorporates the latest refinements developed in statistical machine translation (Hutchins and Somers, 1992). Some technologies, such as webpage translation, are also available but often encounter difficulties with colloquial expressions and culturally dependent terms⁶ (Hutchins, 1986). Hutchins and Somers (1992) implement and refine the methodology in Automatic Translation, utilizing statistical machine translation. In this work, Hutchins and Somers (1992)

⁶ Expressions used in specific geographic regions or communities with a shared cultural background.





relate existing technologies to the tools available on the Internet. The authors present the advancements generated by the automatic translation applied to online tools, contributing substantially to the field of communication and language processing.

Metrics such as the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), are created to expedite the translation process through software, contrasting the high cost and long time required for human translation (Mathur et al., 2020). Papineni et al. (2002) propose a metric for evaluating the automatic translation that is fast, cost-effective, language-independent, and highly correlated with human judgment. The metric involves comparing the candidate solution sentence with the expected reference sentence, counting corresponding terms between the candidate and the correct reference. The translation results between English and Arabic, Chinese, French, and Spanish are evaluated. The work presents a technique capable of generating a robust and simple numerical valuation to classify translations obtained through machine translation.

2. RELATED WORK

Hatim and Munday (2019) define translation from two perspectives. The first focuses on the rules of translation between a source text T_s and the target text T_t . The second perspective is centered on the tangible product generated by the translator (Hatim and Munday, 2019). However, confusion arises between literal translation and the intended meaning of the message to be conveyed. The potential confusion stems from the interpretation between the direct meaning and the intended significance (Malmkjaer and Windle, 2011). Even when interpretation is excluded, the fields covered by translation are vast and complex, considering all the nuances of meaning in expressions used in different languages (Shuttleworth, 2014).

There are three types of translation: i) intralingual translation, ii) interlingual translation, and iii) intersemiotic translation. Intralingual translation occurs when there is translation within the same language, involving rearrangement or paraphrasing of the content.





Interlingual translation is when translation takes place between two different languages (Wilks, 2009). Intersemiotic translation involves the transmission of a translated message through non-verbal signs, such as music or image. Most authors consider only the interlingual model as a translation (Hatim and Munday, 2019).

Automatic translation is the computational technique used to transform digital text, written in a particular source language, into resulting text in the target language, different from the language of the source text (Bhattacharyya, 2015). The original conception of automatic translation addresses the bilingual translation model (Hutchins and Somers, 1992). The process undertaken by automatic translation systems can employ fully automatic methods and models that involve human intervention.

For every automatic translation system, the primary aim is to achieve high-quality translation that remains faithful to the original meaning. Similar to translations done by humans, computer generated translations are often reviewed by other translators before being published or used on a large scale (Dave et al., 2001). Sometimes, the result of computerized translation can be used as a pre-translation or a draft for the subsequent human-led process, ultimately leading to the generation of the final text (Hutchins, 2003). The basic flow of automatic translation is illustrated in Figure 1, where L_S represents digital text in the **source language**, L_T represents digital text in the **target language**, and AT refers to the computerized automatic translation system.

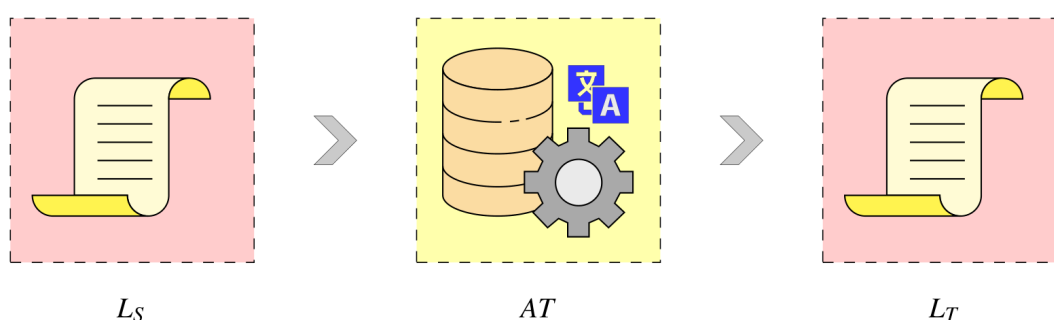


Figure 1: Basic flow of the automatic translation system.



The BLEU evaluation system is considered the standard metric for evaluating digital text translation (Agrawal and Jolly, 2020). The method relies on the direct exploitation of multiple references. Two items are required for this system to be applied: i) the ability to numerically estimate translation similarity, and ii) a reference translation database (Babych, 2014). The score assigned by this algorithm is calculated by counting the number of letters or n -gram word sequences. Each n -gram represents the smallest unit analyzed during the scoring process (Papineni et al., 2002). The BLEU metric utilizes the count of each n -gram in the candidate translation, taking into account reference phrases, and applies a brevity penalty by comparing the n -gram count to the reference texts, regardless of the order of each n -gram (Agrawal and Jolly, 2020).

The METEOR metric is a variant of the BLEU algorithm. This scoring model addresses and corrects several identified flaws in the BLEU model (Banerjee and Lavie, 2005). BLEU metric is precision-oriented, while METEOR method is call-oriented. To assign scores, the precision of n -grams is calculated in relation to the reference texts, and the order of the n -grams between the candidate text and the reference texts is also taken into account (Su et al., 2020; Islam and Mukta, 2022). As the final result of its ranking, the METEOR method generates a score value for each candidate translation separately (Banerjee and Lavie, 2005). It then selects the candidate translation with the highest score. Compared to the BLEU metric, the METEOR metric does not directly check multiple reference sources at once (Zhao et al., 2022). The BLEU and METEOR systems generate evaluations relative to the reference texts, always within the range [0 1], with a value of 1 considered the maximum achievable score. Both methods do not fully consider word order and are not capable of evaluating synonyms, being entirely dependent on the reference sources (Agrawal and Jolly, 2020).

