Machine translation and audiovisual products: a case study

Adrià Martín-Mor and Pilar Sánchez-Gijón, Universitat Autònoma de Barcelona

ABSTRACT

Much has been said and written about the effects that machine translation (MT) is having on all kinds of published products. This paper discusses the introduction of MT in the localisation of audiovisual products in general and particularly voiceover documentaries. Incorporating MT into the translation of voiceover documentaries would boost the dissemination of non-commercial or minority products, and could enhance the spread of culture. While it might at first seem that MT could be easily integrated into translation for documentaries, some particular aspects make it difficult to use MT to translate for dubbing or for voice-overs. We therefore designed an exploratory study using a corpus containing different texts of a film, in order to compare the results of different automatic measures. The preliminary results show that different results are obtained for different types of speech and that the application of automatic metrics produces similar results. In this article, we will present furthermore the methodological design, which might be considered useful for other studies of this kind[1].

KEYWORDS

Machine Translation, postediting, audiovisual translation, dubbing, voice-over.

1. Introduction

The main goal of this research is to explore the possibility of incorporating machine translation (MT) and postediting into the translation workflow of audiovisual (AV) material, in particular the translation of documentaries. This kind of AV product has been chosen because the text to be translated can be easily separated from the AV product and processed through a MT system. Although a parallelism with the MT of subtitles might at first come to mind, some aspects make it difficult to use MT to translate for dubbing or voice-over. These include technical issues such as whether the script (if any) should be used as the source text, or instead the audio should be transcribed specifically for MT, how interactions among characters such as an interviewer and interviewee should be dealt with, and noise or ambient sound affecting automatic voice recognition or human transcription (and therefore the accuracy of the scripts' transcription). They also include difficulties caused by the characteristics of oral discourse (interruptions, unfinished sentences, hesitations, etc.). We therefore designed an exploratory study using a corpus containing different kinds of texts. Because the characteristics of these texts differ, different results are expected when using MT. Several automatic metrics were used to measure the MT engines' performance.

As a secondary objective of this case study, we are interested in developing a methodology to test whether MT performs differently, depending on the type of discourse (see 4, Tools and methodology).

2. Describing the genre

From the point of view of textual linguistics, discourse might be characterised under different criteria. MT outputs seem to improve when source texts have particular characteristics. MT systems (mainly ruled-based systems, but also statistical-based systems) work best with source texts produced with controlled language constraints. Since both the language model used in a rule-based machine translation (RBMT) system and the corpora used in a statistical machine translation (SMT) system usually represent a standard and neutral style, the less figurative and idiosyncratic an original text is, the better MT result is to be expected. Documentaries were therefore the best type of document for this study, since they tend to use standard language in terms of formality (although they might include technical terms) and contain different types of formal speech (dialogues, narration, etc.).

As repeatedly stated in the literature, however, documentaries have traditionally been neglected in research on audiovisual translation (AVT) (Matamala 2009: 119), perhaps because they belong to a "relatively new academic field" (Espasa 2004: 183). The term *documentary* has even been associated with negative connotations, to the extent that terms such as *non-fiction* have been proposed to avoid those connotations (Espasa 2004: 186). However, there is a fairly broad consensus among researchers that documentaries differ from other audiovisual products in that they deal with real facts, as opposed to fiction films, even if, as Matamala points out, "separating fiction and reality is not always easy and documentaries, although based upon reality, usually offer a subjective vision" (2009: 109).

Another feature of documentaries is the wide range of speakers. Matamala (2009: 115) identifies three main types of speakers based on their relationship with the person they are speaking to and the degree of spontaneity in their discourse: third-person narration, talking heads — i.e., people being interviewed, usually by the narrator —, and dialogues and spontaneous interventions. The third-person narrator is the most frequent type of speaker in documentaries. The narrator tends to be off-screen and, as Matamala puts it, "narrators do not normally pose big problems with regard to the mode and tenor of discourse, since it is usually a planned script and the language register is formal" (2009: 116). Talking heads and dialogues usually use more spontaneous language. Since their interventions do not tend to be scripted, their speeches include oral markers, such as hesitations, repetitions, false starts, etc.

