Article

CADERNOS de TRADUÇÃO

The role of metrics to assess the quality of British teenage language translation into Spanish and Italian using machine translation tools

Andrés Canga Alonso
University of La Rioja
Logroño, La Rioja, Spain
andres.canga@unirioja.es
https://orcid.org/0000-0002-1578-1626

María Cira Napoletano
University of La Rioja
Logroño, La Rioja, Spain
manapole@unirioja.es
https://orcid.org/0009-0003-6644-8000

Abstract: The rapid evolution of adolescence language, characterized by slang and idiomatic expressions, presents a significant challenge for machine translation systems. Existing research has extensively covered the translation of languages in general; however, there remains a gap in understanding these systems' ability when faced with adolescent language. This study aims at (i) the evaluation and the comparison of the accuracy of the translations of colloquial language by Bing Translator, DeepL and HelsinkiNLP from English into Spanish and Italian, (ii) the validity and reliability of two different metrics (i.e., BLEU, METEOR) to assess the accuracy and quality of MT tools with informal language, and (iii) the analysis of how specific features of teenage slang influence the ability of online tools to generate precise and comprehensible translations 1000-character excerpts from the Linguistic Innovators Corpus were translated in Spanish and Italian using DeepL, Bing Translator, and HelsinkiNLP and assessed using BLEU and METEOR metrics to verify their quality and reliability. Our findings show that teenage slang poses challenges for all tools, particularly with phrasal verbs and idioms. Our results also reveal that METEOR seems to be more reliable to assess British teenage language into Spanish and Italian.

Keywords: machine translation; teenage language; quality assessment; BLEU; METEOR.

1. Introduction

Recent advancements in deep learning algorithms and the availability of vast linguistic data have significantly enhanced the accuracy of machine translation (MT) tools over the past few years. Since Hutchins and Somers (1992) seminal work, there has been a widespread belief that the idea of machines capable of translating human language was unachievable. This perspective arose because the primary challenges in translation were not computational but linguistic; in fact, machines often

DOI https://doi.org/10.5007/2175-7968.2025.e101329

struggled with "[...] lexical ambiguity, syntactic complexity, and vocabulary differences between languages" (Hutchins & Somers, 1992, p. 2). Consequently, as more in-depth studies emerged in the 20th century, it became clear that one of the significant limitations of MT was its ability to handle figurative and creative aspects of language, such as adolescent slang.

Tagliamonte and Denis (2010), Palacios Martínez (2011, 2020, 2021) and Rosyadi Za *et al.* (2023) have examined the use of teenage language in various contexts. However, we consider that there is a lack of research on how these tools address the translation of this type of language into Spanish and Italian. As a matter of fact, most studies focus primarily on English (Nicholas & Bhatia, 2023), not only because it is spoken globally but also because it dominates the computational linguistics' field; as a result, there is more data available in English than in any other language.

Hence, this study aims at (i) the evaluation and the comparison of the accuracy of the translations of colloquial language by Bing Translator, DeepL and HelsinkiNLP from English into Spanish and Italian, (ii) the validity and reliability of two different metrics (i.e., BLEU, METEOR) to assess the accuracy and quality of MT tools with informal language, and (iii) the analysis of how specific features of teenage slang influence the ability of online tools to generate precise and comprehensible translations.

The paper comprises five sections. The first section provides a brief overview of the evolution of machine translation, focusing on today's most used automatic metrics and the typical features of informal youth language. In the second section, the research methodology is detailed, giving information about LIC corpus and the data analysis procedures. Following this, the third section describes the principal findings of the study. The fourth section offers the examination and interpretation of its results. The paper ends by proposing areas for future research addressing its main shortcomings.

2. Literature review

2.1. Advancements in machine translation

Machine translation has been, for years, a fundamental subject of research in the artificial intelligence (AI) field (Das, 2018; Moneus & Sahari, 2024). The emergence of Deep Learning models, particularly neural models, has marked significant milestones in improving the quality of translations generated by automatic tools (Duan *et al.*, 2021; Son & Kim, 2023). In the early stages, Banitz (2020) and Song (2022) assert that the systems prominently featured were the *rule-based machine translation* (RBMT) and the *statistical-based machine translation* (SBMT). The former involves three approaches: (i) the direct method, producing a literal translation; (ii) the Interlingua approach, entailing the conversion of the source language (SL) into an abstract representation; and (iii) the transfer approach (Hutchins & Somers, 1992; Banitz, 2020). Thus, Wang (2023) pinpoints that the RBMT interprets the meaning of the SL by decoding it, then blend it with the linguistic characteristics and grammar regulations of the target language (TL) to produce the translation. Consequently, it relies heavily on parallel corpora and faces challenges to translate texts from a more specialized language or with limited resources, which requires extensive manual effort for rule maintenance.

On the other hand, the SBMT operates within statistical frameworks, involving the acquisition of translation principles and patterns from extensive parallel corpora. These corpora consist of pairs of bilingual sentences, featuring one sentence in the source language and its corresponding counterpart in the target language. In this vein, Zhao (2022) shows that the system computes probabilities for various translation options, selecting the translation with the highest probability as the final result. Translation quality, as stated by Sharma *et al.* (2023), depends on the availability and quality of parallel data, with multiple possible translations assigned to words, chosen subjectively by the translator. However, SBMT's performance is limited by data quality and quantity, requiring extensive training data to struggle with lexical ambiguities.

Recently, these traditional systems (RBMT and SBMT) have gradually transitioned to the use of neural models. Specifically, the introduction of the Transformer Model has marked the beginning of a new phase in technological advancement. Pimentel and Pires (2024) highlighted that Transformer-based models trained on domain-specific corpora could achieve better performance than general-purpose systems such as Google Translate. Their study emphasized not only the effectiveness of these specialized models but also their viability, demonstrating that they could be developed using open-access tools and relatively modest computational resources.

With the advancement of deep learning and neural networks, Zhu *et al.* (2020) and Wang (2023) began exploring the application of neural networks in MT around the year 2014. The neural machine translation system (NMT) consists of two fundamental elements: an encoder network, which maps the structure of the original sentence into a vector of values, and a decoder network, which generates the translation from this vector. According to Wang *et al.* (2021), this model could be characterized as functioning analogously and comparably to the way the human brain processes information. In other words, it first comprehends the entirety of the original sentence and then generates the translation based on this understanding.

Vaswani *et al.* (2017) and Zhu *et al.* (2020) proposed a novel system known as Transformer. This innovative network comprises three main components: a self-attention layer, which scrutinizes the words in a sentence one by one, considering their context at each step; an encoder-decoder attention, which connects the input sentence with the output sentence; and a feed-forward layer which nonlinearly transforms the data. Wang *et al.* (2021) reported that this mechanism had demonstrated a significant enhancement in the quality of translations, solidifying NMT as a groundbreaking technology in this domain. In their study, Pimentel and Pires (2024) implemented this architecture to train a translation model using a legal English–French corpus. Despite the limited size of the dataset, the model achieved promising results, comparable to those obtained by larger, general-purpose systems. Therefore, this new model will be analysed in the following section.