3. BEPHS: BILINGUAL EVALUATION BY PHRASE STRUCTURE FOR MACHINE TRANSLATION

The proposed methodology aims to address the deficiencies of other existing methods. The method is entirely based on the structures of translated sentences between two languages.

Revista *OWL Journal*, Campina Grande - PB, v. 4 n. 10 (2026) - ISSN 2965-2634

A Revista *OWL Journal* está licenciada com uma Licença Creative Commons Atribuição (CC BY)





The validation sequence of the proposed methodology comprises the following steps: selecting an online AT that provides an Application Programming Interface (API) for software integration, selecting an online dictionary with an API, choosing a baseline metric for evaluating candidate translations, defining a new proposed metric, evaluating the translated texts using the baseline and new metric, and evaluating the results by human translators. The proposed textual translation evaluation metric is called Bilingual Evaluation by Phrase Structure (BEPHS). The method is divided into two blocks, with the first block using the initial parameters of the source language sentence T_S and the translated sentence T_T . With these parameters, a query is made to the phrase structure along with the equivalent phrase structures $PS_{(L_S, L_T)}$ in the source language L_S and the target language L_T , as illustrated in Figure 2.

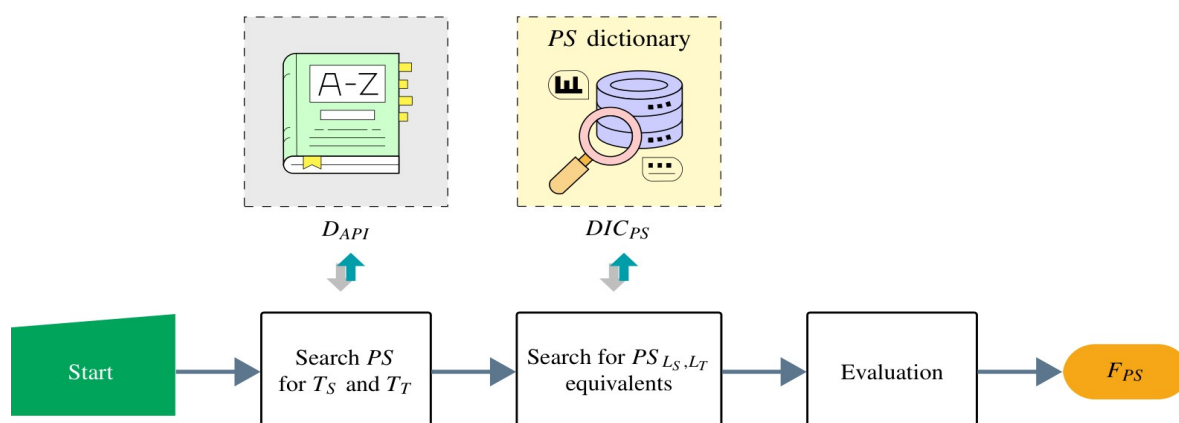


Figure 2: Block for phrase structure evaluation.

The classification of each segment of the received sentences is performed through a query to an external dictionary. Then, a search for the equivalent sentence structures PS_{L_S, L_T} is conducted, and as the final step of this block, a comparison is made between the sentence structures of T_T and the expected PS , given T_S . Finally, a word comparison method is applied (Wagner and Fischer, 1974), resulting in a value F_{PS} within the range [0 1].

In the second block of the BEPHS method, the process of evaluating the translation candidate for solution is continued. This block receives T_S , T_T , PS_{T_T} , PS_{L_S, L_T} and F_{PS} as input



parameters. With these parameters, a search for the PS_{L_T} closest to PS_{T_r} is carried out using the word comparison method. Having chosen the PS closest to the PS_{T_r} , each term of the T_T sentence is compared with the equivalent translation of the T_S . For each validated equivalent translation, F_x is incremented by one unit. At the end of processing all T_T elements, the final value of F_{BEPHS} is given by:

$$F_{BEPHS} = \frac{F_x}{count_p} \cdot F_{PS} \quad (1)$$

where F_{PS} is the evaluation value considering only the phrase structure, $count_p$ is the number of corresponding words found given the reference phrase structure PS_{T_r} , F_x is the sum of correct translations found in T_T and F_{BEPHS} is the final evaluation value within the range [0 1]. Figure 3 illustrates the steps to obtain the final F_{BEPHS} value.

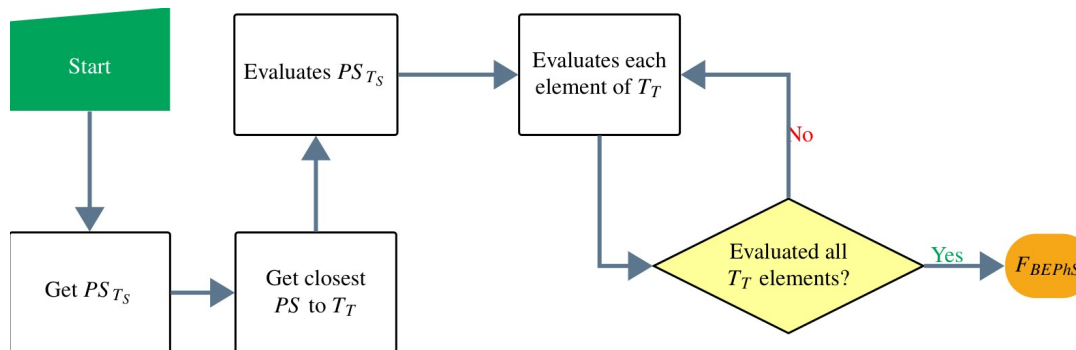


Figure 3: Evaluation block diagram based on translation.