Voice-overs are frequently used to translate talking heads and dialogues. Nonetheless, according to Gambier and Suomela-Salmi (1994: 243), the fact that "research has mainly been concerned with the subtitling and dubbing of fictive stories/fiction films" shows the prevalence that literary translation plays in Translation Studies. Attempts to define voice-over seem to generally encompass two main features: the presence of two voices in

two different languages and synchronisation. As Orero points out, "voice-over is viewed as the final product we hear when watching a programme where a voice in a different language than that of the original programme is heard on top of the original soundtrack" (2009: 132). Despite voice-over's "alleged disregard for synchronisation between source and target texts" (*ibid.*: 132), Orero states that "a different type of synchrony has to be respected" (Orero 2004: 82), namely with the start and the end of the original discourse and with the body language. While the latter refers to the synchrony between the discourse and the gestures of the interviewee, inserts on screen, etc., the former deals with the original soundtrack, which traditionally starts a few seconds before the voice-over and finishes a few seconds after. This procedure, as Franco states, creates an *authenticity illusion* which makes voice-over especially appropriate for documentaries:

> [T]he type of delivery we hear in voice-over translation is an important strategic way of reassuring viewers that what they are being told in their own language is what is being said in the original language, although it is known that what they will be listening to is in fact only a representation of the original discourse (Franco 2001: 292).

Orero (2009) explains that translation for voice-over can take place at one of two moments in the production workflow of audiovisual products. While it usually occurs during the post-production phase, i.e. when the product is finished, sometimes it is part of the production process, in which case it is referred to as *translation for production*. Orero says this is an "important market" (*ibid*.: 135), adding: "In the case of translation for production […], the translator has to work with rough, unedited material which will undergo several processes before being broadcast"[2].

3. Corpus description

The corpus used for the purpose of this exploratory study has been extracted from the film Fahrenheit 9/11 by Michael Moore (Moore 2004). We deemed this film to be appropriate for the purpose of our research because it is a documentary recorded in English with different types of speech and dubbed into Spanish for its official release in Spain, which indicates that the quality of the translation should be guaranteed. For this research it was crucial to select a product genuinely created in the source language and translated by humans into the target language so that the human translation could be used as a benchmark against which machine-translated versions could be compared (see section 4, Tools and methodology, below)[3]. Furthermore, it uses general vocabulary including some legal terms that can be considered part of the standard vocabulary. Various fragments of the film were manually transcribed from English and Spanish. These fragments totalled eight minutes, which was deemed enough to include various representative kinds of texts for our exploratory study. Since transcribing an oral discourse always implies somewhat subjective decisions, we decided not to transcribe oral markers such as

word repetitions and hesitations and to use full stops rather than commas, colons and semicolons where possible. These decisions were taken to meet the requirements of the MT metrics tool used for evaluation, Asia Online (see section 4, Tools and methodology, below). Although subjective, these decisions helped to achieve a more standard and neutral source text to be MT. Furthermore, the same criteria were used for both the source and the target texts in order to reduce the effects they could have on the results. Throughout the transcribed passage three types of speech were identified: narration, dialogue, and talking heads. The table below shows quantitative data about each category:

	Words	Sentences	Avg. length (words per sentence)
Dialogue			
D1	95	9	10.56
D2	133	18	7.38
D3	153	18	8.5
Total	*381*	*45*	*8.46*
Narration	398	24	16.58
Talking heads			
TH1	109	7	15.57
TH2	47	2	23.5
TH3	67	3	22.33
TH4	45	3	15
Total	*268*	*15*	*17.86*