2.2 The automatic evaluation

The assessment of the quality of translation is a critical aspect in the development and refinement of MT systems. Traditionally, this assessment has been carried out through two main approaches: human evaluation and automatic evaluation using specific metrics (Pym, 2020). In the present paper, we will focus exclusively on evaluation through automatic metrics. In its early stages, evaluation relied primarily on human judgments, where translators manually assessed the quality of the translation by considering adequacy, which measures whether the translation preserves the same meaning as the original text; and fluency, which evaluates the grammatical correctness of the translation (Mayor Martínez *et al*., 2009). However, this approach is time-consuming, making it more costly and subjective, which can result in inconsistent outcomes (e.g. Mayor Martínez *et al*., 2009; Lee *et al*., 2023). To overcome these limitations, Dorr *et al.* (2010) created automatic metrics to compare the machine-generated output to reference translations. The most common metrics include BLEU, METEOR, TER, WER, and NIST.

The Bilingual Evaluation Understudy (BLEU) is an automated metric that assesses the similarity between a machine translation output and a reference translation. It assesses both how accurately words are translated and how smoothly they are put together using word sequences of different lengths, ranging from single words to sequences of four words (Lee *et al*., 2023). BLEU is calculated based on three factors: matching word sequences, penalizing overly short translations, and adjusting for repeated words (Lee *et al*., 2023). However, it also shows some limitations. Dorr *et al.* (2010) highlighted its lack of recall that it works better with a large amount of data. Consequently, BLEU scores for individual sentences are considered unreliable.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an evaluation method tailored to overcome certain weaknesses found in BLEU (Dorr *et al*., 2010). While BLEU focuses on precision, METEOR is more geared towards recall. Actually, it considers both precision and recall, emphasising on recall to calculate the harmonic mean. Recent research has shown improved correlations with human judgments by fine-tuning these parameters for specific languages. METEOR employs various stages of word matching between the system output and reference translations. These stages include exact matching, stem matching (i.e. words with the same root are aligned), and synonym matching (e.g. Lavie & Denkowski, 2009; Dorr *et al*., 2010). The matcher aligns words between the hypothesis and reference strings incrementally through these stages, each corresponding to a specific word-mapping module within METEOR.

Moreover, Mathur *et al.* (2020) consider the Translation Edit Rate (TER) to be a successful tool to assess machine translation performance by quantifying the edits required to align the machine-generated output with the reference translation. It exclusively examines word-level correspondence, overlooking semantic similarity. Furthermore, TER neglects fluency assessment, solely focusing on word accuracy. Thus, a translation system might achieve a high score despite producing grammatically correct yet awkward translations.

In terms of the drawbacks associated with these metrics, Zhou *et al.* (2008) and Chatzikoumi (2019) identifies limitations such as the requirement for reference translations, which restricts the amount of data available for evaluation; the failure to recognize subtle nuances; challenges in interpreting scores and the inability to provide detailed insights into the specific strengths and weaknesses of machine translation. In addition, Peña Aguilar (2023) emphasized that automatic evaluation metrics often overlook important grammatical and semantic differences between languages. This is particularly relevant to assess the performance of translation systems such as Google Translate, Bing, and DeepL. Her empirical study shows that human evaluation can identify significant linguistic issues that are not captured by metric scores, such as abstract concepts or differences in how countable and uncountable nouns are expressed in Spanish and English. Hence, Chatzikoumi (2019) posits that human post-editing remains indispensable in the translation process, involving the identification of translation errors by comparing the source and target texts, the identification and correction of linguistic errors in the target language, and the review of the edited text.

In this section, we have explored the automatic evaluation of translation quality, emphasizing metrics such as BLEU, METEOR and TER. These tools have proven to be fundamental to analyse the accuracy and fluency of automatic translations in conventional contexts. Considering the specific focus of our research on colloquial language, it is essential to provide a detailed analysis of how BLEU and METEOR perform in this context. Therefore, to maintain a focused and in-depth examination, this study will concentrate on these two metrics, as they offer a comprehensive balance between precision and recall, which is crucial to assess teenage language.

Given the established importance of automatic metrics in evaluating translation quality and the distinct characteristics of adolescents' language, the following section will delve into the peculiarities of this linguistic variety and the intricacies MT systems need to overcome to produce meaningful translations.

2.3 Teenage language

Machine translation faces unique challenges when it has to cope with the language of adolescents due to its ever-changing nature and usage in specific social contexts. According to Rosyadi Za *et al.* (2023), teenagers are the most active group on social media platforms and employ a variety of language styles for communication and social interaction. These young individuals often incorporate the language they use on the internet into their daily communication (e.g. Rosyadi Za *et al.*, 2023; Silalahi & Silalahi, 2023). As mentioned by Eckert (2003), this varied language, known as slang, is characterized by innovative vocabulary that is commonly understood only by members of their community—in this case, adolescents. Examples include frequent use of abbreviations and acronyms such as *bro*, *cuz*, *bae*, and *lol* (e.g. Nuraeni & Pahamzah, 2021; Silalahi & Silalahi, 2023).

In recent decades, Palacios Martinez (2020) has highlighted the frequent use of intensifiers such as *really* and *so*. There is also a tendency to use taboo words as vocatives, including *fool*, *stupid*, *bastard* and *bitch* (Palacios Martínez, 2011).

In colloquial English, it is also typical to face with vague language, mainly characterized by expressions like *and things*, *and stuff*, *or something* found at the end of a sentence and referred to as general extenders (e.g. Cheshire, 2007; Tagliamonte & Denis, 2010). Within the category of vague language, placeholders such as *thingy*, *stuff* and *thingybob* are frequently used (Palacios Martínez, 2011), employed when the speaker does not remember the name of a thing or a person.

Moreover, among the features of teenagers' slang, there is a high prevalence of negative expressions. In fact, Palacios Martínez (2010, 2013) asserts that it is common to find expressions like *ain't*, *innit* and *dunno*, which have been integrated into youth discourse due to the diverse ethnic groups present in London. Napoletano and Canga Alonso (2023) point out how the abovementioned linguistic aspects complicate translation when using MT tools. Therefore, there is a persistent emphasis on the requirement for human translators to review and correct the machine output.

In the light of the reviewed literature, there is a lack of research in regard with how Bing Translator, DeepL and HelsinkiNLP address the translation of this type of language into Spanish and

Italian. Hence, this paper intends to answer the following questions: RQ1: Which of the three machine translation tools produces the most accurate translations? RQ2: Which metric (i.e. BLEU and METEOR) provides the most reliable results? RQ3: How do specific features of youth language affect the ability of MT tools to generate precise and comprehensible translations?