The metric uses the concept of bidirectional translation $L_S \leftrightarrow L_T$ in order to obtain the translated text T_T . The evaluation process is divided into two translation steps and one evaluation step. In the first step, T_S is translated using the $L_S \rightarrow L_T$ process through the AT_{API} . This translation is the forward translation. As a result of the first step, T_T is obtained in the target language L_T . The second step performs a backward translation of T_T using the $L_T \rightarrow L_S$



process, resulting in a new translated text T'_s . The third and final step utilizes a database of sentence structure rules $PS(LS,LT)$ and applies the evaluation function to T_T and T'_s , as shown in Figure 4.

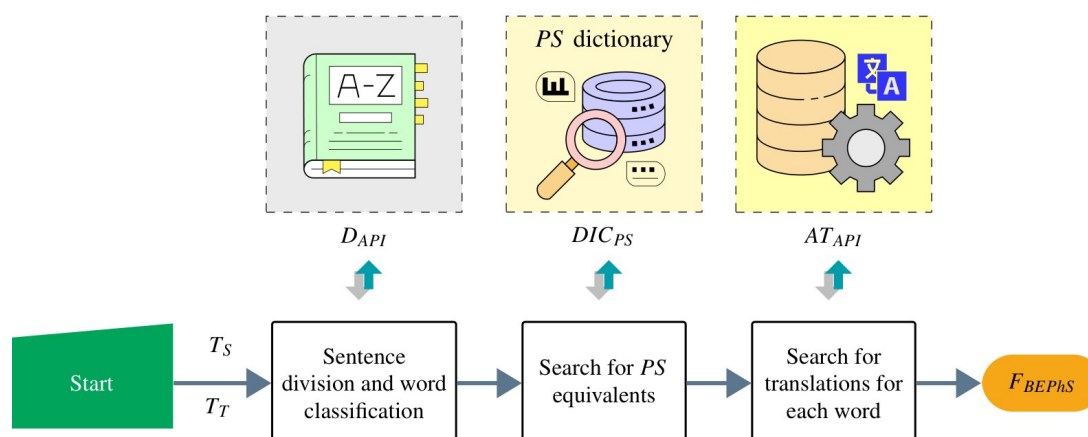


Figure 4: Bilingual Evaluation by Phrase Structure.

The BEPhS method is based on equivalent sentence structures between two languages. Provided that there are API for querying word classifications and a dictionary of sentence equivalence between the two languages, the system will be able to generate evaluations considering the positioning of n -grams, synonyms, and sentence structures.

4. RESULTS

The first action to evaluate machine translation (MT) is to choose the dictionary and translation API application programming interfaces (API). The Oxford Dictionary API was used as the dictionary API, which is a set of tools built for developers. This toolset provides up-to-date monolingual and bilingual data in an accessible and scalable environment, making it suitable for various domains.

As an auxiliary software in MT, the Translate API was used. It was developed by Yandex Technologies and allows developers to integrate machine translation into applications.



The Translate API supports over 90 languages, and it can translate individual words or entire texts, making it possible to incorporate it into various types of applications.

4.1 Case study 1: the book paragraph

In the first case study, the translation is evaluated using the proposed BEPhS metric, the BLEU metric and the METEOR metric. In this case study, a passage from the book “Trading in the zone: master the market with confidence, discipline and a winning attitude”, written by the author Douglas (2001), published in English (source text T_S) and with a printed edition, is analyzed. The official Portuguese version (translated text T_T) was translated by human translators and thus, a passage from the same book in both languages was used to evaluate the metrics. The passage selected to be analyzed appears on page 58 of the English version and page 81 of the Portuguese version, in which both passages are shown in Figure 5.

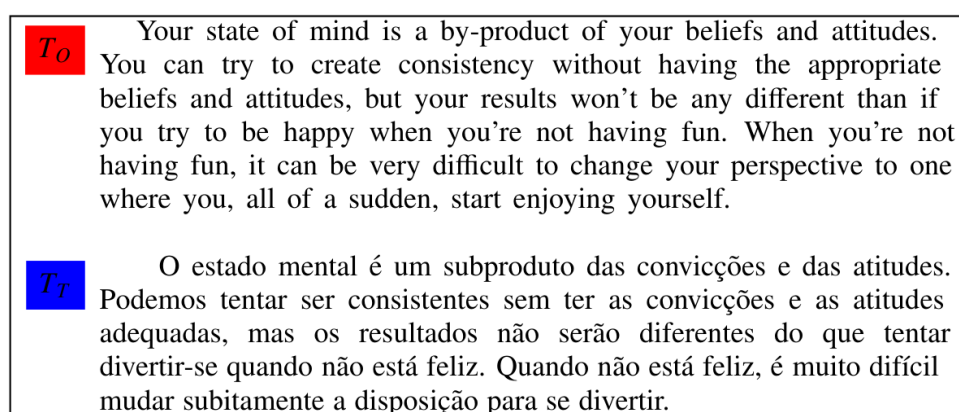


Figure 5: Passage evaluated by BEPhS, BLEU and METEOR metrics.

