Human- or machine-translated subtitles: Who can tell them apart?

Ekaterina Lapshinova-Koltunski, Sylvia Jaki, Maren Bolz, Merle Sauter
University of Hildesheim
Correspondence: lapshinovakoltun@uni-hildesheim.de, sylvia.jaki@uni-hildesheim.de

Abstract

This contribution investigates whether machine-translated subtitles can be easily distinguished from human-translated ones. For this, we run an experiment using two versions of German subtitles for an English television series: (1) produced manually by professional subtitlers, and (2) translated automatically with a Large Language Model (LLM), i.e., GPT4. Our participants were students of translation studies with varying experience in subtitling and the use of machine translation. We asked participants to guess if the subtitles for a selection of video clips had been translated manually or automatically. Apart from analysing whether machine-translated subtitles are distinguishable from human-translated ones, we also seek for indicators of the differences between human and machine translations. Our results show that although it is overall hard to differentiate between human and machine translations, there are some differences. Notably, the more experience the humans have with translation and subtitling, the more able they are to tell apart the two translation variants.

1 Introduction

Although Machine Translation (MT) has arrived in audiovisual translation somewhat later than in some other fields of translation, it has in fact come to play an important role in various translation forms such as subtitling, dubbing, etc. Idiomatic and enjoyable target texts are particularly crucial when it comes to the entertainment values that are typically associated with those types of translation, which is why there is skepticism among audiovisual translators concerning the quality of MT in this field (e.g. Jaki et al., 2024). On the other hand, the quality of MT has increased considerably over the last years, and it is common practice to use post-edited MT (MTPE), especially within the field of subtitling.

The question has therefore arisen whether MT subtitles are still recognisable as such. For this contribution, we analysed linguistic differences based on automatic annotation, as well as overlaps in words. This step involves a comparison of two translation variants using quantitative information on linguistic features. In addition, we asked human evaluators to recognise the method (manual or automatic) with which the subtitles at hand were produced, building on the results of Calvo-Ferrer (2023). For this step, students were asked to identify human and MT subtitles for an English TV series. Apart from the visible surface differences between the two translation variants measured by either linguistic information or human judgement, we are interested in further influencing factors, such as the quality of the subtitles or the test persons' level of expertise. For instance, it is interesting to know if dedicated instruction in subtitling increases the ability to recognise machine-translated subtitles and if other competences may play a role.

Thus, for our study, we formulate three research questions (RQs):

RQ1 Are there any differences between human and machine translation variants of the same subtitles?

RQ2 Does the quality play a role in the differentiation between human and machine-translated subtitles?

RQ3 Does the level of expertise play a role in the ability to tell apart human and machine translation?

In this study, we address the language pair English-German. Although both English and German are high-resource languages with much training data and existing MT solutions performing better than for other language pairs, we still believe

Proceedings of Machine Translation Summit XX Volume 1, pages 496–505, June 23-27, 2025

that looking into this language pair is important. The results of our study are particularly valuable for higher education institutions that train English-German subtitlers, since the information on the differences between MT and human subtitle translation is a great asset for this context.

The remainder of this paper is organised as follows: In Section 2, we give an overview of related work. Section 3 describes the data as well as the methodological design of this study. The results are presented in Section 4, which is organised along the RQs. We summarise the results as well as the limitations of this study, and we provide an outlook for future work in Section 5.

2 Related Work

2.1 MT technology for subtitling

Etchegoyhen et al. (2014)'s seminal work in the project SUMAT has marked a common strand of research in the automatic translation of subtitles that focuses on leveraging the quality of MT for subtitling, in part with feedback from professional subtitlers.

Over the time, neural machine translation (NMT) has taken the stand in the language industry as well as in research trying to boost these systems. Hiraoka and Yamada (2019), for example, obtained positive results for the translation pair Japanese-English by working with a set of pre-editing rules. Likewise, context has been increasingly considered in the improvement of MT systems. While Matusov et al. (2019) obtained positive results by including inter-sentence context, Vincent et al. (2024), in contrast, focused on including extra-textual information such as meta data into the MT model, working with MTCue (a multi-encoder transformer for contextual NMT). Their results imply that contextual data can improve the quality of MT for subtitles. Other researchers have chosen to use visual information to boost NMT performance, for example, Li et al. (2023) who successfully introduced SAFA, a new model for video-guided MT. As the focus of this study is not, however, a technological one, the remainder of the literature overview will go into more detail about the translation product, as well as the production and use of machine-translated subtitles.