3. Method

3.1 Instruments and procedure

In this paper, the authors adopt a corpus-driven approach based on inductive analysis. This method allows linguistic patterns to arise directly from the data, rather than being guided by predefined theoretical frameworks (Tognini-Bonelli, 2001). Thus, the study considers the Linguistic Innovators Corpus (LIC), compiled by Lancaster University, which gathers data from 2004 to 2007 (Torgersen *et al*., 2011). It is noteworthy that this corpus contains spoken language data, consisting of interviews with young people aged 16 to 19 from Hackney, located in London. Due to the extensive corpus, it was necessary to select fragments which contained idioms, phrasal verbs, negative expressions and intensifiers. For this reason, the authors selected only 1000 characters with a twofold aim: (i) to maintain coherence in the dialogue when introducing the fragment into the online tool, and (ii) to avoid coherence and cohesive translation problems derived from the character limitations of the tools used. The selection criteria focus mainly on adolescents' expressions, such as idioms, vague language, and negative expressions. Although the eight examples analysed happen to include these features, they were not deliberately chosen for that reason. Instead, the selection reflects the natural flow of spontaneous language. It should also be noted that before translating the English fragments, the corpus was edited and non-content words such as hashtags, numbers and registration marks were deleted, to improve the quality of the data (Rahm & Do, 2000). As for online translators, we will use exclusively Transformers systems that allow free access to everyone. This decision is based on the proven effectiveness of Transformer-based models, which have demonstrated high performance in handling complex linguistic structures and providing high-quality translations compared to earlier models (Vaswani *et al*., 2017). Additionally, their free accessibility ensures the relevance and applicability of our study to a broad audience. Therefore, DeepL, Bing Translator and HelsinkiNLP will be used.

DeepL is one of the most widely used translators today (Birdsell, 2022); it features a neural network-based translation system developed by the German company DeepL GmbH (Tavosanis, 2019), which was launched in 2017 (Roiss & Zimmermann González, 2020). This tool presents a 1500-character limit. On the other hand, Bing Translator, developed by Microsoft, is considered the most effective translator when translating cultural equivalents into English (e.g. Jibreel, 2023; He *et al*., 2024). This tool enables fewer characters, limited to 1000.

However, the most innovative system is HelsinkiNLP, undertaken by the University of Helsinki. Researchers, such as Smirnov *et al*. (2022), have focused on training various bilingual language models for translating multiple languages to and from English. Unlike the other two translators, it facilitates the inclusion of longer texts by supporting up to 6000 characters.

Overall, the selection of these online tools provides a diverse and meaningful representation of the current state of MT. It encompasses widely used commercial translators as well as approaches based on academic research. This will enable a comprehensive evaluation of translation quality and the effectiveness of different systems using standard metrics.

3.2 Analysis

The selected fragments had been translated from English into Spanish and Italian. The choice of these languages stems from their status as Romance languages, sharing linguistic roots and structural similarities with English (e.g., subject-verb-object word order and similar parts of speech), even though there are differences in grammar and vocabulary. Studying how MT manages these distinctions can provide valuable insights into the specific challenges confronting this technology. Additionally, Spanish and Italian are of growing interest in the research and development of MT technologies, given their significance on both European and global scales (Russo *et al*., 2012).

Besides, the chosen excerpts differ in length, ranging from 3 to 55 words, and have been taken from a 1000-character conversation involving three speakers (Sue, Ahmed and Tina). They were copied and pasted into each online translator (DeepL, Bing Translator and HelsinkiNLP). Then, the output was evaluated according to the results provided by BLEU and METEOR.

4. Results

As for our first research question (RQ1: Which of the three machine translation tools produces the most accurate translations?). The findings displayed in Table 1 show a high degree of similarity among the outputs of the three tools.

Table 1: Translation outputs into Spanish and Italian

Source Text	Bing Translator	DeepL	HelsinkiNLP
Sue: So, what are you doing in the college at the moment? **Ahmed: I'm coming to see Mark and Tina, innit?** **Mark: He's wandering.**	Sue: Entonces, ¿qué estás haciendo en la universidad en este momento? Ahmed: Voy a ver a Mark y a Tina, ¿no? Mark: Está deambulando.	Sue: ¿Qué estás haciendo en la universidad en este momento? Ahmed: Vengo a ver a Mark y Tina, ¿no? Mark: Está deambulando.	Sue: Entonces, ¿qué estás haciendo en la universidad en este momento? Ahmed: Voy a ver a Mark y Tina, ¿no? Mark: Está vagando.
	Sue: Allora, cosa stai facendo al college in questo momento? Ahmed: Vengo a trovare Mark e Tina, no? Mark: Sta vagando.	Sue: Allora, cosa fai al college al momento? Ahmed: Sono venuto a trovare Mark e Tina, no? Mark: Sta vagando.	Sue: Allora, cosa ci fai al college al momento? Ahmed: Vengo a trovare Mark e Tina, no? Mark: Sta vagando.
Ahmed: Know erm fucking erm job centre, but i'm not doing it. They're sitting there and talking about getting to know each other.	Ahmed: Sé erm maldito erm centro de trabajo, pero no lo voy a hacer. Están sentados allí y hablan de conocerse. Estoy como "a la mierda con eso", no voy a hacer eso. Por ejemplo, ¿por	Conocer erm maldito erm centro de trabajo, pero yo no lo estoy haciendo. Están sentados allí y hablando de conocerse. Estoy como "a la mierda con eso", no voy a hacerlo. ¿Por qué quiero	Ahmed: Conocer el puto centro de trabajo, pero no lo estoy haciendo. Están sentados allí y hablando de llegar a conocerse. Soy como "joder eso" que no estoy haciendo eso.

