Can You Tell the Difference? A Study of Human vs Machine-Translated Subtitles

Author: José Ramón Calvo-Ferrer

Affiliation: Universidad de Alicante, Campus de San Vicente del Raspeig, Ap. 99, 03690 Alicante, Spain

Email: jr.calvo@ua.es

Abstract: While machine translation offers the potential for improved efficiency and cost savings, there are concerns about its accuracy and reliability compared to human translation. This study aims to investigate the potential of machine translation systems by analysing viewers' ability to distinguish between subtitles generated by ChatGPT and those created by human translators in the English to Spanish language pair. The study involved 119 Translation and Interpreting degree students who watched eight subtitled clips containing puns, cultural references, humour, and irony: five of these were generated by ChatGPT and the remaining three were created by a human translator. Results indicate that participants were unable to accurately distinguish between ChatGPT-generated and human-generated subtitles, although lower quality subtitles were associated with non-human translation. Factors such as experience with ChatGPT and exposure to subtitled content were not significant predictors of the ability to identify ChatGPT-generated subtitles. However, year of study was found to be a significant predictor, suggesting that translation expertise is a crucial factor for non-human subtitle detection. Overall, these results have important implications for the use of machine translation in subtitle generation and the quality of subtitled content.

Keywords: machine translation; subtitling; ChatGPT; translation quality; human vs machine translation

1. Introduction

In recent years, artificial intelligence (AI) has become an increasingly prevalent technology in a wide variety of fields, from healthcare and finance to entertainment and education. One area that has seen significant development in AI technology is translation. Among the different types of AI, neural machine translation (NMT) is one of the most commonly used types for translation. It uses neural networks to learn the statistical relationships between words in different languages, which allows it to generate more accurate and natural-sounding translations than traditional machine translation methods. This technological advancement has revolutionised the translation field, enabling an increasingly accurate automation of the translation process. Many businesses and organisations are turning to machine translation systems to meet their translation needs, due to their potential to improve efficiency and reduce costs. However, questions remain about its accuracy, reliability, and overall effectiveness compared to human translation. This has led to a debate within the translation community about the role of human translators in the age of AI and the extent to which machines can replace or supplement human translation.

1.1. Subtitling quality and the role of machine translation

Subtitles are an integral part of audio-visual content, allowing viewers to comprehend dialogue or narration in films, TV shows, and other forms of video content. The quality of subtitles is essential in ensuring that the audience understands the meaning and context of foreign audio-visual content. However, subtitling is often constrained by limited space on the screen, requiring translators to be precise and efficient in their translation process without compromising the dialogue's meaning (Díaz-Cintas, 2003). According to Ivarsson (1992), each subtitle must be a coherent logical or syntactical unit, emphasizing the importance of linguistic accuracy. Translating subtitles can also be challenging due to cultural differences, requiring translators to adapt their text to make sense in the target language and culture (Díaz-Cintas, 2003). This adaptation may involve modifications to idiomatic expressions, cultural references, or even character names. As Díaz-Cintas (2003) notes, an ideal translation ensures that the viewer feels they receive accurate information consistent with the source text.

The rise of machine translation (Castilho et al., 2017; Forcada, 2017; Hidalgo-Ternero, 2020; Yulianto & Supriatnaningsih, 2021) has brought about an increasing interest in using automated systems to translate subtitles. The use of machine translation (MT) in subtitling has been gaining popularity due to its potential to improve audio-visual translation workflows. However, subtitling poses challenges for MT—Karakanta, Negri and Turchi (2020) explore the use of NMT in subtitling and the challenges that come with creating proper subtitles in terms of timing and segmentation. The SUMAT project (2014) involving seven European language pairs concluded that abundant work is still required before MT can lead to real improvements in audio-visual translation workflows.

Recent studies have begun to explore the audience reception of machine-translated subtitles. For instance, Karakanta (2022, p. 19) highlights the importance of testing "the impact of fully automatic subtitles on perception and comprehension." In line with this, Tuominen et al. (2023) delve into the gaps in media accessibility and discuss the potential of automated subtitling as a solution. Their findings indicate that while viewers generally accept machine-translated subtitles, they experience a higher cognitive load when reading these subtitles compared to human-translated ones. This increased cognitive load is attributed to shortcomings in the quality of automated speech recognition, translation, and timing of subtitles. These studies underscore the need for further research into the audience reception of machine-translated subtitles to better understand their effectiveness and potential limitations.

1.2. ChatGPT and potential applications

ChatGPT is an intelligent chatbot developed by OpenAI that builds on InstructGPT, a model designed to provide detailed responses to prompts (Dwivedi et al., 2023). As per the official statement, ChatGPT can handle follow-up questions, recognise its errors, challenge incorrect assumptions, and decline inappropriate requests in a conversational format. It incorporates multiple natural language processing capabilities such as answering questions, telling stories, reasoning logically, debugging code, and machine translation. Since its release, ChatGPT has been the subject of intense interest and study, with numerous research papers exploring its capabilities and potential applications (Bhattacharya et al., 2023; Lund et al., 2023; Tlili et al., 2023; Wölfel & Taecharungroj, 2023). Its ability to generate human-like text and carry out complex natural language tasks has opened new avenues for research in a variety of fields, from machine translation to natural language understanding and beyond (Jiao et al., 2023; Kasneci et al., 2023; Liebrenz et al., 2023; Lund & Wang, 2023; Rudolph et al., 2023).

ChatGPT's ability to pass off as human has been demonstrated in various settings, including passing exams—some studies have reported successfully using ChatGPT to take tests (Bommarito & Katz, 2023), which highlights the model's advanced natural language processing capabilities. Popel et al. (2020) found that most participants in a Translation Turing test were unable to distinguish between translations made by a neural-based translation system and humans. In fact, it is believed that deep learning approaches have the potential to revolutionise the translation industry and improve the efficiency and accuracy of translation workflows. In exploring the impact of automatic subtitling on professionals in the field, Karakanta et al. (2022) found that many subtitlers have a neutral to positive user experience with the technology, appreciating its potential to save time and effort. However, AI-generated translations still suffer from various errors that can affect viewer comprehension and enjoyment, and impact cognitive load and learning negatively (Chan et al., 2019).

