Audiovisual Translation through NMT and Subtitling in the Netflix Series *Cable Girls*

Lucía Bellés-Calvera[1][0000-0002-1329-6395] and Rocío Caro Quintana[2][0000-0003-2275-2679]

[1] Universitat Jaume I
lucia.belles@uji.es
[2] University of Wolverhampton
R.Caro@wlv.ac.uk

Abstract. In recent years, the emergence of streaming platforms such as Netflix, HBO or Amazon Prime Video has reshaped the field of entertainment [1], which increasingly relies on subtitling, dubbing or voice-over modes [2] [3]. However, little is known about audiovisual translation when dealing with Neural Machine Translation (NMT) engines. This work-in-progress paper seeks to examine the English subtitles of the first episode of the popular Spanish Netflix series *Cable Girls* and the translated version generated by Google Translate and DeepL. Such analysis will help us determine whether there are significant linguistic differences that could lead to miscomprehension or cultural shocks. To this end, the corpus compiled consists of the Spanish script, the English subtitles available on Netflix and the translated version of the script. As regards data analysis, errors have been classified following the DQF/MQM Error typology and have been evaluated with the automatic BLEU metric. Results show that NMT engines offer good-quality translations, which in turn may benefit translators working with audiovisual entertainment resources.

Keywords: Audiovisual translation, Neural Machine Translation (NMT), Errors.

1 Introduction

Over the past few years, the rise of Netflix, HBO, Amazon Prime Video and other streaming platforms has made it necessary to rethink entertainment media [1]. Accessibility to their catalogues not only offers the audience the opportunity to choose among a variety of films, series, documentaries and other audiovisual resources but also to make use of subtitling and dubbing options [2, 3]. However, even though these audiovisual translation practices are meant to meet the needs of different markets and users [4], the quality of the translation may be affected by errors when translators are given tight deadlines. Machine Translation (MT) is widely used in the translation industry, especially in technical fields because the texts tend to be repetitive, and studies have shown that it increases translators' productivity [5, 6] by post-editing the MT output. However, despite the fact that platforms like Netflix announced that they are using MTPE in their subtitling workflows three years ago, research on this topic is

Translation and Interpreting Technology Online, pages 142–148,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_015

still scarce in creative fields, such as literary or audiovisual texts. It might be assumed that NMT will not work when dealing with Audiovisual Translation due to its time and character constraints, and especially in media entertainment where cultural aspects are prevalent. Given that the quality of NMT, although still not at the human level [7, 8], is improving every day, some issues need to be considered. Would it be beneficial for audiovisual translators to use MT and post-edit the texts? Or is scratch translation still the best solution for this field?

In any commercial deployment of MT in a subtitling workflow, a bespoke engine would be used. In fact, there are already subtitling specialised MT systems available in the market like AppTek, Omniscien and XL8. However, growing volumes of audiovisual content, short turnaround times or lack of access to this type of engines are some of the challenges novice translators need to overcome. These issues have been addressed in previous studies where MT may serve as a possible solution [9] [10]. Numerous publications arised from the SUMAT project, a large-scale EU-funded project that inspected the creation of high-quality parallel corpora of subtitles through MT [10] [11]. Matusov et al. [12], for example, analysed improvement in productivity after integrating MT in audiovisual translation.

This ongoing project aims to ascertain the quality of Google Translate and DeepL translations (i.e. open MT resources) when compared to the subtitling of TV series in the source language. On this account, the current study draws from the following research questions: RQ(1) *How do English subtitling and translations from NMT differ from the source text in Spanish? What types of errors can be found?* and RQ(2) *Does the integration of MT on the audiovisual translation workflow benefit translators?*

The section below delves into the methodological procedure followed in this study. Later on, the discussion of the preliminary results as well as the conclusions and next steps of this ongoing project will be provided.

2 Methodology

The corpus under study revolves around the Spanish Netflix original series called Cable Girls. This drama, premiered in 2017 and set in the 1920s, tells the story of four women working as operators for the National Telephone Company at a time of social changes. The first season consists of eight episodes, with a length of 47 to 64 minutes.

For the compilation of this small corpus, the focus has been on the first 10 minutes of the first episode of the first season released in Spanish. The Spanish script and the official English subtitles incorporated in the streaming platform have been transcribed from Netflix [13] and examined for the purpose of this preliminary study. In addition, the Spanish transcript was translated with Google Translate [14] and DeepL [15] to analyse the quality of these NMT engines.

Google Translate is an MT engine that provides the translation of texts and files into more than 100 languages, including English, Spanish, Greek, Belarusian, Afrikáans or Chinese [14]. The fact that Google offers these services has caught the attention of scholars who have been concerned with error analysis on MT output.

Evidence may be found in Trzaskawka's study [16], which explored the accuracy of this tool in the translation of contracts in English and Polish. Issues related to the quality of the translation output have also been explored in specialised areas, such as literature [17] and scientific writing [18]. However, research on entertainment media seems to be scarce, with studies delving into the dubbing and subtitling of TV series [19] and documentaries [1].