I'm like "fuck that" I ain't doing that. Like why do I wanna know these people for it's not as if I'm gonna chat to them outside college. Do you get me?	qué quiero conocer a estas personas, porque no es como si fuera a charlar con ellas fuera de la universidad? ¿Me entiendes?	conocer a esta gente? No es como si fuera a charlar con ellos fuera de la universidad. ¿Me entiendes?	Como por qué quiero conocer a esta gente porque no es como si fuera a hablar con ellos fuera de la universidad. ¿Me entiendes?
	Ahmed: Conosco ehm cazzo di centro per l'impiego, ma non lo sto facendo. Sono seduti lì e parlano di conoscersi. Sono tipo "fanculo", non lo farò. Ad esempio, perché voglio conoscere queste persone perché non è come se stessi andando a parlare con loro fuori dal college. Mi capisci	Ahmed: Conosco il fottuto centro per l'impiego, ma non lo faccio. Sono seduti lì e parlano di conoscersi. Io dico: "Fanculo", non lo faccio. Perché voglio conoscere queste persone, non è che ci parlo fuori dall'università. Mi capisci?	Ahmed: Conosco erm cazzo centro di lavoro erm, ma non lo sto facendo. Sono seduti lì e parlare di conoscere l'un l'altro. Sono come "cazzo che" non lo sto facendo. Come perché voglio conoscere queste persone perché non è come se ho intenzione di chiacchierare con loro al di fuori del college. Mi capisci?
Sue: So, you like one of the guys in the painting and decorating, do you? **Tina: He's butters.**	Sue: Entonces, ¿te gusta uno de los chicos de la pintura y la decoración, verdad? Tina: Es mantequilla.	Sue: Así que te gusta uno de los chicos de la pintura y la decoración, ¿verdad? Tina: Es Butters	Entonces, te gusta uno de los chicos de la pintura y la decoración, ¿verdad?
	Sue: Allora, ti piace uno dei ragazzi che dipinge e decora, vero? Tina: È butters.	Sue: Allora, ti piace uno dei ragazzi che dipingono e decorano, vero? Tina: È Butters	Allora, ti piace uno dei ragazzi del dipinto e della decorazione, vero?
Ahmed: But what do you call it? He's got a bit of an attitude problem when he's in front of his mates. When he's on his own he just went quiet like a little puppy	Ahmed: ¿Pero ¿cómo lo llamas? Tiene un pequeño problema de actitud cuando está frente a sus compañeros. Cuando está solo, se queda callado como un cachorro	Ahmed: ¿Pero cómo lo llamas? Tiene un pequeño problema de actitud cuando está delante de sus compañeros. Cuando está solo, se queda callado como un cachorrito.	Ahmed: ¿Pero cómo lo llamas? Tiene un pequeño problema de actitud cuando está delante de sus compañeros. Cuando está solo se quedó callado como un perrito
	Ahmed: Ma come lo chiami? Ha un po' di problemi di atteggiamento quando è di fronte ai suoi compagni. Quando è da solo si calma come un cucciolo	Ahmed: Ma come lo chiami? Ha un po' di problemi di atteggiamento quando è di fronte ai suoi compagni. Quando è da solo, si calma come un cucciolo.	Ahmed: Ma come lo chiami? Ha un piccolo problema di atteggiamento quando è di fronte ai suoi compagni. Quando è da solo, è andato tranquillo come un cucciolo
Ahmed: So, why should I waste my breath on him? Well, like it's quite obvious he doesn't want to me and Mark	Ahmed: Entonces, ¿por qué debería desperdiciar mi aliento en él? Bueno, es bastante obvio que no nos quiere a mí y a Mark por la forma en que se presenta.	Ahmed: Entonces, ¿por qué debería gastar mi aliento en él? Bueno, como es bastante obvio que no quiere que yo y Mark debido a la forma en que viene a través.	Ahmed: Entonces, ¿por qué debería desperdiciar mi aliento en él? Bueno, como es bastante obvio que él no quiere para mí y Mark debido a la forma en que se encuentra.

because of the way he comes across.	Ahmed: Allora, perché dovrei sprecare il mio fiato con lui? Beh, come se fosse abbastanza ovvio che non vuole me e Mark a causa del modo in cui si presenta.	Allora, perché dovrei sprecare il mio fiato con lui? Beh, è ovvio che non vuole me e Mark per il modo in cui si presenta.	Ahmed: Allora, perché dovrei sprecare il fiato su di lui? Beh, come se fosse abbastanza ovvio che lui non vuole me e Mark a causa del modo in cui si imbatte.
Tina: Maybe he doesn't like hanging about with us cos we're too in your face.	Tina: A lo mejor no le gusta andar con nosotros porque estamos demasiado en tu cara.	Tal vez no le gusta salir con nosotras porque somos demasiado inofensivas.	Tal vez no le gusta estar con nosotros porque estamos en tu cara.
	Tina: Forse non gli piace stare con noi perché siamo troppo in faccia.	Tina: Forse non gli piace stare con noi perché siamo troppo presenti.	Forse non gli piace stare con noi perche' siamo troppo in faccia a te.
Ahmed: I don't fancy him. He's a minger anyway butters. They two are quite clingy over him. I'm like "No, he ain't coming". "Oh, what a shame" and then will go. Make out that he's doing something else now because he ain't fucking go.	Ahmed: No me gusta. De todos modos, es un minger. Los dos son bastante pegajosos a él. Le digo: "No, no va a venir". "Oh, qué vergüenza" y luego se irá. Da cuenta de que ahora está haciendo otra cosa porque no se va a ir de puta vez.	No me gusta. De todos modos, es un idiota. Ellos dos son muy pegajosos con él. Yo digo: "No, no va a venir". "Oh, qué pena" y luego se irá. Haz como que está haciendo otra cosa ahora porque no va a ir.	Ahmed: No me gusta. De todos modos, es un picadero de mantequillas. Los dos son bastante pegajosos con él. Yo digo: "No, no va a venir". "Oh, qué pena" y luego se irá. Haz que esté haciendo otra cosa ahora porque no se va.
	Ahmed: Non mi piace. È comunque un minger butters. Loro due sono piuttosto appiccicosi su di lui. Io rispondo: "No, non viene". "Oh, che peccato" e poi se ne andrà. Fai finta che ora stia facendo qualcos'altro perché non se ne andrà, cazzo.	Ahmed: non mi piace. E' un mingherlino, in ogni caso. Loro due sono piuttosto appiccicati a lui. Io dico: "No, non viene". "Oh, che peccato" e poi se ne va. E poi dice che sta facendo qualcos'altro perché non se ne va, cazzo.	Ahmed: Non mi piace lui. E' comunque un tritacarte. Sono due piuttosto appiccicosi su di lui. Sono come "No, lui non viene." "Oh, che peccato" e poi andrà. Fai in modo che stia facendo qualcos'altro ora perché non se ne va, cazzo.
Tina: Yesterday he was acting like nothing happened. I don't know like I never existed little fucker. So no believe me today, we will have words.	Tina: Ayer estaba actuando como si nada hubiera pasado. No sé, como si nunca hubiera existido, pequeño hijo de puta. Así que no, créanme hoy, tendremos palabras.	Tina: Ayer actuaba como si nada hubiera pasado. No sé, como si nunca hubiera existido pequeño cabrón. Así que no me creas hoy tendremos palabras.	Tina: Ayer estaba actuando como si nada hubiera pasado. No sé, como si nunca hubiera existido, pequeño hijo de puta. Así que no me creas hoy, vamos a tener palabras.
	Tina: Ieri si comportava come se nulla fosse. Non lo so, come se non fossi mai esistito, piccolo stronzo. Quindi non credetemi oggi, avremo parole.	Ieri si comportava come se non fosse successo nulla. Non lo so, come se non fossi mai esistito, piccolo stronzo. Quindi no, credimi, oggi parleremo.	Tina: Ieri si stava comportando come se non fosse successo niente. Non so come se non fossi mai esistito piccolo stronzetto. Quindi non credetemi oggi, avremo parole.

Source: Authors (2025)

However, there are some informal expressions that are not translated correctly or are completely omitted by the system. It is the case of the phrase ‘He’s butters’, which was translated literally by Bing Translator into Spanish, left untranslated by DeepL, and omitted by HelsinkiNLP. In contrast, in Italian, both Bing Translator and DeepL failed to translate the expression, while HelsinkiNLP omitted it altogether. In this respect, Vilar *et al.* (2006) suggested that omissions in the target text commonly occur with nouns or verbs. Our findings seem to concur with this observation. Another noteworthy aspect is that in some fragments, DeepL and HelsinkiNLP delete the speaker’s name in their outputs, whereas Bing Translator demonstrates greater precision in this regard. According to Goto and Tanaka (2017), when MT is dealing with longer sentences, it tends to remove some information in the translated text.