Each metric has a specific characteristic, and they depend on the initial dataset to perform the evaluation. In this regard, the excerpt T_S was translated into Portuguese T_T using the online translators from Google and Bing. The sentence structures were extracted from the translations generated by the online MT and from the English and Portuguese texts of the book. In order to guarantee the equality of the database for the evaluations, the metrics



received the translations from the MT and the translated text from the Portuguese version of the book for comparison and subsequent evaluation.

Considering that the translation of the book excerpt is from the official version that has undergone several human reviewers and considering that the translation is the best possible, the metric that presents the closest value to the maximum score can be considered the most effective metric. Table 1 displays the evaluations obtained for each metric based on the T_S and T_T datasets. The metrics present values in the interval [0 1] in which 1 is indicative of precise translation and or total accuracy. The evaluation results present a difference of approximately 42% between the metrics, in which the METEOR metric presented the worst evaluation, followed by the BLEU metric, and the BEPhS metric was the one that presented the highest score, indicating that the T_T is likely to be the translation from T_S .

Table 1: Evaluations for the translated text TT.

	BEPhS	BLEU	METEOR
Evaluation	0.46785	0.37827	0.27060

4.2 Sentence translation evaluation and comparison between metrics

For the other case studies, three source texts (T_S) were used, two in Portuguese and one in English, as shown in Table 2. In these cases (Table 2), the sentence structures between the source language (L_S) and the target language (L_T) have corresponding phrase structures PS_{L_S, L_T}

in a predefined database.





Table 2: The source texts used to evaluate the translations.

#	Source Text
T_{S1}	O cavalo é branco.
T_{S2}	If you don't find a way to make money while you sleep, you will work until you die.
T_{S3}	A manga da camisa rasgou.

The $PS_{T_{S1}}$ and $PS_{T_{S3}}$ texts have a single sentence structure, while the $PS_{T_{S2}}$ text has two sentences. The evaluation procedure is performed on each sentence structure separately when using the BEPhS metric. The result of the evaluations for each sentence is calculated through the weighted average between the resulting evaluations and the number of elements in each sentence, given by:

$$F_{BEPhS} = \frac{\sum_{i=1}^n F_{BEPhS_i} \cdot El_i}{\sum_{i=1}^n El_i} \quad (2)$$

where n is the number of evaluated sentences, El_i is the number of elements/words in the evaluated sentence, F_{BEPhS_i} is the evaluation for the i sentence with $i = 1, 2, \dots, n$, and F_{BEPhS} is the weighted average representing the final score for the evaluated text.

For the application of the BLEU and METEOR metrics, reference phrases are needed in order to enable the evaluation calculation. In texts T_{S1} and T_{S3} phrases related to the same theme of the source text were used, swapping verbs and nouns and varying the placement of elements for each reference. In text T_{S2} phrases translated by the automatic translators Google, Bing, and Yandex were used. For the application of the BEPhS metric, previously defined phrase structures were used, taking into account possible translations with similar phrase structures to those of the translated texts.





The human translators were provided with the TO and the expected translations for each pair (T_s , T_T). Ten human translators were consulted, and the final human score for each translation candidate was calculated by taking the arithmetic mean of the scores given by the translators. All scores were assigned in the interval between [0, 1], where 0 represents the worst evaluation, and 1 represents the maximum evaluation for the translation candidate sentence.

4.2.1 Case study 2: simple sentences

For the evaluation of T_{S1} , the phrases presented in Figure 6 are used, which include the source text T_{S1} , the desired translation T_{T1} , the reference sentences for the BLEU and METEOR metrics, and six translation candidates T_{C1i} . Since these are short sentences, human translators were consulted about the possibilities of variations needed to test the ability to evaluate the metrics. The following suggestions were made: i) word repetition, ii) swapping the positions of sentence elements, iii) word removal, and iv) changes in the positions of nouns and verbs.

T_{S1}	–	<i>“o cavalo é branco”</i>
T_{S1}	–	<i>“the horse is white”</i>
Refe.	–	<i>“a white cow” “the horse is galloping” “the horse is alive” “a cow is falling”</i>
Cand.	–	$T_{C1a} \rightarrow$ <i>“the white horse”</i>
		$T_{C1b} \rightarrow$ <i>“the chicken is white”</i>
		$T_{C1c} \rightarrow$ <i>“the the the the”</i>
		$T_{C1d} \rightarrow$ <i>“this horse is black”</i>
		$T_{C1e} \rightarrow$ <i>“white horse is the”</i>
		$T_{C1f} \rightarrow$ <i>“the horse is white”</i>

Figure 6: Data for evaluation of T_{S1} .

Table 3 displays the obtained evaluation scores *Eval* and the absolute distance *Dist* between the evaluations obtained by the metrics and the evaluation performed by a human for



each T_{C1i} candidate of the T_{S1} translation. As shown in Table 4, the BEPhS metric presented the lowest absolute distance when compared to the human evaluation, and in the count of times that the metrics presented the lowest $Dist$, the BEPhS metric obtained the lowest value for five of the six candidates.

Table 3: Evaluations for T_{S1} .