2.2 Product-oriented studies

Hagström and Pedersen (2022) present a more product-oriented analysis of subtitles quality. They demonstrate a lower quality of subtitles since the 2020s, which they attribute to the increased use of MT. Other authors of product-oriented studies, in contrast, emphasise the general good quality of machine-translated subtitles, such as (Bellés-Calvera and Caro Quintana, 2021) for the English translation of the Spanish series *Cable Girls*. Martínez and Vela (2016) carry out an analysis of the quality in human- and machine-translated subtitles. They point out that although manual error analysis is time-consuming, it still provides interesting insights into the nature of human and machine translation in subtitling.

2.3 MT and subtitlers

Karakanta et al. (2022) focus on the subtitler's perspective and how MT influences their productivity. In this context, they test automatic subtitling (with MT as a part of automatic subtitling) with professional subtitlers and conclude that the subtitlers' post-editing experiences were "neutral to positive" (Karakanta et al., 2022, 9). Koponen et al. (2020) analyse the subtitling process in comparison between MT and HT and find that MTPE generally required fewer keystrokes that HT, but that there were considerable differences when it comes to language pairs, which emphasises the need for comprehensive research for a large variety of language pairs. Xie (2023)'s study of subtitler's effort in MTPE as part of automatic subtitling for the language pair English-Chinese concentrates particularly on the difference between videos with much information coming from the image in contrast to videos where most of the information stems from the verbal input. The author concludes that both require approximately the same time for MTPE, but that "the subtitlers spent more effort on revising spotting and segmentation than translation when they post-edited texts with more non-verbal information", and adds that MTPE was seen rather positively by the test persons (Xie, 2023, 63).

2.4 MT and end users

Other authors have focused on the end user's experience. For instance, Schierl (2023) shows in an analysis of Finnish and German subtitles that human translation in subtitles outperforms MTPE subtitles in terms of perceived quality, but that this does not mean that the end users need more time for reading MTPE subtitles (Schierl, 2023, 50). Calvo-Ferrer (2023) performs an experiment on the detectability of machine-translated subtitles for

the language pair English-Spanish. The approach is interesting as it combines a kind of Turing test with MT evaluation research. However, the experiment does not strictly address end users, as the test persons were 119 students of a translation study program. They were provided with eight clips with humorous content and were asked to classify those either as MT or HT. The results suggest that machine-translated subtitles have become difficult to identify. They also show that experience with translation seems to be a decisive factor: The fourth year students outperformed their fellow first year students in this classification task. The study also indicates that clips with poor subtitling quality are more frequently attributed to MT, and those of better quality to HT.

Our study directly builds on the results in Calvo-Ferrer (2023). We aim to find out whether we can find similar tendencies for the language pair English-German and if translation experience also plays a role. Whilst our experiment is designed to be comparable to the previous results by Calvo-Ferrer (2023) in the questions addressed, we also add linguistic analysis of the differences between human and machine translations, as well as the direct comparison of the outputs using the BLEU score (Papineni et al., 2002). Also, the data at hand differs from the data used in the previous research.

3 Methodology

3.1 Data

Subtitles For the experiment, we used freely available data provided on the homepage of IWLST (International Conference on Spoken Language Translation) for shared tasks on automatic subtitling (https://iwslt.org/2024/subtitling). IWLST obtained the data with the kind permission of ITV Studios, which has 60 labels in twelve countries and includes UK's largest commercial broadcaster (https://www.itvstudios.com/). The data set contains seven episodes of three different television series, with an approximate duration of seven hours in total, as well as their subtitles in English, German, and Spanish. To restrict the material, we selected seven clips that contained cultural references, puns, idioms, jargon-specific vocabulary, colloquial terms and elements of orality. In addition, we only chose one of the series and only scenes where the subtitles followed the subtitling guidelines provided by IWLST, therefore eliminating clips with subtitles up to three lines. Due to reasons of feasibility (as surveys need to be strictly limited in time, among other things to avoid fatigue effects), the material was again narrowed down to seven scenes, each with four to eleven subtitles in the German HT.

