Automatically building translation memories for subtitling

KATHERIN PÉREZ ROJAS

Universitat Autònoma de Barcelona, Spain
University of Wolverhampton, United Kingdom

This article describes a methodology to automatically create translation memories for subtitling, using translated books adapted into films and recognising extra-linguistic markers to differentiate character interventions from narration. This methodology includes the automatic identification, extraction, and alignment of the dialogues. The aligned bi-texts served as translation memories in the subtitling of the adapted films. Results show an overall 95% extraction rate for English dialogues and 85% for Spanish dialogues. Alignment showed an accuracy of 90%. Results for the translation memory performance showed that hits between 70% and 100% matches accounted for 15% of the corpus. The results reinforce the claim that dialogues in books can be used as reference material for the translation of subtitles.

Keywords: audiovisual translation; translation memories; subtitling; natural language processing

Introduction

The use of automatic tools to help in the translation process is a field with a great amount of academic research and well proven usability in the market. However, in audiovisual translation, and more specifically in subtitling, there are very few tools to support this process by granting access to the creation of subtitles (Mejías 2010). The lack of resources for automatic subtitle creation may result from the difficulty of processing an ever changing language. Unlike scientific, legal, or even literary texts, audiovisual material does not follow guidelines and cannot be framed in a single type of language.

This research focuses on developing a methodology to build, automatically, translation memories to assist in the subtitling process.

From Translation Research Projects 5, eds. Esther Torres-Simon and David Orrego-Carmona,
Tarragona: Intercultural Studies Group, 2014. pp. 51-62.
http://isg.urv.es/publicity/isg/publications/trp_5_2014/index.htm

Literature review

The Code of Good Subtitling Practice (Carroll and Ivarsson 1998) establishes guidelines for the subtitling task. The authors propose a standardisation of technical features including length of subtitles, number of characters per line, duration on screen, font and size of the letters, and of language features like coherence, cohesion, and treatment of interjections. However, with the increasing demand for subtitling, short deadlines and tight budgets, translators often need the help of more automated or semi-automated methodologies to help them follow good practices and meet deadlines (SUMAT 2011).

In recent decades, researchers in audiovisual translation and machine translation have begun to combine the two fields in order to provide audiovisual translation techniques with tools that focus not just on the technical aspects but also on the translation. The support computer-based systems offer to the subtitling tasks mainly focuses on mechanical aspects such as time coding and word processing, while the possibility to reuse previous translations in new translation assignments remains unaided.

Previous studies have approached the reusability of translations in audiovisual material and automated subtitling. The STAR project developed a rule-based machine translation system to produce Japanese subtitles for English news programmes and Japanese subtitles for newswire translation services (Sumiyoshi et al. 1995: 4). The project Global Translation Systems (GTI), Inc. (Díaz Cintas and Remael 2007: 20-21) uses SYSTRAN to translate, in real time, English subtitles into Spanish subtitles on selected television programmes. SUMAT (2011) is an on-going project to develop an online service for subtitling through machine translation in nine different European languages, with the aim being to semi-automatize the subtitle translation processes on a large scale.

However, when it comes to films, specifically those that are adapted from novels, the translator could rely on the same source the scriptwriters used: the written novel. Harrington (1977) estimates that a third of all films ever made have been adapted from novels, without including other literary forms (such as plays or short stories). The Writing Studio (2001-2004) confirms that over fifty percent of feature-length films for both cinema and television are adaptations of novels, short stories, plays or nonfiction journalism, which together account for 25 percent of all adapted feature films.

It is safe to say that almost all great literary works have been adapted at least once in cinema history. When adapting a novel into a script for a film, and especially when writing a screenplay, "[t]he essence of dialogue and subtext should stay the same […] despite several tricks to "cull and shape the cinematic elements" (Online Film School). When adapting, the writer should compress all dialogues "so that it has the economy and directness of screen dialogue" (ibid). Dialogue in literature is mimetic (as opposed to diegetic), the writer tries to "create the illusion that it is not he who speaks", therefore dialogue in novels is generally direct discourse; a *quotation* of a character's words (Rimmon-Kenan, 1983). Also, Lotman (1989) (as quoted by Rauma 2004) argues that cinematic dialogue is equivalent to dialogue in novels or plays and is thus an indistinctive property of the film medium.

Methodology

It may be possible to use the data contained in novels that have been adapted into films as reference material for the translation of the subtitles for the films. The present research thus seeks to ascertain whether that material can create translation memories suitable for the production of translated subtitles.

To work on this hypothesis, a series of experiments were designed and conducted with the objective of defining a method to search for relevant data in the novels, extract it and create translation memories. We assessed the efficacy of the search and extraction of the information, the alignment, and the matching percentages in the translation memory.