Candidates	BEPhS		BLEU		METEOR		Human V_h
	Eval	Dist	Eval	Dist	Eval	Dist	
T_{C1a}	0.333	0.222	1.000	0.445	0.384	0.171	0.555
T_{C1b}	0.675	0.000	0.750	0.075	0.638	0.037	0.675
T_{C1c}	0.062	0.013	0.750	0.075	0.125	0.050	0.075
T_{C1d}	0.500	0.125	0.500	0.125	0.468	0.157	0.625
T_{C1e}	0.500	0.225	1.000	0.725	0.500	0.225	0.275
T_{C1f}	1.000	0.000	1.000	0.000	0.992	0.008	1.000

Table 4: Analysis with respect to human evaluation for T_{S1} .

	BEPhS	BLEU	METEOR
$Dist_T$	0.58417	1.54500	0.64556
$Dist_M$	0.09736	0.25750	0.10759
$Dist_p$	20.27%	36.17%	24.47%

In Table 4, the total distance $Dist_T$ is the sum of $Dist$ for all candidate phrases (in this case, $n = 6$) for each metric, in relation to the evaluation of human translators. The mean distance $Dist_M$ is obtained by dividing $Dist_T$ by the number of evaluated candidate phrases. The mean percentage distance $Dist_p$ is calculated as follows:

$$Dist_p = \frac{1}{n} \cdot \sum_{i=1}^n \frac{|Eval_i - V_h|}{Eval_i} \quad (3)$$



where V_h represents the value obtained from human evaluation. The distribution of distance values is shown in Figure 7. When considering all metrics in relation to human evaluation, the BEPhS metric did not yield a higher *Dist* for any of the candidates, the METEOR metric had the highest *Dist* for two candidates, and the BLEU metric had the highest *Dist* for four out of six candidates. The BEPhS metric proved to be more effective than the BLEU and METEOR metrics in all analyses conducted on T_{S1} , being closer to human translation with $Dist_p \approx 4\%$ lower than the second best metric.

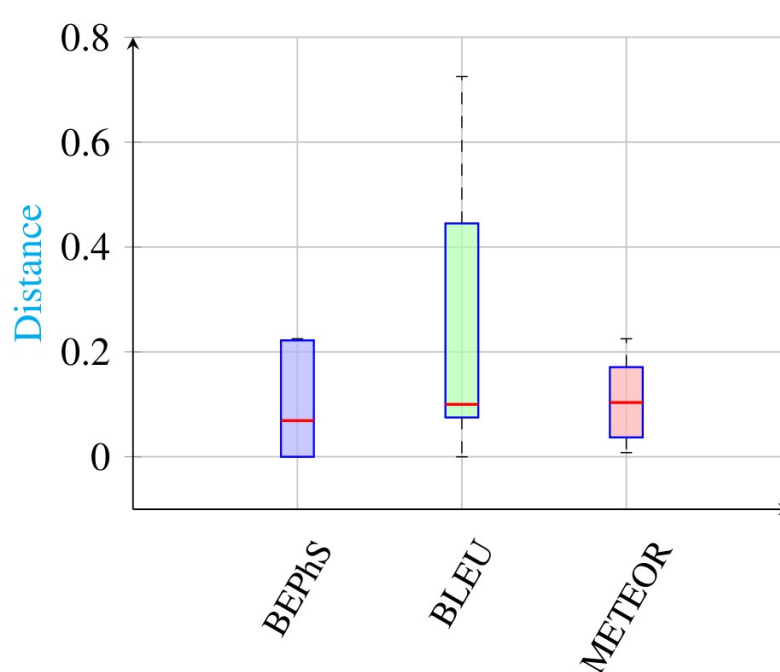


Figure 7: Distance variation analysis for T_{S1} .

4.2.2 Case Study 3: multiple sentences and ordering

For the evaluation of T_{S2} , the data presented in Figure 8 were used. In this specific study, it was necessary to use T_{S2} with multiple sentences separated by commas. As it is a larger sentence compared to T_{S1} , the sentence translated using the MT were provided as a reference base for the BLEU and METEOR methods: i) Google Translator, ii) Bing Translator



and iii) Yandex Translator. Table 5 presents the obtained evaluation values *Eval* and the absolute distances *Dist* for all candidates T_{C2i} evaluated in T_{S2} , and the distribution of distances are shown in Figure 9. From the analysis of Table 5, the BEPhS metric showed the lowest *Dist* for four out of the six translation candidates, while the BLEU metric had the best *Dist* for two out of the six candidates. The BEPhS metric showed a higher *Dist* for the T_{C2b} candidate and the METEOR metric showed a higher *Dist* for all other translation candidates. Table 6 displays the total distance $Dist_T$, the mean distance $Dist_M$, and percentage distance $Dist_P$ given by (3).

T_{S2}	– “if you don’t find a way to make money while you sleep, you will work until you die”
T_{S2}	– “se você não encontrar um jeito de ganhar dinheiro enquanto dorme, você vai trabalhar até morrer”
Refe.	– “se você não encontra uma maneira de ganhar dinheiro enquanto dorme, você vai trabalhar até morrer” – Google Translator “se você não encontrar uma maneira de ganhar dinheiro enquanto você dorme, você vai trabalhar até que você morrer” – Bing Translator “se não encontrares uma maneira de ganhar dinheiro enquanto dormes, trabalharás até morreres” – Yandex Translator
Cand.	– T_{C2a} → “se não achar um jeito de fazer dinheiro enquanto dorme, trabalhará até morrer” T_{C2b} → “você irá trabalhar até morrer, se não achar um jeito de ganhar dinheiro enquanto dorme” T_{C2c} → “se tu não encontrares uma forma de ganhar dinheiro enquanto dormes, trabalharás até morrer” T_{C2d} → “se você não encontrar um caminho para fazer dinheiro enquanto você dorme, você trabalhará até você morrer” T_{C2e} → “se não encontrar você meio para criar dinheiro quando dorme você, trabalhar até morrer você irá” T_{C2f} → “se você não encontrar um jeito de ganhar dinheiro enquanto dorme, você vai trabalhar até morrer”

Figure 8: Data for evaluation of T_{S2} .