Figure 1. Corpus description

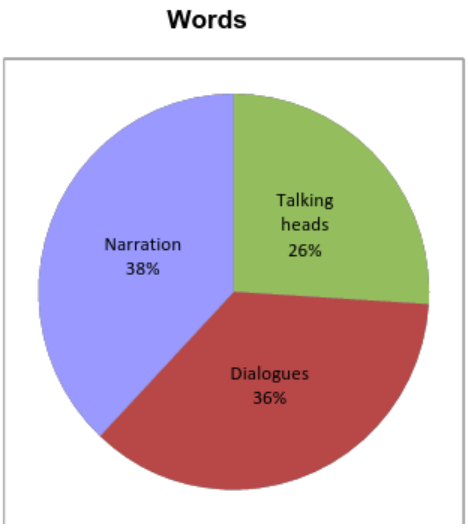

Figure 2. Distribution of the percentage of words within the sample corpus

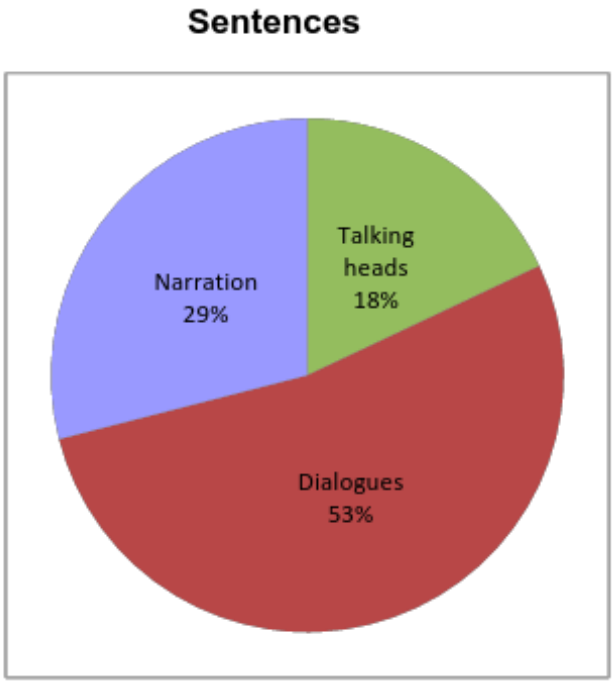

Figure 3. Distribution of the sentences within the sample corpus

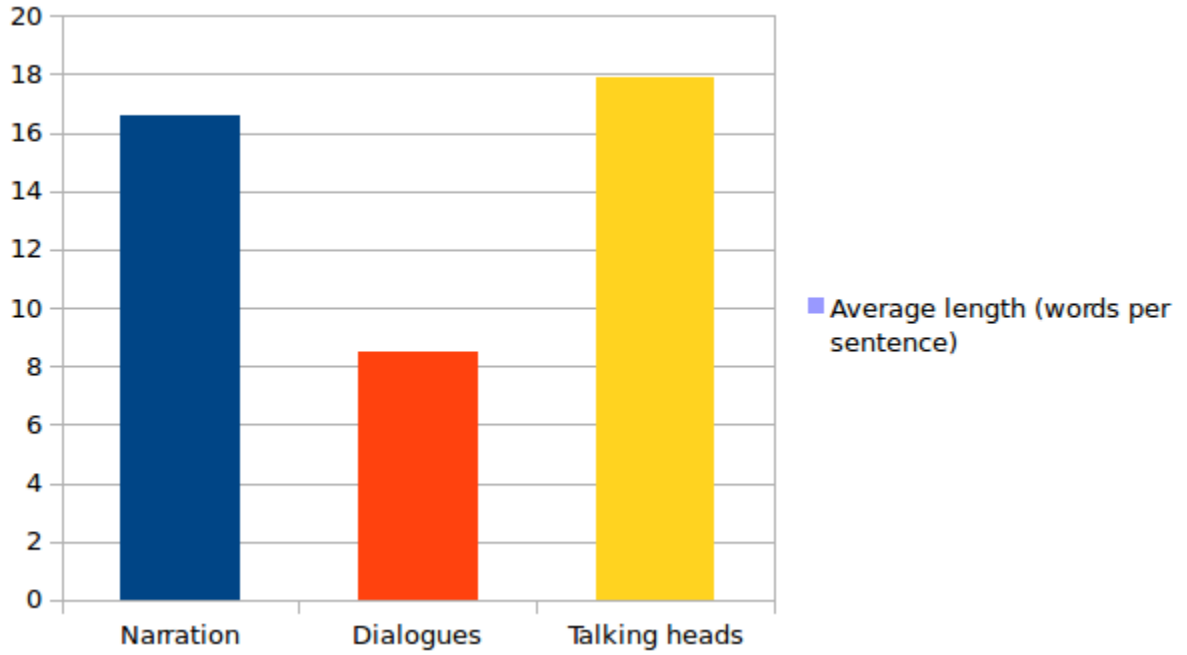

Figure 4. Average length of sentences in each subcorpora

These three categories of text have different characteristics and account for different degrees of formality. Narration tends to consist of relatively long, well-structured sentences, with subordinate clauses and coherence and cohesion markers. In other words, all the characteristics of formal, neutral texts created to be read aloud. Dialogue and talking heads are both formed of spontaneous discourse, but they vary in structure, especially sentence length. The narration passage was tagged as N, dialogues were tagged as D1, D2 and D3, and talking heads were tagged as TH1, TH2, TH3 and TH4.