1.3. Machine translation quality

The increasing use of AI in translation has sparked a debate about the role of human translators in the age of machine translation, as shown by the recent number of studies that have evaluated the accuracy and reliability of machine translation systems such as Google Translate and DeepL (Cambedda et al., 2021; Hidalgo-Ternero, 2020; Minervini, 2021; Papa & Tavosanis, 2020).

Research has focused on evaluating the quality of different systems and comparing them to human translation. One widely used metric for measuring the quality of machine translation is the BLEU (Bilingual Evaluation Understudy) score, which measures the degree of overlap between a machine-generated translation and a human reference translation (Papineni et al., 2002). However, it's worth noting that BLEU, while useful, primarily assesses lexical similarity and doesn't fully capture semantic accuracy. This has led to the development of newer, neural metrics that aim to better align with human judgements, which is still considered the benchmark. In line with this, different studies have shown that while machine translation systems have improved significantly in recent years, they still lag behind human translation in terms of overall quality and accuracy (Al-Kabi et al., 2013; Al-Rukban & Saudagar, 2017; Hagström & Pedersen, 2022; Mathur et al., 2020).

The quality of machine-translated subtitles can vary significantly, particularly in capturing cultural nuances and idiomatic expressions. Post-editing of machine-translated subtitles can improve their accuracy (Koponen, 2016), but the outcome may not always meet expectations. As a matter of fact, the use of automatic translation in subtitling has recently sparked controversy in the context of the popular Netflix series "Squid Game." Critics on social networking sites have accused the English subtitles of failing to accurately capture the essence of the series. In Spain, a multinational company relied on machine translation and minimal post-editing to produce the subtitles of the series. This approach has been criticised as it is considered to impoverish the sector and to result in lower quality subtitles (Llanos Martínez, 2021). Furthermore, it is claimed that this practice benefits only large companies that aim to reduce costs, raising broader questions about the role of technology and automation in the translation industry. While such tools can enhance efficiency and quality, there is concern that they may also be used to impoverish quality in favour of economic benefits.

Nonetheless, despite both the growing interest and controversy surrounding the use of machine translation, few studies have specifically examined whether viewers can actually distinguish between translations produced by machines and those produced by humans. Investigating this ability is crucial in shedding light on the limitations of current MT systems and the continued importance of human translators in the translation industry.

1.4. Research questions and hypotheses

Based on the literature review, the following questions were examined:

1. Can viewers differentiate between subtitles generated automatically by ChatGPT and those generated by humans?
2. How does the quality of ChatGPT-generated subtitles compare to that of human-generated subtitles, according to viewers?
3. Do viewers attribute poor quality subtitles to non-human translation?
4. What factors influence viewers' ability to distinguish between automatically generated and human-generated subtitles?

To address these questions, the hypotheses of the study were formulated as follows:

H1. Participants will be able to distinguish between ChatGPT-generated and human-generated subtitles.

H2. ChatGPT-generated subtitles will be rated lower in quality than human-generated subtitles.

H3. Subtitles perceived as of poor quality will be attributed to ChatGPT.

H4. Viewers' ability to identify ChatGPT-generated subtitles will be positively affected by factors such as experience with ChatGPT, translation expertise, and exposure to subtitled content.

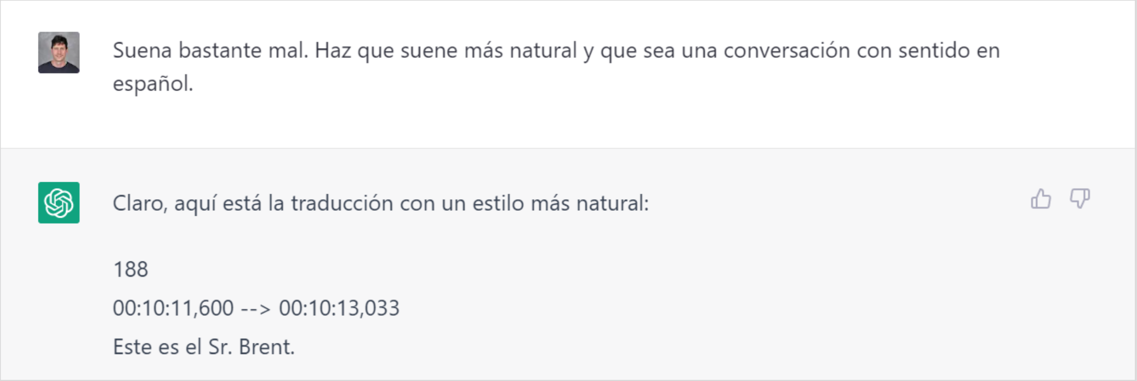
1.5. The current study

These hypotheses were tested in an experiment where students from the degree in Translation and Interpreting watched video clips with both ChatGPT-generated and human-generated English to Spanish subtitles. Participants were asked to rate the quality of the subtitles and identify which had been machine-generated. They also completed a questionnaire assessing their experience with ChatGPT, translation expertise, and exposure to subtitled content. The results of the study are expected to have implications for translator training in the field of audio-visual translation, as well as to contribute to the understanding of the strengths and limitations of automated subtitle generation and the factors that influence its perception and acceptance. To date, there appears to be no prior research investigating the impact of ChatGPT-generated versus human-generated subtitles on user satisfaction and detection.

2. Materials and method

2.1. Sample

The participants for this study were students of the degree in Translation and Interpreting at the University of Alicante (N = 119) either in their first (n = 66) or final year of study (n = 53), of which 96 identified as female (80.7%) and 23 as male (19.3%).