Table 5: Evaluations for T_{S2} .

Candidates	BEPHS		BLEU		METEOR		Human V_h
	Eval	Dist	Eval	Dist	Eval	Dist	
T_{C2a}	0.933	0.002	0.615	0.320	0.179	0.756	0.935
T_{C2b}	0.000	0.645	0.561	0.084	0.178	0.467	0.645
T_{C2c}	0.939	0.134	0.857	0.052	0.278	0.527	0.805
T_{C2d}	0.944	0.084	0.705	0.125	0.251	0.609	0.860
T_{C2e}	0.640	0.205	0.562	0.725	0.135	0.710	0.845
T_{C2f}	0.939	0.061	0.875	0.000	0.282	0.718	1.000

Table 6: Analysis with respect to human evaluation for T_{S2} .

	BEPHS	BLEU	METEOR
$Dist_T$	1.03030	0.91708	3.68385
$Dist_M$	0.17172	0.15285	0.61398
$Dist_p$	23.91%	18.03%	73.80%

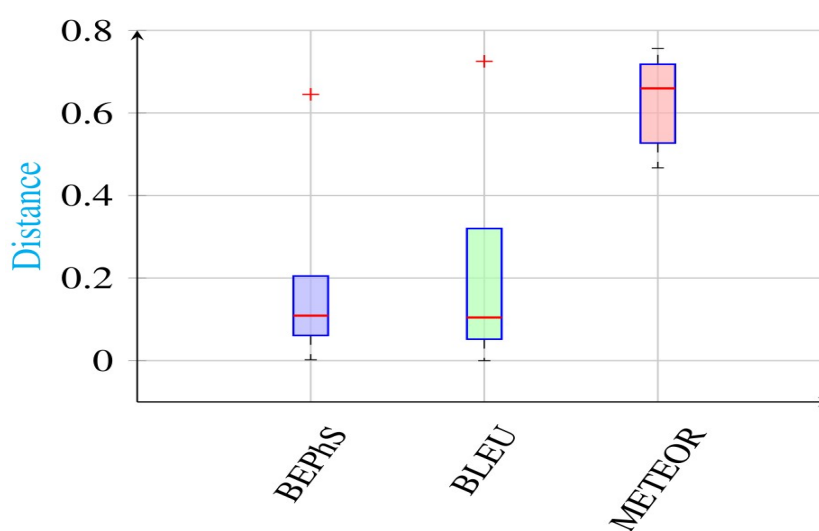


Figure 9: Distance variation analysis for T_{S2} .



Considering the evaluation of the metrics and taking into account the human evaluation of the translations, the METEOR metric presented a higher *Dist* for five out of the six candidates. The BEPhS metric showed the highest *Dist* for one of the six candidates while the BLEU metric did not show the highest *Dist* for any of the candidates. The BEPhS metric proves to be more accurate than the BLEU and METEOR metrics when considering the number of times it presented the lowest *Dist*. However, it presented $Dist_p > 5\%$ of the BLEU metric, which obtained only two smaller *Dist* results. This difference is due to the fact that the references provided to the BLEU and METEOR methods have a high number of coinciding *n*-grams (words) with the expected translation.

4.2.3 Case Study 4: synonym

For the evaluation of T_{S3} , the data presented in Figure 10 were used. In this case study, words with multiple meanings were chosen, which depending on the application can completely change the final meaning of the phrase. Human translators were consulted to provide coherent references for the BLEU and METEOR methods, as well as translation candidates that could test the accuracy of the applied metrics.

T_{S3}	–	<i>“a manga da camisa rasgou”</i>
T_{S3}	–	<i>“the shirt sleeve tore”</i>
Refe.	–	<i>“the shirt” “the shirt is tore” “the shirt is black” “a shirt can tear” “shirt sleeve”</i>
Cand.	–	$T_{C3a} \rightarrow$ <i>“the tore sleeve”</i>
		$T_{C3b} \rightarrow$ <i>“the collar has tear”</i>
		$T_{C3c} \rightarrow$ <i>“the short sleeve”</i>
		$T_{C3d} \rightarrow$ <i>“the shirt is torn”</i>
		$T_{C3e} \rightarrow$ <i>“shirt sleeve the rent”</i>
		$T_{C3f} \rightarrow$ <i>“the shirt sleeve tore”</i>

Figure 10: Data for evaluation of T_{S3} .



Table 7 presents the evaluations obtained by the three metrics and by the human translator. For T_{S3} , the BEPhS metric obtained the lowest *Dist*, followed by the METEOR metric, and then the BLEU metric. When considering the number of times that the metrics had the lowest *Dist* compared to the human evaluation, the BEPhS metric had the lowest *Dist* for four out of six candidates, the BLEU metric for two candidates, while the METEOR metric had the lowest distance for one candidate, as shown in Figure 11.

Table 7: Evaluations for T_{S3} .