Tag	Description
D1	Retired FBI agent Jack Cloonan interviewed by Michael Moore (dialogue with interviewer).
D2	Retired FBI agent Jack Cloonan interviewed by Michael Moore (dialogue with interviewer).
D3	Prince Bandar, ambassador of the Kingdom of Saudi Arabia to the United States, interviewed by Larry King (dialogue with interviewer).
TH1	James Moore, Investigative Reporter (answer to narrator's question).
TH2	James Moore, Investigative Reporter (answer to narrator's question).
TH3	Craig Unger, Author, House of Bush, House of Saud (answer to narrator's question).
TH4	George W. Bush

Figure 5. Description of differences between D and TH

All three dialogue passages involve an interviewer and interviewee, who interact and even interrupt each other, but the talking heads passages involve just one person, with the interviewer not taking part in the conversation. The talking heads passages have the highest number of words per sentence, as shown in Figure 4. Therefore, even though these discourses are as unprepared and as natural as dialogues, they are more formal and more spontaneous.

4. Tools and methodology

In order to research MT engines' performance in each of the three types of speech, the distance between the Spanish translation obtained with different MT systems and the official published translation was measured. We began by transcribing the English source texts and their official Spanish translations. We then machine-translated the source texts using different MT engines to obtain a raw MT output. Finally, the official, human-translated Spanish texts were compared with the raw machine translated excerpts. The unedited MT output was therefore compared against human-translated texts, rather than against postedited MTs, so that the distance between raw MT output and the published human translation could be measured.

Two main types of tools were used for the purpose of this study: MT engines and a quality-assessment module of an MT engine. As for MT engines, five systems were tested: two statistical machine translation (SMT) engines, two rule-based machine translation (RBMT) engines, and a hybrid system. The next figure shows the engines used and their nature:

MT engine	Type
Apertium	Rule-based

Bing Translator	Statistical
Google Translate	Statistical
Lucy LT	Rule-based
Systran	Hybrid (statistical/rule-based)

Figure 6. Machine Translation engines used and their type

For the purpose of this research the demo or free versions of these systems were used, and none of them were specifically trained.[4] MT quality metrics were calculated using Language Studio Pro 3.0.2.0 by Asia Online, a tool for which an academic license was obtained and that provided an integrated interface with all the metrics needed.

5. Automatic metrics

Along with the development of MT, research efforts were put in different ways to automatically evaluate the quality of raw MT output. This has resulted in the creation of several measures, such as BLEU, TER, METEOR or F-measure, among others.

One of the first and still best known automatic measures is BLEU (Bilingual Evaluation Understudy). BLEU is based on the underlying assumption that, according to Papinesi, Roukos, Ward and Zhu (2002: 1), "[t]he closer a machine translation is to a professional human translation, the better it is." Other metrics have been developed since in order to improve or complement previous existing instruments. According to Snover *et al*. (2006), TER "measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation." METEOR (Metric for Evaluation of Translation with Explicit Ordering), which was designed to fix some known issues in BLEU, "counts the number of exact word matches between the system output and reference" (Snover *et al.*, 2006). The F-measure, instead, as claimed by Turian, Shen and Dan Melamed (2003: 1) is a "standard measure" that corresponds to the average between precision and recall. Even if the results of automatic metrics might be used as a reference value, they should not be considered quality indicators *per se,* since the quality of the translated message is not always correlated with the editing distance between two sentences.

6. Results

Both the entire corpus as a whole and the three individual subcorpora were analysed. This section describes the results obtained using the above-mentioned automatic metrics: BLEU, TER, METEOR and F-measure. It was decided to use several metrics (the ones presented previously) in order to increase the accuracy of the results through triangulation. MT outputs were

evaluated using Asia Online's parallel —untrained— English-Spanish corpus. Importantly, this corpus does not include any parts of the documentary we are dealing with, since this would interfere with the results of the raw MT output metrics.

Results with BLEU: Since this study compares raw MT output against human translation, the scores obtained are expected to be a little lower than it would if postedited translations were used as the benchmark.