In addition, Bing Translator and HelsinkiNLP render a literal translation for the idiomatic expression ‘we’re too in your face’ in Spanish as ‘estamos en tu cara’ and in Italian as ‘troppo in faccia’, whereas DeepL provides an incorrect translation, interpreting the expression as referring to someone being unable to cause any harm to someone else. As stated by Sharou and Specia (2022), this discrepancy underscores a fundamental characteristic of MT tools, which often generate translations that are inaccurate, incomprehensible, or different from the original meaning. Specifically, the Italian translation by DeepL maintains a formal tone, while in Spanish, the meaning is misconstrued, leading to a significant alteration in semantic interpretation. Similarly, the idiomatic expression ‘we will have words’, which is literally translated into Spanish as ‘tendremos palabras’ and into Italian as ‘avremo parole’, fails to capture the idiomatic meaning and results in a loss of the original one.

Lastly, it is also worth noting how these tools handle swear words. The results show that the phrase ‘he ain’t fucking go’ is translated into Spanish only by Bing Translator, while it is omitted by DeepL and HelsinkiNLP. However, all three systems achieve an accurate translation in Italian (‘cazzo’), highlighting variability in performance across different languages.

As shown in Table 1, it is difficult to establish a precise percentage of fully accurate or inaccurate translations, as each fragment tends to contain both correct and incorrect elements. In other words, while no fragment is entirely mistranslated, none is entirely accurate either. Typically, a portion of the text is translated appropriately, whereas certain expressions—particularly idiomatic or informal ones—continue to pose significant challenges for MT systems. One illustrative example is the fragment Maybe he doesn’t like hanging about with us ‘cos we’re too in your face, which was translated in Spanish as “A lo mejor no le gusta andar con nosotros porque estamos demasiado en tu cara”. While the initial part of the sentence is accurately rendered, the expression we’re too in your face is not appropriately translated, indicating a failure to convey its idiomatic meaning.

As for RQ2 (Which metric (i.e. BLEU and METEOR) provides the most reliable results?) and RQ3 (How do specific features of youth language affect the ability of MT tools to generate precise and comprehensible translations?), Table 2 visually shows the reliability of BLEU and METEOR, and it also serves to reflect how teenage language affects the accuracy of online translation systems. It should be noted that we take Hadla *et al.* (2015) scale which considers that the closer the result is to 1, the more the generated translation approaches a correct translation and resembles human references.

The results generated by Bing Translator indicate that for English to Spanish translations, the highest score is 0.3375 with METEOR, suggesting good semantic adequacy and fluency. In general terms, the system successfully produces an output with correct grammar and accurately transfers the meaning of the original text into the target text, but it sometimes translates colloquial sentences in a formal way as it was mentioned in the examples provided to answer RQ1.

Table 2: Metrics' evaluation

Metrics	**Bing Translator EN-SP**	**Bing Translator EN-IT**	**DeepL EN-SP**	**DeepL EN-IT**	**HelsinkiNLP EN-SP**	**HelsinkNLP EN-IT**
BLEU	0.0732	0.0738	0.0714	0.0728	0.0729	0.0728
	0.0228	0.0202	0.0225	0.0220	0.0220	0.0238
	0.0744	0.0754	0.0839	0.0844	0.0174	0.0167
	0.0121	0.0119	0.0269	0.0273	0.0269	0.0271
	0.0338	0.0344	0.0332	0.0182	0.0332	0.0339
	0.0341	0.0352	0.0118	0.0319	0.0123	0.0154
	0.0566	0.0264	0.0125	0.0549	0.0546	0.0383
	0.0347	0.0346	0.0371	0.0223	0.0383	0.0313
METEOR	0.3375	0.3647	0.3257	0.3527	0.3384	0.3530
	0.1221	0.0986	0.1238	0.1506	0.1019	0.1830
	0.3144	0.3197	0.3438	0.3602	0.0797	0.0806
	0.0602	0.0801	0.1159	0.1563	0.1063	0.1566
	0.1541	0.1759	0.1440	0.0737	0.1440	0.1666
	0.1352	0.1696	0.0271	0.1381	0.0270	0.0534
	0.1950	0.2409	0.1692	0.1879	0.1937	0.1757
	0.1748	0.1605	0.1620	0.1215	0.1890	0.1441

Source: Authors (2025)

As for the English to Italian translations, the highest METEOR score observed is 0.3647, showing high accuracy in meaning. Nevertheless, BLEU gives a score of 0.0119, indicating potential difficulties with text coherence. This result seems to show a failure in the system to accurately translate a multi word-expression. Thus, MT systems generally prioritize simpler sentences that are easier to translate sequentially (Volk, 1998).

As shown in Table 2, the analysis presented by DeepL reveals that the English into Spanish translations achieved a minimum BLEU score of 0.0118. This score implies a clear deficiency to attain exact correspondences with human references, particularly in terms of grammar and sentence structure nuances characteristic of informal language. As Lotz and Van Rensburg (2016) suggest, the MT tools may not have been sufficiently trained to handle these types of texts optimally.

Additionally, for the English into Italian outputs, the lowest observed METEOR score is 0.0271. Similar to BLEU, this metric indicates challenges in accurately capturing idiomatic expressions, such as 'we're too in your face', within the Spanish context. However, the same phrase achieved a METEOR score of 0.1381 in Italian ('siamo troppo presenti'), indicating a relatively closer structural similarity between the translated output and the human reference text, despite the overall low score.

Evaluating HelsinkiNLP, we can contend that translations from English into both Spanish and Italian struggle with contextual understanding, as indicated by low BLEU scores of 0.0174 and 0.0167, respectively. Specific failures include omitting phrases like 'He's butters', which suggests a lack of contextual understanding in both language outputs. Such errors highlight the problems these MT systems encounter with contextual references and the need for further development to address this translation challenge (Lotz & Van Rensburg, 2016).

Regarding METEOR scores, translations from English into Italian preserve the meaning and style of the source text. However, one of the highest yet still low scores was 0.1757. The system consistently fails to translate phrases such as 'he's a minger anyway butters', resulting in mistranslations. For instance, the tool translated this phrase to 'tritacarte', which means 'paper shredder'. This output implies that the system selects a wrong word that is not related to the context or the preceding word which concurs with Costa *et al.* (2015). Likewise, there is a mistranslation in Spanish, where HelsinkiNLP translated it as 'picadero de mantequilla'. However, the correct translation into Spanish was provided only by DeepL: 'es un idiota'.

These findings portray a consistent trend wherein METEOR scores are higher than BLEU scores; thus, METEOR seems to give a better assessment of translation quality in terms of how well the meaning and fluency of the text are preserved, whereas BLEU, which focuses on exact matches, may not fully capture these aspects.