Automatic translation To produces machine-translated alternatives to the provided German subtitles, we used generative AI. More specifically, we performed tests with different models and on different web services: ChatGPT-4o mini on the Open AI web service[1], GPT-4o on a local university web service[2], as well as Meta LLaMA 3.1 8B Instruct[3] and GPT-o4 mini on ChatAI web service(Doosthosseini et al., 2024)[4]. In the end, we used the output of the best resulting translation as assessed by the authors of this paper, who have a background in linguistics, translation studies, and subtitling. Note that the outputs were not systematically compared with scores. Instead they were manually checked and the main attention was paid to the formal requirements for the subtitles as well as to linguistic accuracy. In our case, the best output was delivered with ChatGPT-4o mini. The results for this system were obtained by several prompts in German that we provide in Table 1, translated into English.

Prompt 2 was used to improve the result obtained from Prompt 1. For the human translation, we considered the subtitles provided by ITV as a gold standard, as we are dealing with human subtitles produced for a highly experienced broadcaster with a global outreach. None of the subtitles underwent any form of post-editing before the experiment, in order to avoid data manipulation. The subtitles were displayed to the test persons within the video clips, i.e., in their multimodal context.

Automatic annotation The collected human and machine translations were automatically annotated with parts-of-speech tags and syntactic functions with the help of the dependency parser using the Stanford NLP Python Library Stanza (v1.2.1)[5] with all the models pre-trained on the Universal Dependencies v2.5 datasets. We collected occurrence distribution of automatically tagged parts-of-speech (based on universal part-of-speech tags or UPOS) and selected syntactic functions that are assigned to

[1] https://chatgpt.com/
[2] Anonymised URL
[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[4] https://docs.hpc.gwdg.de/services/chat-ai/index.html
[5] https://stanfordnlp.github.io/stanza/index.html

System prompt: Your are a subtitler. For the translation of the English subtitles that I will be providing, please use the following rules for the output: The in- and out cues from the input as well as the time stamps are maintained like in a template; therefore they are not supposed to be changed. There is a maximum of 17 characters per second (including blanks). Each subtitle has a maximum of two lines (like in the input). There is a maximum of 42 characters per line (including blanks). The subtitles are produced for an American TV series from the Genre thriller or crime drama.
Prompt 1: Please translate the subtitles from English to German. Please stick to the rules indicated above.
Prompt 2: Please adjust the subtitles so that there are really only two lines per subtitle. Make sure they sound more colloquial and natural, like a conversation among colleagues. Don't forget to stick to the rules indicated above.

Table 1: Prompts used to translate subtitles into German.

the nominal category in the Universal Dependency classes (UD)[6], see de Marneffe et al. (2021) for more details. The occurrence of these categories was then compared between the two variants of translations.

3.2 Survey design

Building on the results by Calvo-Ferrer (2023), the survey was conducted among students at the University of Hildesheim in Germany. All the test persons were students of translation programs, i.e., they all had a very good command of English and native or near-native knowledge of German. One group was composed of 24 BA students. We assumed that they were not familiar with the art of subtitling yet. The other group consisted of 30 MA students that have already undergone instruction on subtitling, the hypothesis being that the students more experienced with subtitling may have less difficulty distinguishing between the machine-translated and the human subtitles. In order to control for the level of experience, both with MT and subtitling, we asked them how long they were studying, whether they had experience with subtitling and how often they were working with MT. We also asked students for their proficiency in English and German, mainly to understand potentially why grammatical mistakes or the like may have been overlooked in the subtitles. Please note that not all students finished the survey; fragmentary questionnaires were excluded from analysis. Consequently, the corpus of analysis consists of 21 answers from BA students and 25 from MA students.