BLEU results	APERTIUM	BING	GOOGLE	LUCY	SYSTRAN
Narration	23.09	38.62	36.6	26.28	30.47
Dialogue	10.68	24.28	25.00	16.72	18.87
Talking heads	17.44	28.92	33.62	21.32	20.65
Whole corpus	17.92	31.56	32.03	21.95	24.07

Figure 7. BLEU results

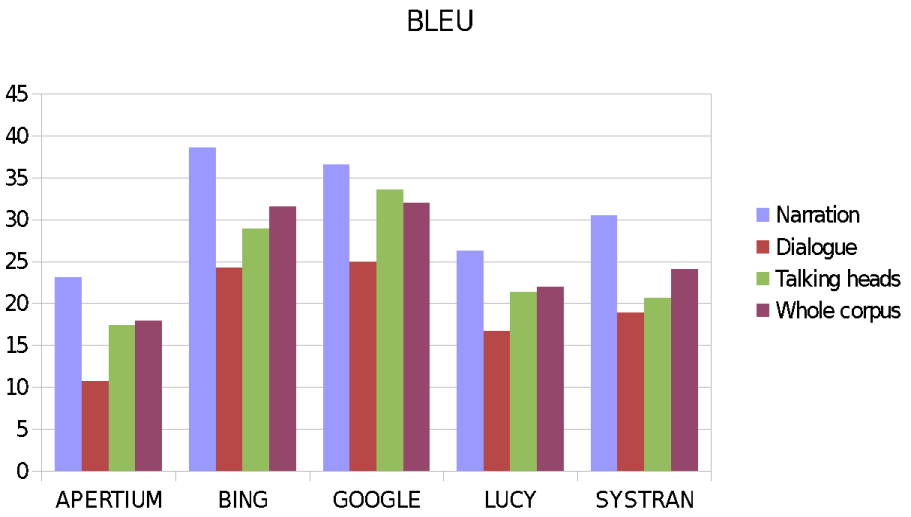

Figure 8. BLEU results

The BLEU results show that translations of narration and talking heads seem to achieve better results than translations of dialogue. Moreover, they also show that output obtained from some statistical MT systems may be high enough for postediting purposes. Lavie (2010) says that scores above 30 usually reflect that the MT output is understandable. The Bing, Google and Systran translations of narration and talking heads scored more than 30. None of the translations of dialogue scored more than 30.

Results with TER: Scores obtained with TER (Translation Edit/Error Rate) are comparable to those obtained with BLEU:

TER results	APERTIUM	BING	GOOGLE	LUCY	SYSTRAN
Narration	30.11	56.26	41.76	32.53	41.76

Dialogue	23.5	37.79	38.48	31.34	31.57
Talking heads	32.76	44.83	47.24	38.28	36.9
Whole corpus	28.33	46.74	41.9	33.5	36.81

Figure 9. TER results

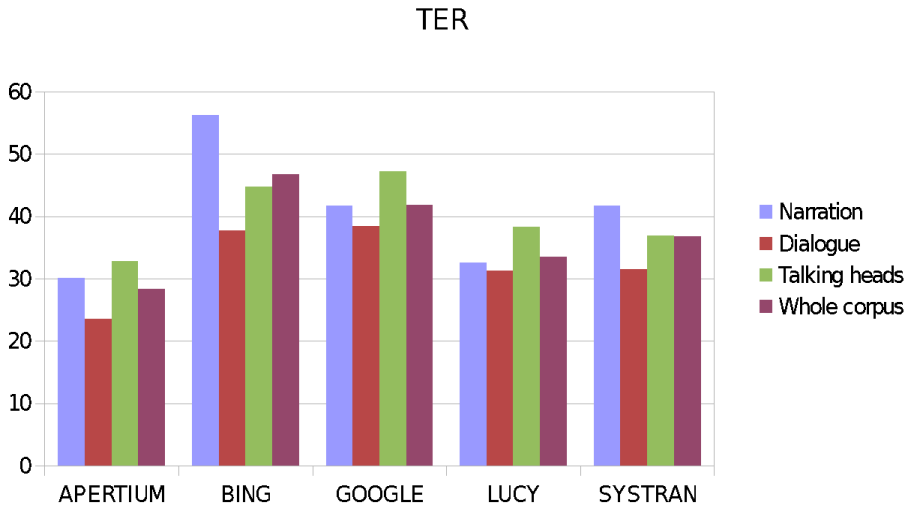

Figure 10. TER results

Again, translations of narration and talking heads scored higher than translations of dialogue.

Results with METEOR: The results obtained with METEOR metrics are comparable to the BLEU and the TER results described above.