5. Discussion

Regarding RQ1 (Which of the three machine translation tools produces the most accurate translations?), our findings indicate the most accurate tool is DeepL. Since it generates translations that are more coherent and faithful to the meaning of the source text. However, DeepL encounters difficulties to translate phrasal verbs in both Spanish and Italian. For instance, the expression 'the way he comes across' is translated into Spanish as 'en que viene a través'. This outcome implies that, as suggested by Thiruumeni *et al.* (2011), the online tool does not recognize multi-word expressions; therefore, it translates the phrasal verb by only looking at the first part of the phrase and treats the preposition like an adverb. Consequently, our findings also concur with Thiruumeni *et al.* (2011) who pointed out that a significant issue in current machine translation systems concerning phrasal verbs is that they often translate only the main verb instead of treating the entire expression as a single unit.

Multi-word expressions such as idioms also present difficulties to be translated by means of a machine translation system. In fact, none of the idioms were translated correctly by any of the three online tools. For instance, the expression 'he just went quiet like a little puppy' was translated literally and formally. The expression was translated into Spanish as 'se queda callado como un cachorrito/perrito' and into Italian as 'calmo come un cucciolo'. These data concur with Gaspari and Zachetta (2011) and He *et al.* (2024) since they are able to grasp the concept, but struggle to fully convey the mood and cultural differences of the source text. Hence, these tools seem to find problems to interpret certain expressions that would be readily understood by human readers, requiring human post-editing to capture cultural differences between languages.

As for RQ2 (Which metric provides more reliable results?), our findings suggest that METEOR offers more reliable results compared to BLEU with a score of 0.0839. METEOR incorporates several features not included in BLEU, such as "[...] stemming, synonym matching, and standard exact word matching" (Hadla *et al.*, 2015, p. 216). These elements are crucial for the success of a metric which needs to compare the quality of a translation from English into Spanish and Italian: the two languages present a richer morphology, and morphemes convey more information, and different forms of the same word may not be as freely interchangeable as in English (Agarwal & Lavie, 2008).

On the contrary, BLEU only checks for exact word matches and does not consider stems or synonyms, making it harder to evaluate translations in languages with varied morphological expressions (Lee *et al.*, 2023). Thus, as shown in previous studies (e.g. Lavie & Denkowski, 2009; Dorr *et al.*, 2010), our findings seem to prove that METEOR attempts to capture semantic similarity by considering synonyms and inflected forms, which aligns more closely with human evaluations of translation quality. However, it is important to note that while METEOR addresses some of BLEU's limitations (such as its focus on exact word matches and disregard for synonyms), it is not without its own shortcomings, especially related to the context of the phrase or speech.

With regard to the last research question (How do specific features of youth language affect the ability of MT tools to generate precise and comprehensible translations?), our data show that it is hard for MT tools to handle teenage slang effectively.

In view of our findings, Bing Translator, DeepL, and HelsinkiNLP encounter difficulties to translate idiomatic expressions, as previously noted. Our data confirms Alawi and Abdulhaq (2017) assertion that these systems tend to produce a literal output with a meaning that is irrelevant or does not fit the context in which it is used. Additionally, none of the MT tools cannot provide an accurate translation for slang, or word used in regional dialects since these translation systems cannot identify the SL word within their databases. As a result, they leave the original phrase untranslated, thus, this approach keeps the original meaning and avoids misunderstandings that might come from incorrect translations (e.g. Costa *et al.*, 2015; Jufriadi *et al.*, 2022).

Alternatively, progress in neural systems is evident in their ability to successfully translate negative and taboo expressions such as 'ain't', 'innit' and 'fuck that' (translated as 'a la mierda eso' into Spanish and 'fanculo' into Italian). However, these expressions have only been accurately translated by Bing Translator, DeepL and HelsinkiNLP have still encountered difficulties in translating 'fuck that,' often omitting the phrase altogether. To conclude, our data seem to confirm Orrego-Carmona's (2022) assertion that machine translation systems have accomplished to produce fluent translations. Nonetheless, achieving accuracy remains a persistent challenge, often requiring post-editing to meet human-level coherence and quality standards. This need becomes especially apparent with the translation of idiomatic expressions, as NMT systems frequently generate literal, word-for-word translations that can result in meaningless or incorrect output (Baziotis *et al.*, 2023). Moreover, these systems often fail to capture cultural nuances and context-dependent meanings, which can lead to translations that do not convey the intended message (Peña Aguilar, 2023). As a result, human post-editing remains essential to address these shortcomings and to ensure that translations are both accurate and contextually appropriate.

6. Conclusion

Our study investigated the performance of three prominent machine translation tools—Bing Translate, DeepL, and HelsinkiNLP—across various metrics and translation challenges. Our findings highlight that *DeepL* emerges as the most accurate tool. However, it faces significant difficulties to translate phrasal verbs in Spanish and Italian, often failing to interpret them as cohesive units due to its handling of prepositions and adverbs. Additionally, our research reveals that all three systems produce comprehensible translations but struggle with phrasal verbs and idioms. These online tools maintain sentence structure and syntax but encounter difficulties with more nuanced aspects of language and culture.

Furthermore, the present study compared the effectiveness of two metrics, BLEU and METEOR. As it can be noted, METEOR proves to be more effective in capturing semantic and stylistic differences, whereas BLEU focuses primarily on text structure. This distinction is critical for languages with complex morphologies like Spanish and Italian, where word forms convey nuanced meanings that exact word matching struggles to capture. However, both metrics have their limitations. BLEU can lead to biased evaluations favouring literal translations, while METEOR, despite considering synonyms and inflected forms, can still present problems in regard with contextual appropriateness. The disparity in results between BLEU and METEOR underscores the importance of using multiple evaluation metrics for a comprehensive understanding of automated translation quality. What is more, the inability of these tools to handle slang and colloquialisms has also been highlighted, underscoring the current limitations in capturing the nuances of informal language.

One limitation of the present study is that the analysis has been carried out taking into account free access tools with a limited number of characters, these constrain limits the corpus size, as only a limited number of fragments were analysed due to this limitation. we do not certainly know whether these limits are also repeated in premium versions, or whether they offer an improved service. Further research is needed to validate or refute this hypothesis. Another limitation concerns the metrics; thus, only the most used metrics, such as BLEU and METEOR, were considered. Although these are standard in assessing machine translations, they have certain shortcomings. For instance, BLEU depends on finding n-gram matches between the machine translation and multiple human references (Lee *et al.*, 2023), which can lead to literal translations. METEOR, on the contrary, attempts to address some of these issues but it can still present problems relating to context. Relying exclusively on these metrics ignores others that could provide a more reliable assessment of translation quality. As an example, metrics like TER assess the number of edits needed for a machine translation to be comprehensible (Mathur *et al.*, 2020). Additionally, newer metrics like BERTScore, which use pre-trained language models (Saadany & Orasan, 2021) could provide a more precise output. Therefore, future research could be enhanced by including a greater variety of metrics for a more accurate evaluation of translation quality.