The automatic creation of the translation memories consisted of four stages. The first stage analyzed five film-adapted novels in English and their Spanish translations and extracted the common dialogue features in both languages. The second stage focused on the creation of rules to edit and standardize the texts for processing, the creation of the dialogue extraction scripts and the testing of eight further novels in English and their translations in Spanish. The third stage focused on the alignment of the dialogues extracted and the creation of a TMX file, fine-tuning the pre-processing and extraction scripts as well as the revision of the extraction using seven additional novels in English and their translations into Spanish. The final stage focused on collecting and analyzing the data, and a first evaluation of the translation memory. In all, 20 novels in English and their corresponding translations were analyzed and tested for dialogue extraction.

Definition and characterization of dialogues in English and Spanish novels

Literary theorists (Bakhtin 1986; Maranhão 1990) define *dialogue* as the conversation, or the literary work in the form of a conversation, between characters, often used as a mechanism to reveal characters and to develop and make the plot advance. Dialogues are the lines that a character speaks in any literary work.

Dialogues can be diferentiated from the surrounding prose by using punctuation marks. Sherlock (2011) states that "in standard American usage, opened and closed double quotation marks [" "] indicate when narration has stopped and a character's dialogue begins." This is supported by Elson and Mckeown (2010), who state that, in English literary novels, quoted speech or dialogue is “considered to be a block of text within a paragraph, falling between quotation marks”.

According to Portolés et al. (2009), Spanish novel dialogues tend to be written directly, without any clarifying introduction, and with very few comments or detailed explanations of the mental state or features of the speaker. In Spanish novels, dialogues open with an em-dash, *raya*, [—] at the beginning of the sentence (but it is not repeated at the sentence closure), and when indicating the speaker; closing only when there is a clarification in the middle.

Tables 1 and 2 present the most common standard combinations of punctuation, narration and character intervention (CI) or dialogue sequences found in English and Spanish novels.

Table 1: Types of narration and character intervention (CI) in English novels

English	**Example**
“CI”	“Oh, yes! I know that! I know that, but do you know what day it is?”
“CI” narration.	“She’s packing them,” explained Mother.
narration "CI".	He said to the driver, “You are early tonight, my friend.”
“CI” narration “CI”.	“But where?” he asked. “Where are we going exactly? Why can't we stay here?”
narration “CI” narration.	She saw, I suppose, the doubt in my face, for she put the rosary round my neck and said, “For your mother's sake,” and went out of the room.

Table 2: Types of narration and character intervention (CI) in Spanish novels

Spanish	**Example**
—CI	—Los Potter, eso es, eso es lo que he oído…
—CI —narration.	—Tendremos a papá y a mamá y a nosotras mismas —dijo Beth alegremente desde su rincón.
narration —CI	Hubo un momento de silencio, hasta que Padre dijo: —¿Y bien? ¿Qué opinas?
—CI —narration—. CI	—No —respondió en tono cortante—. ¿Por qué?
—CI —narration—. CI. —narration— CI.	—Hemos recibido denuncias sobre hombres y mujeres vagabundos que desaparecieron el mes pasado —intervino Banks—. Al principio pensamos que podría ser uno de ellos, pero no es así. —añadió en tono dramático—. La víctima fue una de esas personas de anoche.
—CI —narration—. CI. — CI2.	—Pero ¿adónde? —preguntó—. ¿Adónde nos vamos? ¿Por qué no podemos quedarnos aquí? —Es por el trabajo de tu padre. Ya sabes lo importante que es, ¿verdad?

English and Spanish pre-processing of the novels

English dialogues are represented by single (‘ ’) or double (“ ”) quotation marks. There are two ways of expressing them: 1) The ones with identical form (neutral, vertical, straight, typewriter, or “dumb” quotation marks), typewriter double quotes (" ") and typewriter single quotes (' '). 2) The ones with left and right hand distinction (typographic, curly) typewriter double quotes curly (“ ”) and typewriter single quotes curly (‘ ’). For the sake of standardization, and to enhance recognition of the dialogues, it was decided to convert all quotes to the typewriter double quotes curly (“ ”) format.

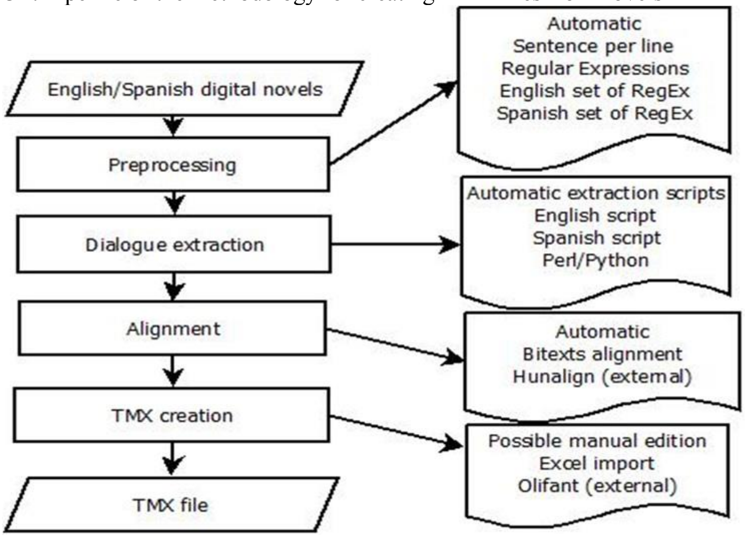
The main feature of the Spanish dialogues is the em-dash (—), but it needs to be in accordance with another feature (punctuation mark, capital letter) to distinguish the caracter intervention from the narrative part. For technical reasons we changed the em-dashes into en-dashes (–).