Suena bastante mal. Haz que suene más natural y que sea una conversación con sentido en español.

Claro, aquí está la traducción con un estilo más natural:

188
00:10:11,600 --> 00:10:13,033
Este es el Sr. Brent.

Figure 1. Prompt used to improve idiomaticity of GhatGPT-generated subtitles

2.2. Materials

The purpose of this study was to investigate viewers' ability to distinguish between subtitles generated by ChatGPT and those created by human translators, together with the factors that may influence this ability. To achieve this goal, eight clips from the first episode of season one of the TV series *The Office* (Gervais & Merchant, 2001) containing instances of puns, cultural references, humour, and irony, which have been identified as common translation in audio-visual translation in previous studies (Attardo, 2002; Chiaro, 2008; Martínez-Sierra, 2006; Pedersen, 2009; Zabalbeascoa, 1996), were selected. To create the materials for the experiment, two sets of subtitles were generated for each clip: one created by a human translator, and one generated automatically by ChatGPT. ChatGPT was selected for this study as it has been found to outperform other encoder-decoder MT models, such as those from DeepL, Google, OPUS, and NLLB, as well as large-scale language models like BLOOM and BLOOMZ, across various language pairs (Moslem et al., 2023). The subtitles in Spanish created by a human translator were extracted from the actual DVD of *The Office* commercialised in Spain. For the ChatGPT-generated subtitles, the English subtitles from the DVD were first translated to Spanish by ChatGPT using the neural machine translation feature. This process began with the creation of eight clips, each similar in length, number of utterances, and containing a comparable number of translation challenges such as humour, puns, or cultural references. The subtitles for each clip, including timestamps, were extracted and input into ChatGPT as a single block of text. This

Table 1. Selected clips from *The Office* Season 1, Episode 1 “Downsize” (Gervais and Merchant, 2001)

Clip	Subtitle source	Beginning	End	Duration	Number of subs.	Word count	Example of RP (Rich points)
1	DVD	00:00:31.320	00:01:46.316	00:01:14,996	42	192	AKA, for you.
2	DVD	00:01:49.800	00:03:04.473	00:01:14,673	29	165	El vino did flow.
3	ChatGPT	00:04:24.080	00:05:21.678	00:00:57,598	30	142	All right. What is it, time of the month?
4	ChatGPT	00:05:24.320	00:06:20.514	00:00:56,194	42	132	Hypocrite warning.
5	ChatGPT	00:10:11.600	00:11:17.158	00:01:05,558	50	208	"Ricky. No, Ricky!" What was his girlfriend’s name on EastEnders?
6	DVD	00:13:25.360	00:15:20.519	00:01:55,159	42	338	You should put him in custard-y!
7	ChatGPT	00:15:22.280	00:16:37.954	00:01:15,674	27	181	I like to have a laugh just as much as the next man,
8	ChatGPT	00:18:16.520	00:19:01.630	00:00:45,110	16	115	You get the knife in behind the windpipe, pull it down like that.

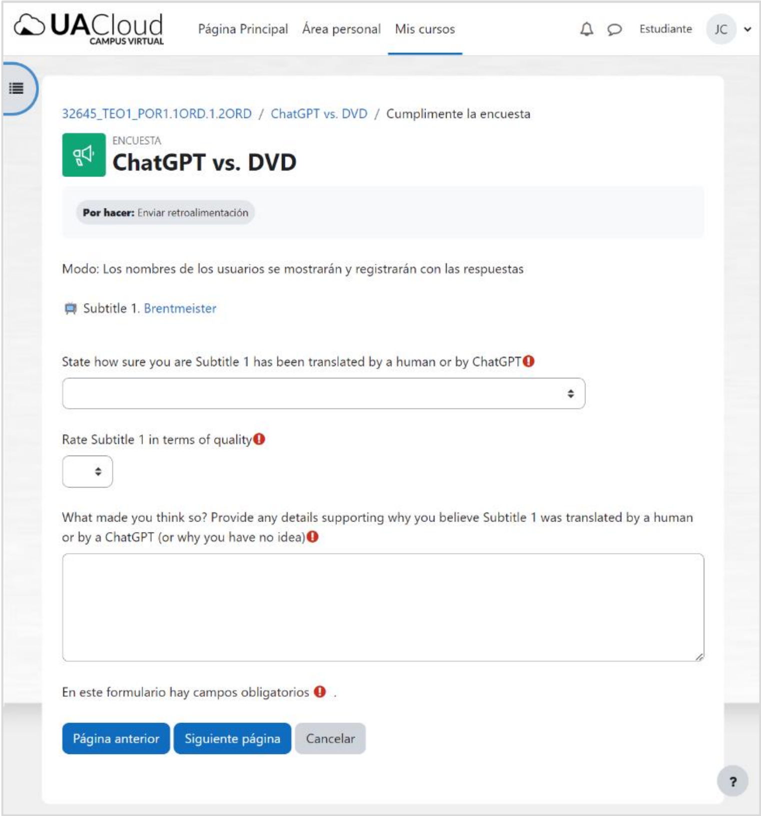
approach ensured that the context of the entire clip was considered during translation, rather than translating individual subtitles in isolation. In an attempt to ensure idiomaticity, ChatGPT was provided with two specific prompts—“Please translate the following subtitles into Spanish. Do not modify the time codes” and “It sounds awful. Please make it sound more natural in Spanish and make sure the conversation makes sense” (Figure 1). Once the translation was generated, the translated subtitles were copied into a new Spanish subtitle file and saved. The resulting subtitles were then burned into the video clips using an open-source video transcoder. Table 1 presents a summary of the eight clips selected, detailing their duration, the quantity of subtitles, the count of source words, and instances of “rich points”. These “rich points” refer to “source text segments containing prototypical translation problems” (Hurtado Albir et al., 2020, p. 114).