To conclude, our findings purport that MT achieves fluency in many contexts, but human intervention is necessary to ensure comprehensible and culturally accurate translations. Future advancements should prioritize improving the ability of these MT tools to interpret idiomatic expressions and colloquialisms, thereby moving towards higher standards of quality in machine translation. It is worth noting, however, that the analysed corpus includes expressions from 2004 to 2007, and language use among teenagers may have evolved since then. As, to our knowledge, no up-to-date corpus of contemporary teenage speech in English appears to be publicly available, future research would benefit from exploring digital platforms and social media, where current youth language is actively used.

References

Agarwal, A., & Lavie, A. (2008). METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. *Proceedings of the Third ACL Workshop on Statistical Machine Translation.* Association for Computational Linguistics.

Alawi, N., & Abdulhaq, S. (2017). Machine Translation: The Cultural and Idiomatic Challenge. *Journal of Al-Azhar University – Gaza (Humanities)*, *19*(2), 1–28.

Banitz, B. (2020). Machine translation: A critical look at the performance of rule-based and statistical machine translation. *Cadernos de Tradução*, *40*(1), 54–71. https://doi.org/10.5007/2175-7968.2020v40n1p54

Baziotis, C., Mathur, P., & Hasler, E. (2023). Automatic Evaluation and Analysis of Idioms in Neural Machine Translation. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 3649–3661). Association for Computational Linguistics.

Birdsell, B. J. (2022). Student writings with DeepL: Teacher evaluations and implications for teaching. In P. Ferguson & R. Derrah (Eds.), *JALT2021: Reflections and New Perspectives* (pp. 117-125). JALT. https://doi.org/10.37546/JALTPCP2021-14

Chatzikoumi, E. (2019). How to evaluate Machine Translation: A review of Automated and Human Metrics. *Natural Language Engineering*, *26*(2), 137–161. https://doi.org/10.1017/S1351324919000469

Cheshire, J. (2007). Discourse Variation, Grammaticalisation and "Stuff like That". *Journal of Sociolinguistics*, *11*(2), 155–193. https://doi.org/10.1111/j.1467-9841.2007.00317.x

Costa, Â., Ling, W., Luís, T., Correia, R., & Coheur, L. (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, *29*(2), 127–161. http://dx.doi.org/10.1007/s10590-015-9169-0

Das, A. K. (2018). Translation and Artificial Intelligence: Where are we heading? *International Journal of Translation*, *30*(1), 1–26.

Dorr, B., Snover, M., & Madnani, N. (2010). Chapter 5.1 Introduction. In B. Dorr (Ed.), *Part 5: Machine Translation Evaluation* (pp. 802–806). DARPA GALE Program Report.

Duan, G., Yang, H., Qin, K., & Huang, T. (2021). Improving Neural Machine Translation Model with Deep Encoding Information. *Cognitive Computation*, *13*, 972–980. https://doi.org/10.1007/s12559-021-09860-7

Eckert, P. (2003). Language and adolescent peer groups. *Journal of Language and Social Psychology, 22*(1), 112-118. https://doi.org/10.1177/0261927X02250063

Gaspari, F., & Zacchetta, E. (2011). Scrittura controllata per la traduzione automatica. In G. Bersani Berselli (Ed.), *Usare la Traduzione Automatica* (pp. 63-79). Clueb.

Goto, I., & Tanaka, H. (2017). Detecting Untranslated Content for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation.* Association for Computational Linguistics.

Hadla, L. S., Hailat, T. M., & Al-Kabi, M. N. (2015). Comparative Study Between METEOR and BLEU Methods of MT: Arabic into English Translation as a Case Study. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *6*(11), 215–223. https://dx.doi.org/10.14569/IJACSA.2015.061128

He, L., Ghassemiazghandi, M., & Subramaniam, I. (2024). Comparative assessment of Bing Translator and Youdao Machine Translation Systems in English-to-Chinese literary text translation. *Forum for Linguistic Studies*. *6*(2), 1–18. https://doi.org/10.59400/fls.v6i2.1189

Hutchins, J., & Somers, H. (1992). *An Introduction to Machine Translation*. Academic Press Limited.

Jibreel, I. (2023). Online Machine Translation Efficiency in Translating Fixed Expressions Between English and Arabic (Proverbs as a Case-in-Point). *Theory and Practice in Language Studies*, *13*(5), 1148–1158. https://doi.org/10.17507/tpls.1305.07

Jufriadi, J., Asokawati, A., & Thayyib, M. (2022). The Error Analysis of Google Translate and Bing Translator in Translating Indonesian Folklore. *FOSTER: Journal of English Language Teaching*, *3*(2), 69–79. https://doi.org/10.24256/foster-jelt.v3i2.89

Lavie, A., & Denkowski, M. (2009). The METEOR metric for automatic evaluation of Machine Translation. *Machine Translation*, 23, 105–115. https://doi.org/10.1007/s10590-009-9059-4

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. *Mathematics,* *11*(4), 1–22. https://doi.org/10.3390/math11041006

Lotz, S., & Van Rensburg, A. (2016). Omission and other sins: Tracking the quality of online machine translation output over four years. *Stellenbosch Papers in Linguistics*, *46*, 77–97. https://doi.org/10.5774/46-0-223

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the Evaluation of Automatic Mahine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Mayor Martínez, A., Alegría Loinaz, I., Díaz de Ilarraza Sánchez, A., Labaka Intxauspe, G.,Lersundi Ayestaran, M., & Sarasola Gabiola, K. (2009). Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve. *Procesamiento del Lenguaje Natural*, *43*, 197–205.

Moneus, A. M., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon*, *10*(6), 1–14. https://doi.org/10.1016/j.heliyon.2024.e28106

Napoletano, M. C., & Canga Alonso, A. (2023). The Translation of Adolescence Language by means of Apertium, Systran and Google Translate. *Revista Electrónica de Lingüística Aplicada*, *22*(1), 148–163. http://dx.doi.org/10.58859/rael.v23i1.585

Nicholas, G., & Bhatia, A. (2023). *Lost in translation: Large language models in non-English content analysis*. Center for Democracy & Technology. https://doi.org/10.48550/arXiv.2306.07377

Nuraeni, F. W., & Pahamzah, J. (2021). An Analysis of Slang Language used in Teenager Interaction. *Litera*, *20*, 313–322. http://dx.doi.org/10.21831/ltr.v20i2.37058

Orrego-Carmona, D. (2022). Machine translation in everyone's hands – Adoption and changes among general users of MT. R*evista Tradumàtica. Tecnologies de la Traducció*, *20*, 322–339. https://doi.org/10.5565/rev/tradumatica.324

Palacios Martínez, I. M. (2011). The language of British teenagers: A preliminary study of its main grammatical features. *Atlantis*, *33*(1), 105–126.