METEOR results	APERTIUM	BING	GOOGLE	LUCY	SYSTRAN
Narration	18.28	29.41	24.15	20.24	25.73
Dialogue	14.53	22.35	23.74	19.48	19.88
Talking heads	19.87	25.33	27.16	21.33	20.82
Whole corpus	17.35	25.8	24.77	20.26	22.31

Figure 11. METEOR results

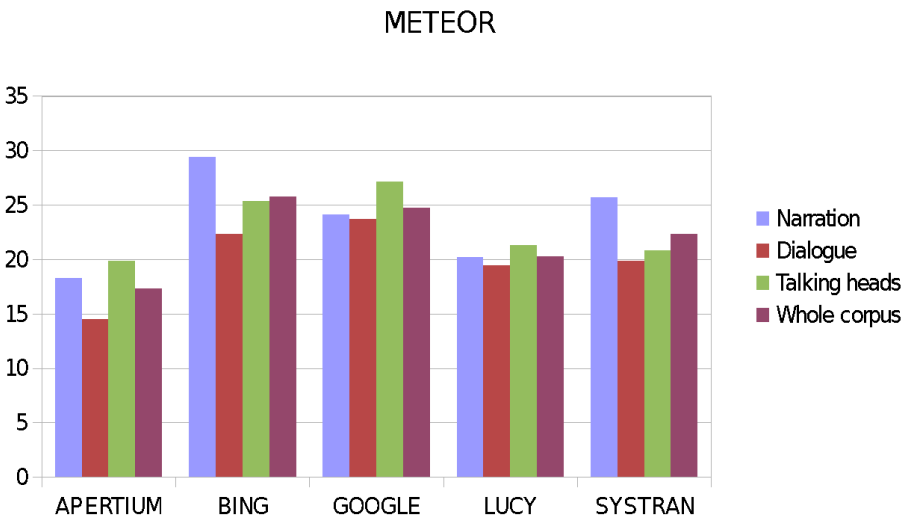

Figure 12. METEOR results

Nevertheless, none of the translations scored above 50, the score that generally reflects that translations are understandable (Lavie, 2010). Moreover, the differences in the scores for the different types of text are smaller than in the BLEU and TER results.

F-measures: The F-measure results follow the same pattern as the BLEU and TER results.

F-MEASURES	APERTIUM	BING	GOOGLE	LUCY	SYSTRAN
Narration	53.81	68.04	63.5	56.19	61.97
Dialogue	41.51	55.29	57.84	52.11	53.23
Talking heads	52.49	62.23	63.74	59.48	58.77
Whole corpus	49.07	61.99	61.50	55.54	57.96