2.3. Instruments

A questionnaire was designed to elicit information from participants regarding their perceptions of the eight clips used in the experiment. The questionnaire included two sets of questions: contextual questions and clip-specific questions. While participants’ translation expertise was inferred from their year in their degree programme, contextual questions asked participants to rate their self-assessed proficiency in using ChatGPT, their ability to spot mistakes while reading texts in Spanish and while watching subtitled films, and their preference for watching dubbed films, which would serve as a proxy for understanding the participants’ exposure to and preference for subtitling. The clip-specific questions were asked after each clip. The first question asked participants to indicate their certainty about the origin of the subtitles on a scale from -3 to 3, where -3 indicated complete certainty that the subtitles were translated by ChatGPT, and 3 indicated complete certainty that the subtitles were translated by a human translator. A second question asked participants to rate the perceived quality of the subtitles on a scale from 0 to 10. The last question after each clip was an open-ended one, to gather any details supporting their opinion on whether the subtitles were ChatGPT-generated or human-generated. The questionnaire was administered through the course Moodle page, as shown in Figure 2.

2.4. Procedure

The experiment was conducted in a classroom setting in two single sessions, with participants in each year watching the videos together, as Figure 3 shows. This approach aimed to ensure that all participants viewed the same video clips under the same conditions, allowing for accurate comparisons between their responses. At the beginning of the session, an unrelated clip was



32645_TEO1_POR1.1ORD.1.2ORD / ChatGPT vs. DVD / Cumplimente la encuesta

ENCUESTA

ChatGPT vs. DVD

Por hacer: Enviar retroalimentación

Modo: Los nombres de los usuarios se mostrarán y registrarán con las respuestas

Subtitle 1. Brentmeister

State how sure you are Subtitle 1 has been translated by a human or by ChatGPT

Rate Subtitle 1 in terms of quality

What made you think so? Provide any details supporting why you believe Subtitle 1 was translated by a human or by a ChatGPT (or why you have no idea)

En este formulario hay campos obligatorios .

Figure 2. Survey question for each of the clips used in the experiment

played on screen to ensure that all students were able to read the subtitles properly. Students who had difficulty viewing the screen were given the opportunity to move closer, although no one decided to do so. Additionally, to further ensure visibility, students were provided with links to each video clip in the questionnaire. This allowed them to access and view the clips individually if needed.

In the study, all participants watched the same set of subtitles. The researcher randomly presented five clips containing subtitles generated by ChatGPT version 3.5 and three others with the commercial subtitles, made by human translators, included in the actual DVD of the series commercialised in Spain. Each clip was played only once. After viewing each clip, participants were asked to rate the quality of the subtitles on a scale of 0 to 10 and to indicate whether they believed they had been generated by ChatGPT or by a human translator, along with some reflections on the reasons underlying their opinions. The rating was intended to capture their overall impression of the subtitles, although no specific assessment rubric was suggested for the sake of conciseness. As a result, the concept of ‘quality’ was subjectively construed by the participants and may have been individually influenced by various factors such as timing, segmentation, and handling of cultural references and humour.

2.5. Data analysis

To investigate whether participants were able to distinguish between ChatGPT-generated and human-generated subtitles, data on the responses to each of the items of the questionnaire was collected and used to create several variables, including number of ChatGPT subtitles identified, average grade received by ChatGPT subtitles, average grade received by human subtitles, perceived subtitle quality, and perceived subtitle authorship.

Bivariate correlations were first computed to explore relationships between variables, and several statistically significant correlations between ability to identify subtitle authorship and other

Figure 3. Participants taking part in the study

measured variables were found. Multivariate linear regressions were then carried out to investigate the predictive strength of such variables. All analyses were performed using SPSS 22.0 statistical software with a significance level of 0.05. Key assumptions of the linear regression model, including normality, linearity, homoscedasticity, and absence of multicollinearity and autocorrelation, were checked and confirmed using the approach recommended by Vilà, Torrado and Reguant (2019). Finally, independent sample t-tests were also conducted to determine if there were any differences in the number of ChatGPT-generated subtitles identified owing to the participants' expertise in translation, as determined by their year of study.

3. Results

3.1. Ability to identify subtitle authorship

In order to investigate whether participants were able to distinguish between subtitles generated by ChatGPT and those created by human translators, two descriptive statistics analyses were performed. The first analysis assessed the number of correct answers in connection with the authorship of the eight subtitles included in the study, whereas the second examined how many of the five ChatGPT-generated subtitles were identified as such by participants and not attributed to a human translator. Furthermore, the students' degree of certainty in their answers (e.g., irrespective of being right or wrong, "I am completely sure it has been translated by ChatGPT" shows a greater degree of confidence than "I would say it has been translated by a human translator") was computed to investigate any potential effects on the dependent variables.

Descriptive statistics for the variables under investigation are presented in Table 2. The mean number of ChatGPT subtitles identified was 2.07 (*SD* = 1.064), indicating that participants correctly attributed an average of two out of the five ChatGPT-generated subtitles. The median and mode were both 2, suggesting that this performance was typical across participants. Similarly, the mean number of total right answers was 3.06 (*SD* = 1.33), which shows that participants accurately attributed an average of three out of the eight subtitles. The median and mode were both 3, indicating also that this level of performance was typical across participants. Regarding the degree of certainty (*M* = 10.63, *SD* = 3.842), participants appeared to be somewhat certain about their attributions of authorship. However, there was wide variability in the level of certainty across individuals. Out of 24, the minimum degree of certainty was 3, indicating that some participants were uncertain about the authorship of the subtitles. On the other hand, the maximum degree of certainty was 24, suggesting that other participants were totally confident in their attributions.