The survey was implemented via Lime Survey[7], which allows for an exportation of data in excel format, for example. We also used the automatic shuffling function of Lime Survey to make sure that each participant would be confronted with seven video clips, with either machine-translated or human subtitles, respectively. For each of the clips, the participants had to indicate whether they thought they were dealing with human or machine-translated subtitles, how sure they were about their assessment in the respective cases, and how good they judged the quality of the subtitles to be. They were also provided with an open question for each of the clips where they had to indicate what their decision (MT vs. HT) was based on.

4 Results

4.1 RQ1: Differences between human and machine translation

We start with the first research question which concerns the differences between human and machine translations. To answer this question, we looked into both the frequency distributions of linguistic features in the two translation variants and into the human judgements collected in the survey.

Linguistic difference In terms of linguistic features, we analysed morpho-syntactic properties of the texts derived from the automatic analyses described in Section 3.1 above. We counted distributions of parts-of-speech (POS) and syntactic functions.

Figure 1 demonstrates the distributions of adjectives (ADJ), adpositions (ADP), adverbs (ADV), auxiliaries (AUX), connectives (CCONJ) and subjuncts (SCONJ), determiners (DET), common and proper nouns (NOUN, PROPN), pronouns (PRON),

[6] https://universaldependencies.org/u/dep/

[7] https://www.limesurvey.org

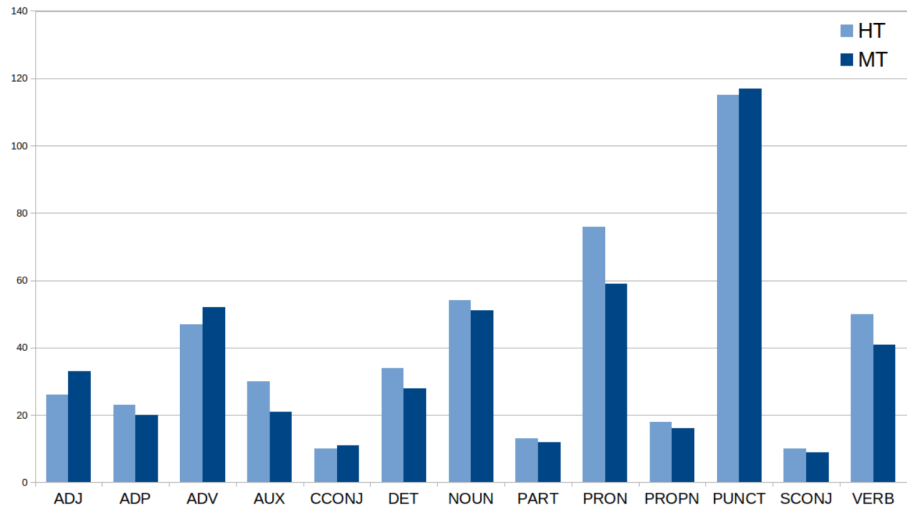

Figure 1: Distributions of parts-of-speech in human and machine translations.

verbs (VERB), particles (PART) and punctuation (PUNCT). The barplots reveal a number of differences in the distributions: While human translations contain more nouns, pronouns and verbs, machine-translated texts contain more adjectives and adverbs. However, the overall difference is not significant as confirmed by Pearson's chi-square test (p-value of 0.94).

We observe a similar tendency in terms of the distributions of the selected syntactic functions. They include nominal subjects (nsubj), direct objects (obj), indirect objects (obl), nominal modifiers (amod and nmod summarised as a-nmod in the figure), nominal modifiers functioning as appositions (appos), as well as adverbal modifiers (advmod), see Figure 2. It is obvious from the figure that the distributions of the categories are similar in both translation variants, with HT utilising more of those constructions. The most prominent difference is observed for the distribution of subjects, which prevail in human translations. However, the overall difference is not significant (p-value of 0.79).