Figure 13. F-measures

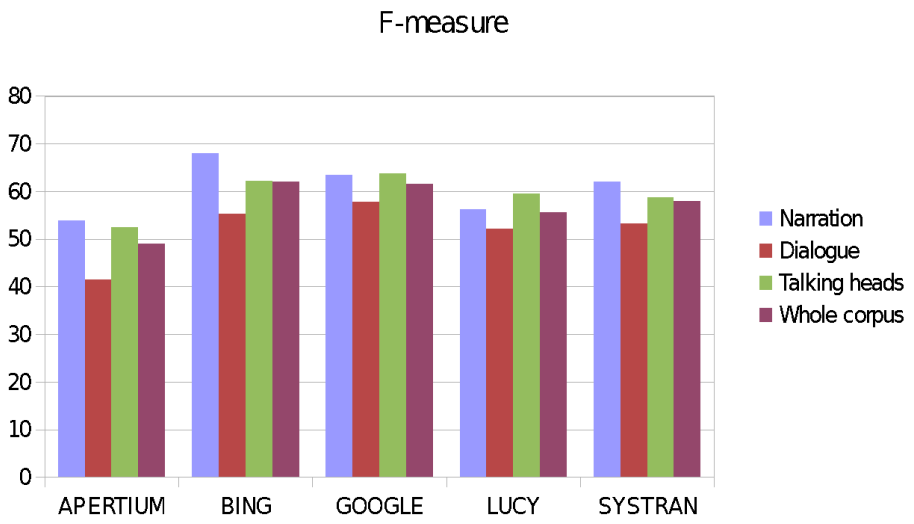

Figure 14. F-measures

As seen in the results above, the translations of narration and talking heads achieved the best scores using all the metrics. Such texts typically have long sentences (16 to 18 words on average). Their vocabulary and style are less formal than written text (they have been produced to be heard, not to be read) but less spontaneous than conversation. These kinds of texts seem to bring out the best in untrained SMT systems, which suggests they could be used in the production processes of multilingual documentaries. Consequently, the performance of untrained SMT systems based on corpora that include a wide range of topics would be acceptable for postediting. They would be able to translate properly middle length sentences (under 20 words), as well as to face the lexical variety used in documentaries on any topic. As for RBMT engines, they seem to get the lowest results in all categories and all metrics. In our opinion, it should not be concluded from these results that RBMT engines are not valid for translating documentaries, but they indicate that some extra rules should be created according to the kind of texts being translated.

7. Discussion and conclusions

This paper discusses the introduction of MT in the localisation of documentaries. Our first conclusion is that the results are suggestive that the methodological approach works and could be expanded into a focused study on MT and audiovisual products. Despite the fact that the corpus we set up (with transcripts of different types of speech, narration, dialogue and talking heads) was not a large one, selecting rich passages of the texts might be enough for an explanatory study of this topic.

Each metric used in this study (BLEU, TER, METEOR and F-measure) produced similar results. SMT systems scored slightly higher in all four metrics, and narration and talking heads scored higher than dialogues. These two types of speech share two important features: they are formal oral texts (relatively long but well-constructed sentences, direct style,

among others aspects of discourse), and they are very well-recorded (there is no ambient sound, and spontaneous discourse markers are usually edited out). We could conclude that these oral discourses have a very similar level of formality to that of the written texts contained in the bilingual corpora of SMT systems like Bing Translator and Google Translate. In this regard, it might be interesting to compare, probably in a controlled environment, the time required, cognitive effort and costs of human translation of a formal documentary (without dialogue) versus MT followed by postediting. MT followed by postediting would perhaps perform better than expected for documentaries dealing with a common subject.

Our study shows that MT can be used to translate audiovisual products containing direct, formal discourse, and therefore it might be interesting to research MT potential in the translation of other products, such as training and instructional videos. MT systems (particularly RBMT) might improve their output if they were fed glossaries and terminology databases containing terms used in the field dealt with in the document being translated. It must be reminded that SMT output tends to improve if the engine is trained with domain-specific corpora.

Nevertheless, an important obstacle with audiovisual products is obtaining a written source text that can be fed into the MT system. Either if the audio is transcribed manually or by using voice-recognition technology, texts need heavy editing before they are ready for MT, especially where the soundtrack contains ambient noise. Hesitations and other features of oral discourse pose additional problems, since they are difficult to transcribe into text that can be processed by a MT system and are difficult for such systems to translate. However, if SMT systems were fed with corpora containing such features of oral discourse, perhaps they would be able to deal with them when translating oral texts.

Bibliography

- **Agarwal, Abhaya and Alon Lavie** (2008). "METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output." *Proceedings of the Third Workshop on Statistical Machine Translation at the 46th Meeting of the Association for Computational Linguistics (ACL-2008)*, Columbus, OH, June 2008, 115-118, http://www.aclweb.org/anthology/W/W08/W08-0312.pdf (consulted 16.10.2014).

- **Banerjee, Satanjeev and Alon Lavie** (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)* 29 June 2005, Ann Arbor, Michigan, 65-72, https://aclweb.org/anthology/W/W05/W05-09.pdf (consulted 16.10.2014).

- **Espasa, Eva** (2004). "Myths about Documentary Translation." Pilar Orero (ed.) (2004). *Topics on Audiovisual Translation*. Amsterdam/Philadelphia: John Benjamins, 183–197.

- **Franco, Eliana** (2001). "Voiced-over television documentaries. Terminological and conceptual issues for their research." *Target* 13(2), 289-304.

- **Gambier, Yves and Eija Suomela-Salmi** (1994). "Subtitling: A type of transfer." Federico Eguíluz, Raquel Merino, Vickie Olsen, Eterio Pajares and José Miguel Santamaría (eds) (1994). *Transvases culturales: Literatura, Cine, Traducción: Proceedings of the symposium*. Vitoria-Gasteiz: Universidad del Pais Vasco, 243–252.

- **Lavie, Alon and Abhaya Agarwal** (2005). "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments." *Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic, 23rd June 2007, 228-231.