Table 2. Statistics

		Mean degree of certainty	ChatGPT subtitles identified	Total right answers
N	Valid	119	119	119
	Missing	0	0	0
Mean		10.63	2.07	3.06
Median		10.00	2.00	3.00
Mode		9[a]	2	3
Std. Deviation		3.842	1.064	1.330
Variance		14.760	1.131	1.768
Minimum		3	0	0
Maximum		24	5	6

a. Multiple modes exist. The smallest value is shown

3.2. Participants' subtitle assessment

In order to investigate whether viewers perceived ChatGPT-generated subtitles to be of equal quality to human-generated subtitles, a descriptive statistics analysis was initially conducted to compare the grades given by viewers to the eight subtitles in the study, five of which were ChatGPT-generated and three human-generated. The results of the analysis showed that, on a scale from 0 to 10, the average grade received by ChatGPT-generated subtitles was 6.71 (*SD* = 1.229), while the average grade received by human-generated subtitles was 6.64 (*SD* = 1.274). Subsequently, a Wilcoxon Signed Ranks Test was performed, the results of which showed that there were 42 negative ranks, 38 positive ranks, and 39 ties, as Table 3 shows. Also, the calculated Z-statistic was -0.522 with a p-value of 0.602, indicating that there was no significant difference between the average grade received by ChatGPT subtitles and the average grade received by human subtitles, thus suggesting that viewers do not significantly differentiate between the quality of ChatGPT-generated and human-generated subtitles.

Table 3. Wilcoxon Signed Ranks Test

		N	Mean Rank	Sum of Ranks
Average grade received by ChatGPT subtitles - Average grade received by human subtitles	Negative Ranks	42[a]	36.06	1514.50
	Positive Ranks	38[b]	45.41	1725.50
	Ties	39[c]		
	Total	119		

a. Average grade received by ChatGPT subtitles < Average grade received by human subtitles

b. Average grade received by ChatGPT subtitles > Average grade received by human subtitles

c. Average grade received by ChatGPT subtitles = Average grade received by human subtitles

3.3. Participants' perceived subtitle quality

The third hypothesis of the study aimed to explore whether participants associated poor quality subtitles with non-human translation. To this end, a correlation analysis was conducted between perceived subtitle authorship (i.e., whether participants believed the subtitles were generated by a human or by ChatGPT) and overall perceived quality (i.e., subtitle rating as reported by the study participants). The correlation analysis was followed by a linear regression analysis to assess the strength of the perceived subtitle authorship variable in predicting perceived subtitle quality.

The analysis revealed a positive correlation between perceived subtitle authorship and overall perceived subtitle quality, with a Pearson correlation coefficient of 0.314 and a significance level of 0.01. These results suggest a bidirectional relationship—participants were more likely to rate subtitles as high quality when they believed they were generated by a human translator, and

Table 4. Descriptive statistics for contextual questions

	M	SD
I am proficient in the use of ChatGPT	4.50	1.939
I usually spot mistakes when I watch a film with subtitles	5.34	1.210
I do not like watching dubbed films	4.08	1.867
I typically detect typos when I am reading a text in Spanish	5.50	1.314

Note: $N = 119$

conversely, they were more likely to believe the subtitles were human-translated when they perceived them to be of high quality.

The linear regression analysis confirmed this relationship, with the model including "perceived subtitle authorship" as a predictor variable explaining a significant proportion of the variance in overall perceived subtitle quality (R^2 = 0.099). The regression coefficient for perceived subtitle authorship was 0.633, indicating that the perceived subtitle quality rating increased by an average of 0.633 points for every one-point increase in the perceived human authorship score, after controlling for the constant term. The t-value for the regression coefficient was 3.582, which was statistically significant at the 0.01 level. These findings highlight the complex interplay between perceived subtitle quality and perceived subtitle authorship, suggesting that both factors significantly influence each other.

3.4. Factors predicting ability to identify subtitle authorship

A regression analysis was performed in order to explore whether factors such as experience with ChatGPT, translation expertise, and exposure to subtitled content, as measured by the items in the questionnaire shown in Table 4, could predict the participants' ability to detect ChatGPT-generated subtitles. All in all, the results show that the overall model was not significant (F(3, 115) = 2.128, p = 0.1) and that it accounted for only 5.3% of the variance in the ability to identify ChatGPT-generated subtitles, with an R^2 value of 0.053. Proficiency in the use of ChatGPT and preference for watching dubbed films were not significant predictors of the ability to identify ChatGPT-generated subtitles (β = 0.148, p = 0.162 and β = -0.101, p = 0.299, respectively), while year of study did have some impact on the ability to identify ChatGPT-generated subtitles (β = 0.244, p = 0.017), which suggests that as translation expertise increases, viewers' ability to distinguish between ChatGPT-generated and human-generated subtitles also increases.

To further explore these results, a t-test was carried out to compare the mean ChatGPT subtitles identified between the two groups of participants that took part in the study (i.e., students in their first and in their fourth year of their degree in Translation and Interpreting). The results show that the mean ChatGPT subtitles identified by participants in their fourth year of study (M = 2.28, SD = 1.150) was higher than that of first-year students (M = 1.89, SD = 0.963), t(117) = -2.009, p = 0.047), confirming that there was a significant difference in the number of ChatGPT-generated subtitles identified by both groups.

4. Discussion

This study aimed to investigate viewers' ability to distinguish between subtitles generated automatically by ChatGPT and those generated by humans, to compare the quality of ChatGPT-generated subtitles to that of human-generated subtitles, and to examine whether viewers attribute poor-quality subtitles to non-human translation. Additionally, the study aimed to identify the factors that influence viewers' ability to differentiate between ChatGPT-generated and human-generated subtitles. The findings suggest that participants were not able to accurately distinguish between ChatGPT-generated and human-generated subtitles in the context of the present study and support the notion that AI-generated content may be assumed equal in terms of quality to that generated by humans (Graefe et al., 2018; Serban et al., 2016; Wölker & Powell, 2021). While the overall results showed a similar degree of accuracy, there were some small differences between the subtitles that were identified. These small differences could be due to several factors, such as the specific content of the subtitles, the length of the subtitles, or the participants' individual language skills. Further research is needed to investigate these differences and to understand how they affect the ability of humans to distinguish between ChatGPT-generated subtitles and human-generated subtitles.