Human judgements We proceed with the analysis of the survey results to see if students were able to recognise if the subtitles were translated manually or automatically. Table 2 represents the confusion matrix based on human judgements. The overall accuracy is relatively low (0.5). While human translations were recognised with 47.88% of precision, machine-translated texts seem to be slightly better identifiable - their recognition precision constitutes 51.59%. However, MTs have a lower true positive rate than human translations (0.49 vs. 0.51), which means that they were more frequently labeled as HTs.

true	HT	79	76
	MT	86	81
		HT	MT
		predicted	

Table 2: Confusion matrix: classification as HT and MT by test persons.

Table 3 illustrates the amount of correct judgements by BA and MA students. In general, the recognition rates were relatively low, and varied considerably between the different test items.

	BA		**MA**	
Clip	**total**	**in %**	**total**	**in %**
1	14	67	15	60
2	10	48	9	36
3	12	57	11	44
4	14	67	15	60
5	5	24	12	48
6	8	38	12	48
7	10	48	16	64

Table 3: Recognition rate per study degree.

For the BA students, MT was correctly recognised in 33 cases; HT was recognised correctly in 38 cases. MT was misinterpreted as HT 39 times, and HT as MT 37 times. The results differed considerably between the seven test items. For example, with two items, the subtitles were correctly classified as MT only once, respectively, while it was correctly classified as such 11 times with another test item. For the MA students, MT

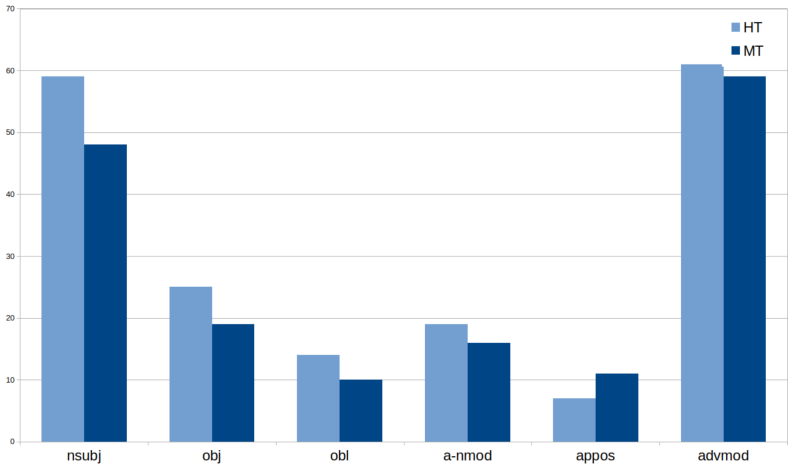

Figure 2: Distributions of syntactic functions in human and machine translations.

was recognised as such in 48 cases, HT in 41 cases. MT was erroneously identified as HT 47 times, and HT erroneously as MT 39 times.

4.2 RQ2: Role of quality

Next, we analysed if the quality of machine translation impacts the recognition rate. We also analysed if humans judge the quality of human and machine translations in a similar way.

MT quality As we were particularly interested in differences between human and machine translations and in their indicators, quality evaluation of MT is not the focus of this study. However, we calculated the automatic evaluation scores to get a general idea of their performance. Moreover, some scores, as e.g. BLEU score (Papineni et al., 2002), also provides information on the overlaps between human and machine translations. So, we used three metrics that can be calculated with the tools provided in MATEO (Vanroy et al., 2023), i.e., BLEU, ChrF (Popović, 2015), and TER (Snover et al., 2006). The numbers are reported in Table 5. All the scores point to dissimilarities between the two translation variants, as both BLEU and ChrF count the overlaps in ngrams between HT and MT (with ChrF taking into account also word order differences) and TER the edits needed fo MT to be overlapping with HT. This means that machine-translated texts in our data differ considerably from human translations in terms of word choices. In Table 4, we demonstrate an example from the data marking overlapping words in bold. As seen from this example, there is not much overlap in word choices between human and machine translations. At the same time, syntactic constructions, e.g. imperative in lines 6 and 7, seem to be similar, which coincides with our result on the linguistic differences measured with parts-of-speech and syntactic function distributions.

Using the data from the judgements by humans, we analysed if the BLEU score correlates with the misclassification cases, i.e., how many students labeled its machine-translated version as a human translation. As seen in Figure 3, we observe a negative correlation, which means that the quality of MT (at least the automatically evaluated quality) does not impact our test persons' decision and even the texts with lower scores can be identified as human translations.