DeepL Translator [15] is an NMT software developed in 2016 with the aim of producing high-quality translated texts. At the moment, DeepL works with more than 20 languages and also offers a formal/informal register for their translations. DeepL has also caught the attention of researchers and several studies compare its quality to other MT engines like Google Translate, Yandex or Microsoft Translator [20, 21, 22].

The quality of the machine-translated texts has been assessed manually following the DQF/MQM Error Typology [23] – the integration of DQF (Dynamic Quality Framework) [24] and MQM (Multidimensional Quality Metrics) [25] – paying attention to the categories labelled as Accuracy and Fluency. For this manual evaluation, 153 segments containing 7 words on average were examined by two annotators with experience in translation (i.e. post-editing) and linguistics. The translated texts were then analysed automatically with the BLEU metric [26], using the original subtitles as the human translation and the NMT output from Google Translate and DeepL.

3 Evaluation: Preliminary results

3.1 Manual evaluation

A total number of 153 segments were analysed manually following the DQF/MQM Error Typology. The most common errors were related to Fluency, Accuracy and Style. The distribution of errors in Google Translate and DeepL are presented in Table 1.

For this ongoing study on audiovisual translation, namely in subtitling, the character constraint – which entails 70 characters distributed in two lines and a maximum on-screen duration of 6 seconds, has not been analysed on the grounds that Google Translate and DeepL are not specialised systems in subtitling. Instead, the focus has been on the quality of the translation. Therefore, as noted in Table 1 above, the manual analysis of the output taken from both engines differs to a great extent. The findings reveal that only 15 errors have been identified in DeepL (10%), as opposed to Google Translate, where meaning was not properly conveyed in 41 segments (27%).

Table 1. Distribution of errors

Category	Number of errors		Sub-category
	Google Translate	DeepL	
Fluency	20	4	Grammar Grammatical register Inconsistency
Accuracy	14	10	Mistranslation Addition Over-translation
Style	5	1	Unidiomatic Awkward
Other	2	0	Culture-specific reference Tone
TOTAL	**41**	**15**	

Most errors in both engines have to do with Fluency and Accuracy. The number of fluency errors is higher in Google Translate, with a total of 20, and only 4 out of 15 in DeepL. Some examples of fluency errors can be seen in Table 2.

Regarding Accuracy errors, DeepL seems to perform better than Google Translate. Only 10 accuracy errors were spotted in DeepL, while these amount to 14 in Google Translate. Some accuracy errors are illustrated in Table 3.

Table 2. Fluency errors

Original	**English Translation**	**Error**
¡Corre!	Come on!	**Runs**! (Google Translate)
Como grites, te juro que te mato.	If you shout, I swear I'll kill you.	**As** you scream, I swear I will kill you (Google Translate)
Pues lo lamento, no se encuentra entre las preseleccionadas.	I'm sorry, you're not on the short list.	Well, I'm sorry, **she**'s not among the shortlisted. (DeepL)

Table 3. Accuracy errors

Original	**English Translation**	**Error**
Tú no te metas. ¡No te metas!	You stay out of this! Stay out of this!	**You do not mess. Do not mess!** (Google Translate)
600 km para poder estar aquí ahora. / -550.	Six hundred kilometers to get here. / Five hundred and fifty.	600 km to be here now. / 550. **550**. (DeepL)
A continuación, tenemos dos plantas para las salas de máquinas.	Next, two floors with the machine rooms.	Next we have two **plants** for the engine rooms. (Google Translate)

These findings suggest that efforts should still be devoted to refine Fluency and Accuracy in MT engines, as they still not work at the human level. In order to improve the quality of NMT outputs, more corpora should be processed.

3.2 Automatic evaluation

The quality of the texts was evaluated with the automatic metric BLEU [26] using the online BLEU score evaluator from Tilde [27]. Thus, the English subtitles employed in the Netflix platform were compared with the outputs generated by Google Translate and DeepL. The BLEU score for Google Translate is 36.44, in contrast to DeepL, which rises up to 40.79. Although these findings are not conclusive due to the size of the sample, DeepL appears to achieve better results than Google Translate when it comes to the translation of audiovisual resources.

4 Conclusions and further research

The research questions attempted to determine the quality of the translations provided by Google Translate and DeepL when dealing with audiovisual media. Hence, the Cable Girls series script in the source language was compared with the MT outputs from Google Translate and DeepL.

As to RQ(1), the findings suggest that the most common errors occur at Fluency and Accuracy levels. In addition, the results show that DeepL outperforms Google Translate in both manual and automatic evaluation.

With regard to RQ(2), the next steps of the project will delve into translators' post-editing efforts: is it useful to use MT for audiovisual texts? In this vein, technical, temporal, and cognitive variables will be considered to prove whether these efforts are higher or lower when integrating MT tools. Accordingly, an eye-tracking device and a keystroke logging tool will be employed.

Limitations in this study should be acknowledged. The small size of the corpus compiled for this preliminary study may affect the validity of the generalisations presented here. Nonetheless, it should be noted that the corpus will be expanded in the near future. Moreover, the MT engines that were used are not trained on subtitling and may contain an enormous amount of noise. DeepL and Google Translate were used to emulate the experience of freelance translators using general MT. Notwithstanding, the use of these MT engines could have a negative impact on translation quality as the length of the segments, a relevant feature in subtitling, is not taken into consideration.