Also, although translations generated by ChatGPT were generally similar to those produced by human translators, students' perceptions of certain strategies varied widely—while some were seen as natural solutions indicative of human translation, others viewed them as unnaturally foreign and unlikely to be chosen by a professional human translator. The findings also align with Hu, O'Brien & Kenny (2020), who discovered that while fully post-edited machine-translated subtitles lead to better reception among viewers compared to raw machine-translated subtitles, human-translated subtitles do not necessarily result in superior reception. However, it's crucial to note that the findings of the present study are specific to the English-Spanish language pair, one of the best-resourced language pairs (Ortega et al., 2020), and that the results may not be generalisable to other language pairs, particularly those that are less resourced. Further research is required to explore the performance of MT-generated subtitles across different language pairs and contexts. Also, future research could benefit from a more interlingual comparative approach, investigating the differences between machine-generated and human-generated subtitles, and examining how these differences influence user perceptions and evaluations.

The findings of the study suggest that the second hypothesis, which proposed that ChatGPT-generated subtitles would be rated lower in quality than human-generated subtitles, was not supported by the data. In fact, the analysis indicated that ChatGPT-generated subtitles received an even slightly higher average grade than human-generated subtitles, and there was no statistically significant difference between the quality ratings of the two types of subtitles. These results suggest that viewers may not be able to differentiate between the quality of ChatGPT-generated and human-generated subtitles, indicating that, if provided with relevant prompts to ensure quality and adequacy, the use of AI-generated subtitles may be a viable option for subtitling in the future, as suggested by Karakanta (2022) and Papi et al. (2022). These results are also in line with previous studies, which suggest that machine translation quality "can either be as high as when manually translating, or that it can even be improved" (Screen, 2017, p. 137).

The findings of the third hypothesis indicate that participants tended to associate lower quality subtitles with non-human translation, as they were more likely to rate subtitles as high quality when they believed they were generated by a human translator as opposed to when they believed they were generated by ChatGPT. These results underscore the significance of authorship in shaping the perceived quality of subtitles and align with previous studies referring to a "stronger preference for human translations" (Läubli et al., 2018, p. 4795). Ensuring high-quality AI-generated subtitles is thus essential, as viewers may reject audio-visual material if they perceive poor quality in non-human translation.

The fourth hypothesis aimed to investigate the factors that predict the ability to detect ChatGPT-generated subtitles. The findings show that proficiency in the use of ChatGPT and preference for watching dubbed films did not have a significant impact on the ability to identify ChatGPT-generated subtitles. However, year of study was found to be a significant predictor—fourth-year students were able to identify subtitle authorship more effectively compared to first-year students, suggesting that translation expertise and experience are crucial factors for detecting ChatGPT-generated subtitles. The controversy surrounding the accuracy of the translation of the series "Squid Game" prompts an interesting question: did the criticism stem from the general public or was it initially brought to light by translation industry experts who noticed the discrepancies in the subtitles and spoke out about them? This stresses the complex dynamics of audience reception, where viewers without translation experience may not notice errors, leading to misunderstandings about the content, and further emphasising the importance of high-quality subtitles. The debate highlights the potential issues with using machine translation and minimal post-editing in producing subtitles and underscores the role of human translators in the translation process.

This study does not come without limitations: first, it must be noted that the sample consisted only of students of the degree in Translation and Interpreting at the University of Alicante, which may limit the generalizability of the results to other populations. While the participants in the study were selected based on their expertise in translation and interpreting, the results may not be representative of the wider population. Second, the experiment was conducted in a classroom setting with participants watching the videos together, which may not reflect the real-world experience of watching videos with subtitles. This might have also caused participants to have been influenced by social cues and the presence of other participants, thus affecting their ratings of the subtitles. Third, it must be noted that the prompts given to enhance the quality of the generated subtitles may have biased the results in favour of ChatGPT, and that other results might have arisen in the absence of such commands. Fourth, the study focused on a limited number of variables, such as the ability to identify subtitle authorship and perceived subtitle quality, which, as suggested by the low R^2 values, may not capture the full range of factors that influence the perception of automated subtitle generation. Fifth, the study used only one TV series as the source of the video clips, which may not be representative of all audio-visual content. Different types of genres, such as comedy, action, thriller, etc., may require different types of subtitles to meet viewer needs. In line with this, *The Office* (Gervais & Merchant, 2001) contains elements which are difficult to translate accurately such as humour, irony, wordplay, and cultural references, etc.—text with less problematic elements might have made the stylistic differences between human-generated and ChatGPT-generated subtitles more evident. As a matter of fact, authorship of human subtitles struggling to convey meaning properly appears to have been attributed to ChatGPT. Sixth, the study relied on participants' subjective ratings of both their proficiency in the use of ChatGPT and their notion regarding the concept of quality, which may be influenced by individual biases and may not accurately reflect actual skills or subtitle adequacy. Also, the audience's familiarity with the original language and culture was assumed to be homogeneous, although different proficiency levels might have led to different capacities to identify non-human subtitles. As discussed above, the study's findings, specific to the well-resourced English-Spanish language pair, may not generalise to less-resourced language pairs, necessitating further research. Finally, it is worth considering the possibility that the participants' awareness that some of the subtitles were generated by ChatGPT influenced their ratings. Although the participants did not know which specific subtitles were machine-generated, they may have been more critical or vigilant in their evaluations, knowing that machine translation was involved.