- **Lavie, Alon** (2010). *Evaluating the output of machine translation systems*. AMTA Tutorial. http://amta2010.amtaweb.org/AMTA/papers/6-04-LavieMTEvaluation.pdf (consulted 16.10.2014).

- **Matamala, Anna** (2009). "Main challenges in the translation of documentaries." Jorge Díaz-Cintas (ed.) *In So Many Words: Translating for the Screen*. London: Multilingual Matters, 109-120.

- **Orero, Pilar** (2004). "The pretended easiness of voice-over translation of TV interviews." *JoSTrans, The Journal of Specialised Translation* 2, 76-96.

- **Orero, Pilar** (2007). "Voice-over: A Case of Hyper-reality." *MuTra 2006. Audiovisual Translation Scenarios: Conference Proceedings*. http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Orero_Pilar.pdf (consulted 16.10.2014)

- **Orero, Pilar** (2009). "Voice-over in Audiovisual Translation." Anderman, Gunilla and Jorge Díaz-Cintas (eds) *Audiovisual Translation. Language Transfer on Screen.* London: Palgrave Macmillan, 130-139.

- **Papineni, Kishore; Salim Roukos, Todd Ward and Wei-Jing Zhu** (2002). "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 7th-12th July 2002. http://aclweb.org/anthology/P/P02/P02-1040.pdf, 311-318 (consulted 16.10.2014).

- **Snover, Matthew and Bonnie Dorr** (2006). "A Study of Translation Edit Rate with Targeted Human Annotation." *Proceedings of Association for Machine Translation in the Americas*. 8th – 12th August 2006. https://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf (consulted 16.10.2014).

- **Snover, Matthew; Nitin Madnani, Bonnie Dorr and Richard** Schwartz (2009). "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric." *Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL-2009*, Athens, Greece, 30th-31st March 2009, 259-268.

- **Turian, Joseph; Luke Shen and I. Dan Melamed** (2003). "Evaluation of Machine Translation and its Evaluation." Proteus technical report #03-005, a revised version of the paper presented at MT Summit IX, New Orleans, LA. http://nlp.cs.nyu.edu/publication/papers/turian-summit03eval.pdf (consulted 16.10.2014).

- **Ulitkin, Ilya** (2013). "Human Translation vs. Machine Translation: Rise of the Machines," *Translator Journal* 17 (1). http://www.bokorlang.com/journal/63mtquality.htm (consulted 16.10.2014).

Primary film source

- Moore, Michael (2004). *Fahrenheit 9/11*. 122m. Producers: Michael Moore, Kathleen Glynn, Jim Czarnecki. New York: Dog Eat Dog Films.

Biographies

Dr. **Adrià Martín-Mor** is Lecturer in translation technologies at the Department of Translation, Interpreting and East Asian Studies at the Universitat Autònoma de Barcelona, where he teaches translation technologies and coordinates the Tradumàtica MA course. He holds a PhD and an MA in Translation Studies. His research interests are CAT tools, machine translation, minoritised languages and FOSS software.

E-mail: adria.martin@uab.cat

Dr. **Pilar Sánchez-Gijón** is a senior lecturer in translation technologies at the Department of Translation, Interpreting and East Asian Studies at the Universitat Autònoma de Barcelona, where she teaches subjects related to CAT tools, corpus linguistics, machine translation, post-editing, and terminology.

E-mail: pilar.sanchez.gijon@uab.cat

Appendix - Passages transcribed from the film

N 0:25:25-0:30:05
D1 0:21:58-0:22:26
D2 0:23:24-0:24:15
D3 0:24:34-0:25:24

TH1 0:27:48-0:28:18
TH2 0:29:05-0:29:22
TH3 0:29:22-0:29:41
TH4 0:29:41-0.29:58

Notes

[1] This article is part of the research project "Sensorial and linguistic accessibility" (FFI2012-31024), funded by the Spanish Ministerio de Economía y Competitividad.

[2] The methodology proposed in this paper would probably suit the translation for production model cited above. However, since this paper presents an exploratory study, the corpus is based on already published films.

[3] It must be stressed, however, that it was part of our task to MT the transcribed scripts, and we do not have any information about whether MT was used during the official translation of the documentary.

[4] In this exploratory study, engines were not to be fed with specific corpora in order to get a general overview of the results which can be extrapolated to documentaries on any topic. The customisation of the MT engines would represent, from the methodological point of view, an uncontrolled variable.