Palacios Martínez, I. M. (2013). Non-standard negation in Modern English: A corpus-based study of four salient features. *ES Review. Spanish Journal of English Studies*, 34, 211–226.

Palacios Martínez, I. M. (2020). Taboo vocatives in the language of London teenagers. *Pragmatics*, *31*(2), 250–277 https://doi.org/10.1075/prag.19028.pal

Palacios Martínez I. M. (2021). Recent changes in London English: An overview of the main lexical, grammar and discourse features of Multicultural London English (MLE). *Complutense Journal of English Studies*, *29*, 1–20. https://doi.org/10.5209/cjes.77504

Peña Aguilar, A. (2023). Challenging machine translation engines: Some Spanish-English linguistic problems put to the test. *Cadernos de Tradução*, *43*(1), 1–26. https://doi.org/10.5007/2175-7968.2023.e85397

Pimentel, C. H. M., & Pires, T. B. (2024). Treinamento e análise de um modelo de tradução automática baseado em Transformer. *Texto Livre: inguagem e Tecnologia, 17*, 1–15. https://doi.org/10.1590/1983-3652.2024.49118

Pym, A. (2020). Quality. In M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology* (pp. 437–449). Routledge.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4), 3–13.

Roiss, S., & Zimmermann González, P. (2020). DeepL y su potencial para el desarrollo de la capacidad de análisis crítico en la clase de traducción inversa. *Hermēneus. Revista de Traducción e Interpretación, 22*, 363–382. https://doi.org/10.24197/her.22.2020.363-382

Rosyadi Za, D., Purnamawati, N., Galuh Dwi Ajeng, A. M., & Hejash, M. (2023). Slang as a Medium of Communication for Adolescents in Social Interaction between Others. *JETA. Journal of English Teaching and Applied Linguistic*, *4*(1), 1–14. https://doi.org/10.52217/jeta.v4i1.1141

Russo, L., Loáiciga, S., & Gulati, A. (2012). Improving Machine Translationof null subjects in Italian and Spanish. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 81–89). Association for Computational Linguistics.

Saadany, H., & Orasan, C. (2021). BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text. *Proceedings of the Translation and Interpreting Technology Online Conference* (pp. 48–56). INCOMA Ltd.

Sharma, S., Diwakar, M., Singh, P., Singh, V., Kadry, S., & Kim, J. (2023). Machine translation systems based on classical-statistical-deep learning approaches. *Electronics*, *12*(7), 1–29. https://doi.org/10.3390/electronics12071716

Sharou, K. A., & Specia, L. (2022). A taxonomy and study of critical errors in Machine Translation. In H. Moniz, L. Macken, A. Rufener, L. Barrault, M. R. Costa-Jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy & M. Fonteyne (Eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Silalahi, E., & Silalahi, N. (2023). Linguistics Realization Analysis on Slang Word; Social Media Whatsapp. *JETAL. Journal of English Teaching & Applied Linguistic*, *5*, 8–13. http://dx.doi.org/10.36655/jetal.v5i1.1120

Smirnov, A. V., Teslya, N., Shilov, N., Frank, D., Minina, E., & Kovacs, M. (2022). Comparative Analysis of Neural Translation Models based on Transformers Architecture. *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022)* (pp. 586–593). https://doi.org/10.5220/0011083600003179

Son, J., & Kim, B-Y. (2023). Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information* *14*(10), 1–18. https://doi.org/10.3390/info14100574

Song, R. (2022). Analysis on the Recent Trends in Machine Translation. *Highlights in Science, Engineering and Technology*, *16*, 40–47. https://doi.org/10.54097/hset.v16i.2228

Tagliamonte, S. A., & Denis, D. (2010). The Stuff of Change: General Extenders in Toronto, Canada. *Journal of English Linguistics, 38*(4), 335–368. https://doi.org/10.1177/0075424210367484

Tavosanis, M. (2019). Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano. In R. Bernardi, R. Navigli & G. Semeraro (Eds.), *CLiC-it 2019. Proceedings of the Sixth Italian Conference on Computational Linguistics.* CEUR.

Thiruumeni, P. G., Anand, K., Dhanalakshmi, V., & Soman, K. P. (2011). An approach to handle idioms and phrasal verbs in English-Tamil Machine Translation system. *International Journal of Computer Applications, 26,* 36–41. https://doi.org/10.5120/3139-4328

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Torgersen, E. N., Gabrielatos, C., Hoffmann, S., & Fox, S. (2011). A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory*, *7*(1), 93–118. https://doi.org/10.1515/cllt.2011.005

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., & Kaiser L. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, *7*, 1–15. https://doi.org/10.48550/arXiv.1706.03762

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Volk, M. (1998). The automatic translation of idioms. Machine translation vs. translation memory systems. In: N. Weber (Ed.), *Machine translation: theory, applications, and evaluation. An assessment of the state of the art* (pp. 167–192). Gardez-Verlag.

Wang, H., Wu, H., He, Z., Huang, L. B., & Church, K. W. (2021). Progress in Machine Translation. *Engineering*, *18*, 143–153. https://doi.org/10.1016/j.eng.2021.03.023

Wang, Y. (2023). Research of types and current state of machine translation. *Proceedings of the 2023 International Conference on Machine Learning and Automation*. EWA Publishing. https://doi.org/10.54254/2755-2721/37/20230479

Zhao, Z. (2022). The Machine Translation Model. *Proceedings of the 2022 5th International Conference on Humanities Education and Social Sciences (ICHESS 2022)*. Atlantis Press. https://doi.org/10.2991/978-2-494069-89-3_247

Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., & Zhao, T. (2008). Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points. *Proceedings of the 22nd International Conference on Computational Linguistics*. Organizing Committee.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li1, H., & Liu, T. (2020). *Incorporating BERT into Neural Machine Translation.* Cornell University.

Notes

Authorship contribution

Conceptualization: A. Canga Alonso, M. C. Napoletano
Data collection: M. C. Napoletano
Data analysis: M. C. Napoletano
Results and discussion: A. Canga Alonso, M. C. Napoletano
Review and editing: A. Canga Alonso, M. C. Napoletano
Supervision: A. Canga Alonso

Research dataset

Not applicable.

Funding

Not applicable.

Image copyright

Not applicable.

Approval by ethics committee

Not applicable.

Conflicts of interest

Not applicable.

Data availability statement

The data from this research, which are not included in this work, may be made available by the author upon request.

License

Publisher

Cadernos de Tradução is a publication of the Graduate Program in Translation Studies at the Federal University of Santa Catarina. The journal *Cadernos* de Tradução is hosted by the Portal de Periódicos UFSC. The ideas expressed in this paper are the responsibility of its authors and do not necessarily represent the views of the editors or the university.

Section editors

Andréia Guerini – Willian Moura

Technical editing

Alice S. Rezende – Ingrid Bignardi – João G. P. Silveira – Kamila Oliveira

Article history

Received: 31-07-2024
Approved: 28-02-2024
Revised: 10-04-2025
Published: 05-2025