However, these limitations may open new avenues for research. As suggested, while this study focused on a single TV series, future research could examine the performance of ChatGPT-generated subtitles on a broader range of audio-visual content, such as films, documentaries, news broadcasts, and online videos. ChatGPT-generated subtitles struggled to convey complex meaning, such as humour, irony, and cultural references. Future research could explore ways to improve the quality of ChatGPT-generated subtitles for these types of content, such as by incorporating more context or cultural knowledge into the algorithm, and by measuring the effect of the viewers' knowledge of the source language and culture. Regarding the notion of quality, future research in the field of reception studies could investigate the relationship between perceived subtitle quality, possibly construed according to the FAR (Functional equivalence – Acceptability – Readability) model (Pedersen, 2017) and its effect on viewers' assumptions regarding subtitle authorship. This could lead to a more nuanced understanding of subtitle quality and inform future comparisons between human-generated and MT-generated subtitles.

It would also be interesting to explore how different prompts may increase its effectiveness in terms of accuracy, speed, and ease of use, both at the professional level and in translator training. In line with this, automated translation tools have the potential to speed up the translation process and, subsequently, to make content more accessible to people with hearing impairments and to those who are learning a new language. Thus, future research could evaluate the impact of ChatGPT-generated subtitles on language learning outcomes and the overall accessibility of audio-visual content, and how it could be implemented or embedded in systems designed for such purposes. Finally, to mitigate the potential influence of participants' awareness of machine-generated subtitles, future studies could initially withhold this information. Following this, a debriefing session could reveal the study's true nature, ensuring unbiased assessments while maintaining ethical transparency.

5. Conclusion

The findings of this study have several implications for the field of automated subtitle generation and the use of AI in translation. First, the study suggests that ChatGPT-generated subtitles may be comparable in quality to human-generated subtitles in terms of perceived quality and accuracy. Although the transcription of dialogue and audio cues for subtitles for the deaf and hard of hearing (SDH) still requires human intervention, the development of AI-generated subtitles could improve accessibility of audiovisual content for individuals who are deaf or hard of hearing. Also, the study highlights the potential benefits of using translation prompts to improve the quality of non-human subtitles. This could lead to different analyses regarding how AI processes user commands and how they can be used in language learning.

All in all, this study emphasises the importance of translator training and suggests that AI-generated subtitles have the potential to provide a viable alternative to human-generated subtitles. However, a cautious approach is necessary when implementing these subtitles, as it is crucial to consider viewers' perceptions and attitudes towards non-human translation.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and following the regulations in force at the University of Alicante (Spain) for studies involving humans: https://web.ua.es/en/vr-investigacio/comite-etica/presentation.html.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: This study analysed publicly available datasets which can be found at http://hdl.handle.net/10045/133278.

Conflicts of Interest: The author declares no conflict of interest.

References

Al-Kabi, M. N., Hailat, T. M., Al-Shawakfa, E. M., & Alsmadi, I. M. (2013). Evaluating English to Arabic Machine Translation Using BLEU. *(IJACSA) International Journal of Advanced Computer Science and Applications*, *4*(1), 66–73.

Al-Rukban, A., & Saudagar, A. K. J. (2017). Evaluation of English to Arabic machine translation systems using BLEU and GTM. *ACM International Conference Proceeding Series*, 228–232. https://doi.org/10.1145/3175536.3175570

Attardo, S. (2002). Translation and humour: An approach based on the general theory of verbal humour (GTVH). *Translator*, *8*(2), 173–194. https://doi.org/10.1080/13556509.2002.10799131

Bhattacharya, K., Bhattacharya, A. S., Bhattacharya, N., Yagnik, V. D., Garg, P., & Kumar, S. (2023). ChatGPT in Surgical Practice—a New Kid on the Block. *Indian Journal of Surgery*, 1–4. https://doi.org/10.1007/S12262-023-03727-X/METRICS

Bommarito, M. J., & Katz, D. M. (2023). GPT Takes the Bar Exam. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4314839

Cambedda, G., Di Nunzio, G. M., & Nosilia, V. (2021). A Study on Automatic Machine Translation Tools: A Comparative Error Analysis Between DeepL and Yandex for Russian-Italian Medical Translation. *Umanistica Digitale*, *10*(10), 139–163. https://doi.org/10.6092/ISSN.2532-8816/12631

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109–120. https://doi.org/10.1515/PRALIN-2017-0013

Chan, W. S., Kruger, J. L., & Doherty, S. (2019). Comparing the impact of automatically generated and corrected subtitles on cognitive load and learning in a first- and second-language educational context. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *18*, 237–272. https://doi.org/10.52034/LANSTTS.V18I0.506

Chiaro, D. (2008). Issues in Audiovisual Translation. In J. Munday (Ed.), *The Routledge Companion to Translation Studies* (pp. 155–179). Routledge. https://doi.org/10.4324/9780203879450-16

Díaz-Cintas, J. (2003). Teoría y práctica de la subtitulación inglés-español. *Ariel Cine*.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/J.IJINFOMGT.2023.102642

European Commission. (2014). *SUMAT: An Online Service for SUbtitling by MAchine Translation | SUMAT Project | Fact Sheet | CIP | CORDIS | European Commission*. https://cordis.europa.eu/project/id/270919

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces*, *6*(2), 291–309. https://doi.org/10.1075/TS.6.2.06FOR

Gervais, R., & Merchant, S. (2001). *The Office*. BBC.

Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Hagström, H., & Pedersen, J. (2022). Subtitles in the 2020s: The Influence of Machine Translation. *Journal of Audiovisual Translation*, *5*(1), 207–225.

Hidalgo-Ternero, C. M. (2020). Google Translate vs. DeepL: *MonTI. Monografías de Traducción e Interpretación*, *6*(Especial 6), 154–177. https://doi.org/10.6035/MONTI.2020.NE6.5

Hu, K., O'Brien, S., & Kenny, D. (2020). A reception study of machine translated subtitles for MOOCs. *Perspectives: Studies in Translation Theory and Practice*, *28*(4), 521–538. https://doi.org/10.1080/0907676X.2019.1595069

Hurtado Albir, A., Galán-Mañas, A., Kuznik, A., Olalla-Soler, C., Rodríguez-Inés, P., & Romero, L. (2020). Translation competence acquisition. Design and results of the PACTE group's experimental research. *Interpreter and Translator Trainer*, *14*(2), 95–233. https://doi.org/10.1080/1750399X.2020.1732601

Ivarsson, J. (1992). *Subtitling for the media: A handbook of an art*. Transedit.