Further research could also focus on other audiovisual resources, including documentaries or realities. Such examination would prove the efficiency of Google Translate in specialised and non-specialised contexts or the quality of other machine translation software like DeepL in audiovisual domains. Other lines of the proposal presented here could involve the role of MT in the translation of humour and cultural aspects, which are prolific in entertainment media.

References

1. Costan Davara, G.: Audiovisual Translation: Subtitling Netflix documentary â Black Hole Apocalypseâ. (PhD dissertation). Università degli Studi di Padova (2020).
2. Oh, K., Noh, Y.: The actual condition and improvement of audiovisual translation through analysis of subtitle in Netflix and YouTube: focusing on Korean translation. Journal of Digital Convergence 19(3), 25–35 (2021).
3. Díaz Cintas J.: Teaching and learning to subtitle in an academic environment. In Díaz Cintas J. (ed.), The Didactics of Audiovisual Translation (pp. 89-103). Amsterdam and Philadelphia: John Benjamins (2008).
4. Campbell V.: Science, Entertainment and Television Documentary. Palgrave Macmillan, London (2016).
5. Guerberof, A.: Productivity and quality in the post-editing of outputs from translation memories and machine translation. Localisation Focus. The International Journal of Localisation, 7(1), 11–21 (2009).
6. Sanchez-Torron M., Koehn P.: Machine translation quality and post-editor productivity. In: Proceedings of AMTA (p. 16-26). Association for Machine Translation in the Americas, AMTA, Austin, Texas (2016)
7. Läubli S, Sennrich R, Volk M.: Has machine translation achieved human parity? a case for document-level evaluation. arXiv preprint arXiv:1808.07048. 2018 Aug 21.
8. Toral, A., Castilho, S., Hu, K., Way, A.: Attaining the unattainable? reassessing claims of human parity in neural machine translation. arXiv preprint arXiv:1808.10432. (2018).
9. Volk, M.: The automatic translation of film subtitles: a machine translation success story? In: Nivre, J; Dahllöf, M; Megyesi, B. Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein. Uppsala, Sweden, 202-214. (2008).
10. Bywood, L., Georgakopoulou, P., Etchegoyhen, T. Embracing the Threat: Machine Translation as a Solution for Subtitling. Perspectives 25(3), 492–508 (2017). doi:10.1080/0907676X.2017.1291695.
11. Bywood, L., Georgakopoulou, P., Volk, M., Fishel, M.: What is the productivity gain in machine translation of subtitles? Paper presented at Languages & The Media, Berlin. (2012).
12. Matusov, E., Wilken, P., Georgakopoulou, Y. Customizing Neural Machine Translation for Subtitling. Proceedings of the Fourth Conference on Machine Translation (WMT), 1, pp. 82–93. Florence, Italy. Association for Computational Linguistics. 2019
13. Campos, R., Neira, G.R.: Las chicas del cable (2017). https://www.netflix.com/es/title/80100929
14. Google Translate, https://translate.google.com/about/languages/
15. DeepL Translator, https://www.deepl.com/translator
16. Trzaskawka, P.: Selected Clauses of a Copyright Contract in Polish and English in Translation by Google Translate: A Tentative Assessment of Quality. International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique 33(3), 689–705 (2020).
17. King, K.M.: Can Google Translate be taught to translate Literature? A case for humanists to collaborate in the future of machine translation. Translation Review 105(1), 76–92 (2019).
18. Suhono, S., Zuniati, M., Pratiwi, W., Hasyim, U. A. A. Clarifying Google Translate Problems Of Indonesia-English Translation Of Abstract Scientific Writing. EAI (24-25), 1–13 (2020).

19. De Nardi, I. “La casa de papel”: comparación entre el doblaje y la subtitulación al italiano del primer capítulo. (PhD dissertation). Università degli Studi di Padova (2018).
20. Rescigno A.A., Vanmassenhove, E., Monti, J., Way, A.: A Case Study of Natural Gender Phenomena in Translation A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. In: Association for Machine Translation in the Americas (AMTA): Workshop on the Impact of Machine Translation (iMpacT 2020) (pp. 62–90). Workshop on the Impact of Machine Translation at Association for Machine Translation in the Americas (AMTA).
21. Cambedda G.: A Study on Automatic Machine Translation Tools: A Comparative Error Analysis Between DeepL and Yandex for Russian-Italian Medical Translation.
22. Hidalgo-Ternero C.M.: Google Translate vs. DeepL: Analysing neural machine translation performance under the challenge of phraseological variation.
23. Harmonized DQF-MQM Error Typology https://www.taus.net/qt21-project#harmonized-error-typology
24. Dynamic Quality Framework https://www.taus.net/data-for-ai/dqf
25. Multidimensional Quality Metrics (MQM) http://www.qt21.eu/mqm-definition/definition-2015-12-30.html
26. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadeplhia, USA (2018).
27. Tilde BLEU Score Evaluator https://www.letsmt.eu/Bleu.aspx