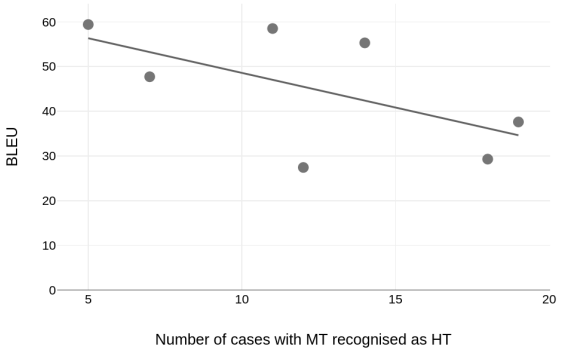

Figure 3: Correlation between human judgements and the quality of machine translation measured with BLEU.

Human quality judgements Participants were also asked to estimate the quality of the translation by labeling the test items with *very good*, *rather good*, *moderate*, *rather poor*, and *very poor*. An overview of the estimation per human and machine translations is given in Table 6. We indicate the percentage of answers normalised against the total

	human	machine
1	You had no idea.	
	Und ihr hattet keine Ahnung **?**	Hättest du nicht gedacht, oder **?**
2	Medal of Valor, Internal Affairs, cameras.	
	Tapferkeitsmedaille, interne Ermittlung, **Kameras**.	Medaille, Interne, **Kameras**... alles dabei.
3	It all seems a tad orchestrated, don't you think?	
	Scheint mir **ziemlich** viel Brimborium zu sein, findest du nicht?	Klingt alles irgendwie **ziemlich** inszeniert, oder?
4	If you think I had anything to do	
	Wenn **du** mir das anhängen willst,	Denkst **du**, ich hab' was damit zu tun,
5	with that, we can just step outside.	
	können **wir** gleich vor die Tür gehen.	dann klären **wir**'s draußen, okay?
6	Relax, Cut. This is not a John Wayne movie.	
	Krieg **dich** wieder ein, **Cut. Das ist kein** John-Wayne-Film.	Beruhig **dich, Cut. Das ist kein** Western.
7	Look at this. Everybody's doing the funny.	
	Sieh **dir das an.** Hier spielt **jeder** den Clown.	Schau **dir das an.** Jetzt macht **jeder** Witze.

Table 4: Example from the data: human (left) vs. machine (right) translation of Clip 1.

Clip	BLEU	ChrF	TER
1	15.7	27.4	84
2	9.7	29.3	75
3	13.6	37.6	75
4	27.6	47.7	62.5
5	32.9	55.3	60.6
6	31.5	58.5	53.7
7	26.8	59.4	65.5
Avg	**22.54**	**45.03**	**68.04**

Table 5: BLEU score per test item.

number of answers for MT and HT separately.

The test persons tended to rate the quality of the translations rather positively than negatively, but indicated a broad range of judgements: 44 times *very good* (13.84%), 135 *rather good* (42.45%), 97 times *moderate* (30.5%), 39 times *rather poor* (12.26%), and three times *very poor* (0.94%). Overall, both the machine-translated and manually produced outputs were rated similarly, with the only noticeable difference being a 6-per-cent higher rating for the human translations for the label *very good* (see Table 6). Nor do the results suggest that participants automatically associated those translations that they qualified as less good to be MTs, or those that they judged to be good to be HTs. This tendency can only be observed for Test Item 6, where the nine translations labeled as rather poor machine translation were all, in fact, human translations. Interestingly, out of the twelve HTs labeled with *very good*, eight were, in turn, MTs.

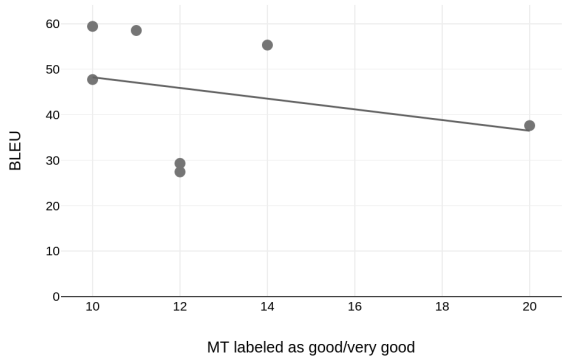

Figure 4: Correlation between human judgements and the quality of machine translation measured with BLEU.