Candidates	BEPhS		BLEU		METEOR		Human V_h
	Eval	Dist	Eval	Dist	Eval	Dist	
T_{C3a}	0.333	0.017	1.000	0.650	0.256	0.094	0.350
T_{C3b}	0.136	0.319	0.500	0.045	0.227	0.228	0.455
T_{C3c}	0.675	0.585	0.500	0.410	0.227	0.137	0.090
T_{C3d}	0.150	0.055	0.500	0.295	0.468	0.263	0.205
T_{C3e}	0.500	0.410	0.750	0.660	0.852	0.762	0.090
T_{C3f}	1.000	0.000	1.000	0.000	0.852	0.148	1.000

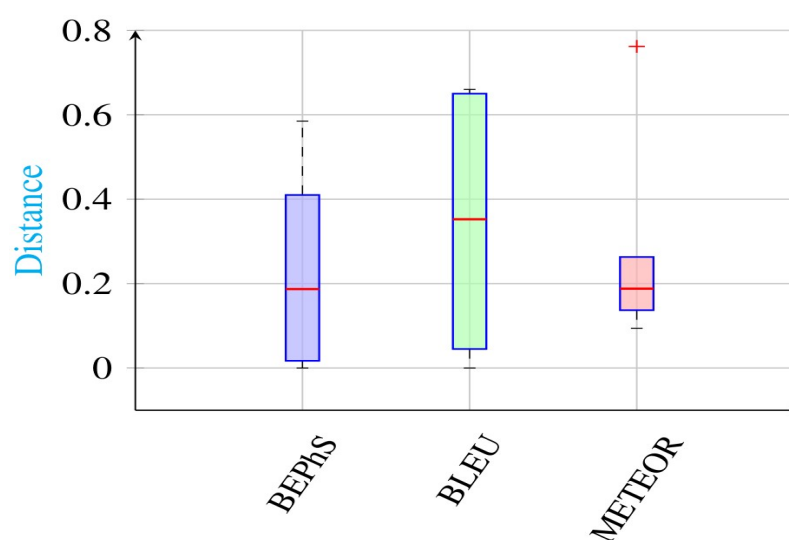


Figure 11: Distance variation analysis for T_{S3} .



In order to verify the difference in scores obtained by each metric compared to the evaluation of human translators, the $Dist_T$, $Dist_M$ and $Dist_P$ were calculated and displayed in Table 8. Considering the evaluation of the three metrics and taking into account the evaluation by human translators, the metrics showed similar $Dist_P$. Each metric presented higher $Dist$ for two of the six candidates. The BEPhS metric was more accurate than the BLEU metric and the METEOR metric, when taking into account the number of times it presented the shortest distance, surpassing the BLEU metric by 5.46%.

Table 8: Analysis with respect to human evaluation for T_{S3} .

	BEPhS	BLEU	METEOR
$Dist_T$	1.38498	2.06000	1.063234
$Dist_M$	0.23083	0.34333	0.27206
$Dist_P$	45.04%	50.50%	49.61%

5. DISCUSSION

The results obtained in the case studies demonstrated the effectiveness of the BEPhS metric as an evaluation mechanism for MT. The scores obtained by the proposed metric surpassed, for the most part, the accuracy of the other commonly used and established metrics in the literature. The BEPhS metric can be employed as an auxiliary tool for MT, since it presented $\approx 5\%$ higher precision compared to the other metrics.

For the application of the proposed metric, some difficulties were found regarding the creation of equivalent phrasal structures. However, once these structures are defined, they can be stored in a simple dataset and used in new translation evaluations. The use of only simple sentences and their corresponding simple phrase structures posed challenges in the process of defining the metric. However, incorporating more elaborate phrasal structures made it possible to debug and improve the metric for application in texts of varied length, given that any text can be divided into parts.





Two different tests were performed to validate the metric. In each test, a different scenario was approached, as follows: i) in the first test, an official book translation was used as an absolute comparison source for the translation, and ii) in the second test (T_{s1} , T_{s2} , and T_{s3}), the absolute comparison source were the translated texts validated by ten different human translators. Analyzing the behavior of the metric, it is observed that it is not as precise as human evaluation, similar to other metrics. However, it demonstrates the ability to measure MT quality. The proposed metric presents limitations when dealing with lengthy texts, requiring them to be divided into smaller parts and limiting the number of verbs within phrases.

At the beginning of the work, mistakes occurred regarding the choice of technologies/tools to be used. For instance, since the aim of the work was to present the proposed metric and MT, everything would be developed in Matlab. However, this tool did not provide the necessary resources. Subsequently, after starting the studies, Python was adopted as the language because it has a library with the BLEU and METEOR metrics, as well as its comprehensive support for integrating the developed system with external API.

In the development of the work, possibilities were identified that can still be explored, such as the use of artificial intelligence (AI) applied in conjunction with the BEPhS method. Due to the existence of advanced studies using neural networks and AI, it would not be plausible to invest in studies applying the BEPhS metric in the evaluation of texts with idiomatic expressions, such as regional expressions and slang.

6. CONCLUSION

This work presents the creation of a metric for the evaluation of text translation and the application of this metric in the evaluation of machine and human translation texts. The proposed phrase structure evaluation method enabled the creation of a new alternative for machine translation evaluation. The initially raised hypothesis was proven through the results





obtained, showing the ability to use the phrase structure of texts as a basis for a MT evaluation tool, contributing to the creation of a new methodology.

The use of API enabled the extraction of data for each word. The extraction of grammatical data allowed for the creation of a dictionary of phrase structures and the processing of information from the original texts and translation candidates. The generation of processed information and the dictionary of phrase structures enabled the development of a metric based on sentence correspondence between different languages, providing a tool for evaluating translation candidates. The evaluations obtained by the proposed metric could be compared to existing metrics and the evaluation of human translators.