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023, January 20). *Is ChatGPT A Good Translator? A Preliminary Study*. ArXiv. https://doi.org/https://doi.org/10.48550/arXiv.2301.08745

Karakanta, A. (2022). *Automatic subtitling: A new paradigm* [Università degli studi di Trento]. https://hdl.handle.net/11572/356701

Karakanta, A., Bentivogli, L., Cettolo, M., Negri, M., & Turchi, M. (2022). Post-editing in Automatic Subtitling: A Subtitlers' Perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 261–270). European Association for Machine Translation.

Karakanta, A., Negri, M., & Turchi, M. (2020). Is 42 the Answer to Everything in Subtitling-oriented Speech Translation? In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)* (pp. 209–219). Association for Computational Linguistics. https://doi.org/https://doi.org/10.18653/v1/P17

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/J.LINDIF.2023.102274

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *JoSTrans. The Journal of Specialised Translation*, *25*. https://www.jostrans.org/issue25/art_koponen.php

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *N Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791–4796. https://www.proz.com

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *The Lancet. Digital Health*, *5*(3), e105–e106. https://doi.org/10.1016/S2589-7500(23)00019-5

Llanos Martínez, H. (2021, October 14). *Los traductores españoles protestan por los "mediocres" subtítulos de 'El juego del calamar', hechos por una máquina*. https://elpais.com/television/2021-10-14/los-traductores-espanoles-protestan-por-los-mediocres-subtitulos-de-el-juego-del-calamar-hechos-por-una-maquina.html

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/LHTN-01-2023-0009

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/ASI.24750

Martínez-Sierra, J. J. (2006). Translating audiovisual humour. A case study. *Perspectives: Studies in Translatology*, *13*(4), 289–296. https://doi.org/10.1080/09076760608668999

Mathur, N., Baldwin, T., & Cohn, T. (2020). *Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics*. https://arxiv.org/abs/2006.06264

Minervini, R. (2021). La traducción automática del género (español-italiano): análisis de ejemplos traducidos con DeepL y Google Traductor. *Rivista Internazionale Di Tecnica Della Traduzione / International Journal of Translation*, *23*, 105–127. https://www.openstarts.units.it/handle/10077/33237

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 227–237). European Association for Machine Translation.

Ortega, J. E., Castro Mamani, R., & Cho, K. (2020). Neural machine translation with a polysynthetic low resource language. *Machine Translation*, *34*(4), 325–346. https://doi.org/10.1007/s10590-020-09255-9

Papa, S., & Tavosanis, M. (2020). Valutazione umana di deepl a livello di frase per le traduzioni di testi specialistici dall'inglese verso l'italiano. *CEUR Workshop Proceedings*, *2769*, 494–525. https://doi.org/10.4000/books.aaccademia.8924

Papi, S., Gaido, M., Karakanta, A., Cettolo, M., Negri, M., & Turchi, M. (2022). Direct Speech Translation for Automatic Subtitling. *ArXiv*. http://arxiv.org/abs/2209.13192

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *2002-July*, 311–318.

Pedersen, J. (2009). Cultural Interchangeability: The Effects of Substituting Cultural References in Subtitling. *Http://Dx.Doi.Org/10.2167/Pst003.0*, *15*(1), 30–48. https://doi.org/10.2167/PST003.0

Pedersen, J. (2017). The FAR model: Assessing quality in interlingual subtitling. *Journal of Specialised Translation*, *28*, 210–229.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, *11*(1), 1–15. https://doi.org/10.1038/s41467-020-18073-9

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1). https://doi.org/10.37074/JALT.2023.6.1.9

Screen, B. (2017). Productivity and quality when editing machine translation and translation memory outputs: an empirical analysis of English to Welsh translation. *Studia Celtica Posnaniensia*, *2*(1), 1–24. https://doi.org/10.1515/SCP-2017-0007

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., & Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, *1*, 588–598. https://doi.org/10.18653/v1/p16-1056

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*(1), 1–24. https://doi.org/10.1186/S40561-023-00237-X/FIGURES/13

Tuominen, T., Koponen, M., Vitikainen, K., Sulubacak, U., & Tiedemann, J. (2023). Exploring the gaps in linguistic accessibility of media: The potential of automated subtitling as a solution. *Journal of Specialised Translation*, *39*, 77–98. https://researchportal.helsinki.fi/en/publications/exploring-the-gaps-in-linguistic-accessibility-of-media-the-poten

Vilà Baños, R., Torrado Fonseca, M., & Reguant Álvarez, M. (2019). Análisis de regresión lineal múltiple con SPSS: un ejemplo práctico. *REIRE*, *12*(2), 1–10. https://doi.org/10.1344/reire2019.12.222704

Wölfel, M., & Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing 2023, Vol. 7, Page 35*, *7*(1), 35. https://doi.org/10.3390/BDCC7010035

Wölker, A., & Powell, T. E. (2021). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, *22*(1), 86–103. https://doi.org/10.1177/1464884918757072

Yulianto, A., & Supriatnaningsih, R. (2021). Google Translate vs. DeepL: A quantitative evaluation of close-language pair translation (French to English). *AJELP: Asian Journal of English Language and Pedagogy*, *9*(2), 109–127. https://doi.org/10.37134/AJELP.VOL9.2.9.2021

Zabalbeascoa, P. (1996). Translating jokes for dubbed television situation comedies. *Translator*, *2*(2), 235–257. https://doi.org/10.1080/13556509.1996.10798976