We also analyse correlation between the BLEU score and the judgements by students. The latter is operationalised as the number of *good* and *very good* labels per MT version of the given video clip. As seen in Figure 4, the BLEU score [8] does not correlate with human judgements in our data, which again confirms the observation on RQ1 above.

4.3 RQ3: Role of the level of expertise

The level of expertise can be measured according to various criteria. All of the MA students in the experiment had had prior experience with the art

[8] We also tested correlation with ChrF and observed the same result as for BLEU.

MT vs. HT	Very good	Rather good	moderate	rather poor	very poor
MT	10.98	43.29	32.93	12.20	0.61
HT	16.56	40.76	28.03	13.38	1.27

Table 6: Ratings of the translations in per cent

of subtitling. At the same time, we included the experience with machine translation as a possible impacting factor too.

Figure 5 illustrates the number of correct judgements grouped by the study degree.

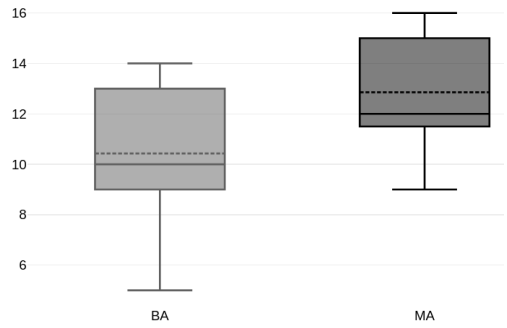

Figure 5: Correct judgements grouped by BA and MA study degree.

Overall, it was easier for the MA students to differentiate between human and machine translation, although the difference is not big.

To control for the **degree of familiarity with MT**, we asked students whether they worked with MT *very often*, *often*, *sometimes*, *rarely*, or *never*, with the hypothesis that it might be easier for students experienced with MT to distinguish between translations produced manually or automatically. The comparison between students who indicated that they worked with MT (1) *very often* or *often* and those who indicated (2) *sometimes* or *rarely* (*never* did not occur) did not produce any significant results: While there were 54 correct and 51 incorrect judgements for (1), it was 105 vs. 105 for (2). The amount of incorrect judgements was slightly higher when we singled out only those students who indicated that they rarely work with MT (with 27 correct and 29 incorrect answers). Therefore, it is fair to say that with the selection of students who are more experienced in MT, the amount of correctly identified translation variants is higher than the incorrect judgements. In contrast, the amount of correct judgements is lower than the amount of incorrect ones for the students who are not used to working with MT. However, the difference is only minor.

When it comes to the **level of confidence** in their answers, the difference between the correct and the incorrect answers is barely noticeable (possible answers: *very confident*, *rather confident*, *rather unsure*, *pretty unsure*): 19.02% of the correct and 14.47 % of the incorrect judgements were accompanied with *very confident*, 45.5% of the correct and 46.54% of the incorrect ones with *rather confident*, 30.06% of the correct and 32.08% of incorrect ones with *rather unsure*, and 5.52% of the correct and 6.96% the incorrect ones with *pretty unsure*.

5 Conclusion and Discussion

One aim of the present study was to see if machine-translated subtitles differ from human-translated ones. We used a number of analyses, including corpus-based frequency distribution of linguistic features, automatic quality scores, as well as human judgements. The overall results show that it is hard to differentiate between manually and automatically translated subtitles. Moreover, both translation variants seem to be similar in terms of the distribution of linguistic features such as parts-of-speech and syntactic functions. This points to structural similarities between the two outputs.

The main differences observed include word choice as indicated by the low BLEU score for machine translations, which implies that there are not so many n-gram overlaps between the two translation variants. Besides that, we showed that the BLEU score did not correlate with human judgements either, as texts with a lower BLEU score were more frequently labelled as human translations. Also, the calculated BLEU score does not necessarily reflect subtitle quality as it is perceived by humans, as our test persons classified the quality of machine translations as good and acceptable frequently, sometimes even more frequently than with the human-translated variants.