The proposed method was used in order to evaluate a paragraph of text extracted from a book, obtaining a better evaluation than the two existing metrics. When applied to individual sentences, the BEPhS metric yielded more accurate evaluations compared to the BLEU and METEOR metrics.

The advantage of the proposed method is the amount of data involved in creating the dictionary of corresponding phrase structures. Due to the proposed structure, the method can be applied in other languages, as long as the corresponding phrase structures and dictionaries between the languages are available. The proposed metric can be used in several segments, such as: i) evaluation of written fluency in immigration programs in which proficiency in certain languages is required, ii) electronic assessment in selection processes, iii) development of improved translation tools and iv) other.

REFERENCES

AGRAWAL, R.; JOLLY, S. K. Big data and machine learning for evaluating machine translation. Cham: Springer International Publishing, 2020. DOI: https://doi.org/10.1007/978-3-030-39119-5_2.

BANERJEE, S.; LAVIE, A. Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. In: PROCEEDINGS OF THE ACL WORKSHOP ON INTRINSIC AND EXTRINSIC EVALUATION MEASURES FOR MACHINE





TRANSLATION AND/OR SUMMARIZATION, jun. 2005, Ann Arbor, Michigan. Proceedings [...]. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65-72. Disponível em: <https://aclanthology.org/W05-0909>.

BABYCH, B. Automated MT evaluation metrics and their limitations. *Revista Tradumàtica: tecnologies de la traducció*, n. 12, p. 464-470, 2014. DOI: <https://doi.org/10.5565/rev/tradumatica.70>.

BHATTACHARYYA, P. Machine translation. 1. ed. Boca Raton: CRC Press, 2015.

DAVE, S.; PARIKH, J.; BHATTACHARYYA, P. Interlingua-based English-Hindi machine translation and language divergence. *Machine Translation*, v. 16, p. 251-304, 2001. DOI: <https://doi.org/10.1023/A:1021902704523>.

DOUGLAS, M. Trading in the zone: master the market with confidence, discipline, and a winning attitude. Upper Saddle River: Prentice Hall Press, 2001.

ESPLÀ-GOMIS, M.; SÁNCHEZ-MARTÍNEZ, F.; FORCADA, M. L. Predicting insertion positions in word-level machine translation quality estimation. *Applied Soft Computing*, v. 76, p. 174-192, 2019. DOI: <https://doi.org/10.1016/j.asoc.2018.11.036>.

HATIM, B.; MUNDAY, J. Translation: an advanced resource book for students. 2. ed. London: Routledge, 2019.

HUTCHINS, W. J. Machine translation: past, present, future. Chichester: Ellis Horwood Ltd., 1986. Disponível em: <https://dl.acm.org/doi/10.5555/6607>.

HUTCHINS, W. J. Machine translation. Chichester: John Wiley and Sons Ltd., 2003.

HUTCHINS, W. J.; SOMERS, H. L. An introduction to machine translation. London: Academic Press, 1992. v. 362.

ISLAM, M. A.; MUKTA, M. S. H. A comprehensive understanding of popular machine translation evaluation metrics. *International Journal of Computational Science and Engineering*, v. 25, n. 5, p. 467-478, 2022. DOI: <https://doi.org/10.1504/IJCSE.2022.126258>.

MALMKJAER, K.; WINDLE, K. The Oxford handbook of translation studies. Oxford: Oxford University Press, 2011. DOI: <https://doi.org/10.1093/oxfordhb/9780199239306.001.0001>.

MATHUR, N.; BALDWIN, T.; COHN, T. Tangled up in bleu: reevaluating the evaluation of automatic machine translation evaluation metrics. arXiv preprint, arXiv:2006.06264, 2020. DOI: <https://doi.org/10.48550/arXiv.2006.06264>.





PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W. J. BLEU: a method for automatic evaluation of machine translation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., jul. 2002, Philadelphia, Pennsylvania. Proceedings [...]. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002. p. 311-318. Disponível em: <https://aclanthology.org/P02-1040>.

SHI, C.; XIANG, Y.; YU, J.; SOOD, K.; GAO, L. Machine translation-based fine-grained comments generation for Solidity smart contracts. Information and Software Technology, v. 153, p. 107065, 2023. DOI: <https://doi.org/10.1016/j.infsof.2022.107065>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950584922001744>.

SHUTTLEWORTH, M. Dictionary of translation studies. 4. ed. London: Routledge, 2014.

SOANES, C.; STEVENSON, A. Concise Oxford English dictionary. 11. ed. Oxford: Oxford University Press, 2004.

SU, C.; HUANG, H.; SHI, S.; JIAN, P.; SHI, X. Neural machine translation with Gumbel Tree-LSTM based encoder. Journal of Visual Communication and Image Representation, v. 71, p. 102811, 2020.

WAGNER, R. A.; FISCHER, M. J. The string-to-string correction problem. Journal of the ACM, v. 21, n. 1, p. 168-173, 1974. DOI: <https://doi.org/10.1145/321796.321811>.

WILKS, Y. Machine translation: its scope and limits. New York: Springer Science & Business Media, 2009.

ZHAO, Y.; KOMACHI, M.; KAJIWARA, T.; CHU, C. Region-attentive multimodal neural machine translation. Neurocomputing, v. 476, p. 1-13, 2022. DOI: <https://doi.org/10.1016/j.neucom.2021.12.076>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231221019317>.

Recebido em: 17/09/2026

Aprovado em: 29/09/2026

Publicado em: 07/10/2026