At the same time, it was interesting to see that the level of expertise measured by the advance in study program does play a role in the ability to correctly differentiate between human and machine translations of subtitles.

However, we are also aware of the limitations of this study. First of all, the number of the data that we included into the study (and also survey) is limited to seven texts (clips) only. This restriction was due to the requirements of the given settings: To avoid fatigue effects (which could have impacted the results), we decided beforehand that the survey time should be restricted to a maximum of 30 minutes. Given that watching the clips, making decisions and answering the questions takes a considerable amount of time, we could not collect data for more than the seven clips at hand.

We plan to extend the data to more clips. Although it is challenging to perform a survey with more texts, we would be able to perform a more extensive quantitative analysis of the linguistic differences between human and machine translations including automatic text classification.

Another drawback of this study is testing translation outputs with an LLM only. More machine-translated outputs, also those produced with traditional MT systems and with other LLMs than GPT is part of our future work. However, we are also aware of the problems of reproducibility, as the future results that build upon our findings may differ from those reported by us, as LLMs are regularly updated and are changing. Another problem of such systems is that we do not have any control over their training data. The dataset used for testing (the selected clips) is probably included into the training data of the LLMs at had, as the dataset is open source and freely available. Producing subtitle translation specifically for the survey would be a better scenario.

Besides that, this study does not provide a deep analysis of the subtitle quality. Although we mention some issues, we do not report on accuracy and other factors. Moreover, pragmatic factors such as transfer of emotions, sentiment, humour, etc. cannot be considered with the methodology applied. However, this can be analysed on the basis of our data in future work.

In future, we would also like to extend the test persons to more experienced groups and include professionals from the subtitling industry.

References

Lucía Bellés-Calvera and Rocío Caro Quintana. 2021. Audiovisual translation through NMT and subtitling in the netflix series 'cable girls'. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 142–148, Held Online. INCOMA Ltd.

José Ramón Calvo-Ferrer. 2023. Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives*, 32(6):1115–1132.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Ali Doosthosseini, Jonathan Decker, Hendrik Nolte, and Julian M. Kunkel. 2024. Chat AI: A Seamless Slurm-Native Solution for HPC-Based Services. *Preprint*, arXiv:2407.00110.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 46–53, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hanna Hagström and Jan Pedersen. 2022. Subtitles in the 2020s: The influence of machine translation. *Journal of Audiovisual Translation*, 5(1):207–225.

Yusuke Hiraoka and Masaru Yamada. 2019. Pre-editing plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 64–72, Dublin, Ireland. European Association for Machine Translation.

Sylvia Jaki, Maren Bolz, and Röther Sofie. 2024. KI-Technologien in der Audiovisuellen Translation. *trans-kom*, 17:320–342.

Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 261–270, Ghent, Belgium. European Association for Machine Translation.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal. European Association for Machine Translation.

Yihang Li, Shuichiro Shimizu, Chenhui Chu, Sadao Kurohashi, and Wei Li. 2023. Video-helpful multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4281–4299, Singapore. Association for Computational Linguistics.

José Manuel Martínez Martínez and Mihaela Vela. 2016. SubCo: A learner translation corpus of human and machine subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2246–2254, Portorož, Slovenia. European Language Resources Association (ELRA).

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Frederike Schierl. 2023. Reception of machine-translated and human-translated subtitles – a case study. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 42–53, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine translation evaluation online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland. European Association for Machine Translation.

Sebastian Vincent, Charlotte Prescott, Chris Bayliss, Chris Oakley, and Carolina Scarton. 2024. A case study on contextual machine translation in a professional scenario of subtitling. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 561–572, Sheffield, UK. European Association for Machine Translation (EAMT).

Bina Xie. 2023. Machine translation implementation in automatic subtitling from a subtitlers' perspective. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 54–64, Macau SAR, China. Asia-Pacific Association for Machine Translation.