English and Spanish dialogue extraction scripts

The scripts to extract the dialogues from the English and Spanish texts were designed to read each line of the novel and recognize the dialogues based on the previously mentioned marks.

Figure 1 is a flow chart graphically showing the process and the description of the modules included in the methodology.

Figure 1: Pipeline of the methodology for creating TMX files from novels



The most common error in the extraction phase of the English texts was due to problems in recognizing the marks when the extraction involved very long segments. These segments contained several character interventions as well as narration that required editing in the alignment phase. The most common error in the Spanish extraction was when there was a character intervention that did not follow the specified mark rules; therefore it was not recognized and extracted. However, since most of these errors usually took place in the second part of the character intervention, the first part of it was always extracted, providing the first sentence of the dialogue.

Evaluation results and discussion

For the dialogue extraction, the notion of accuracy is understood as the amount of the original dialogue that the script is able to recognize and extract. For the TMX file results, accuracy is in terms of the fuzzy matches and the match percentage.

The results of the method will be presented in three different categories:

(i) Results of the dialogue extraction,

(ii) results of the alignment, and

(iii) first results of the TM performance when translating the scripts.

Results of the dialogue extraction

The total amount of dialogues was retrieved using ReGex to recognise the proposed patterns and then they were manually checked to compare the correct recognition of all of them, especially for the Spanish novels were the dialogues needed to be checked to determine the end of a dialogue and the beginning of narrative. Table 3 shows the difference rates between English orignal and extracted amount of CI, Spanish original and extracted amount of CI and the difference rate between both original sets.

On average, the dialogue sequences differ by 14%, which is not surprising because in English there are more dialogue markers (quotation marks) for the same dialogue in Spanish. For English, the difference between the dialogue in the original novel and the dialogues recognised and extracted by the script is 5%. For Spanish, the average extraction difference is 18%. Again, this was expected because the Spanish script recognizes all the character interventions and merges them into one, outputting just one chain of dialogue.

Table 3: Percentage difference between the total numbers of English and Spanish dialogues

Title	Original CI Eng	Extracted CI Eng	Org/Ext CI Eng %	Original CI Spa	Extracted CI Spa	Ori/Ext CI Spa %	% Differ Org/CI Eng/Spa
Atonement	1010	998	1,2	932	836	10	8
The bone collector	4372	4230	3,2	4050	3489	14	7
The boy in the striped pyjamas	1451	1394	3,9	1306	935	28	10
One flew over the cuckoo's nest	1713	1679	2,0	1554	1150	26	9
The devil wears Prada	2312	2254	2,5	2139	1654	23	7
Dracula	1150	1098	4,5	1048	927	12	9
East of Eden	7623	7543	1,0	5898	5361	9	23
The great Gatsby	1438	1348	6,3	1153	922	20	20
Memoirs of a geisha	2875	2740	4,7	2415	1896	21	16
The hitchhiker's guide to galaxy	1862	1828	1,8	1453	1272	12	22
Harry Potter 1	2396	2230	6,9	2205	1700	23	8
Harry Potter 2	2919	2748	5,9	2431	1846	24	17
Harry Potter 3	3826	3708	3,1	3510	2481	29	8
Harry Potter 4	6004	5365	10,6	4940	3773	24	18
Harry Potter 6	6124	5491	10,3	5523	3977	28	10
Little women	1723	1614	6,3	1611	1608	0,1	7
Murder on the Orient Express	2572	2346	8,8	2377	2075	13	8
Pride and prejudice	1783	1597	10,4	1298	1190	8	27
Sense and sensibility	1580	1481	6,3	1089	993	9	31
The hunger games	1537	1536	0,1	1398	1094	22	9
	Average %		5			18	14

Results of the alignment

The results of the alignment of the extracted dialogues showed that most of the books followed a pattern according to which the number of sentences, the disposition of lines, and the continuity of the speech are similar in both languages. Since there is no gold standard for the evaluation of the alignment of these dialogues, the alignment was measured taking into account a shuffled sample corresponding to 10% of each aligned bitext. On average, the alignment was precise for about 90% of the cases, with an erroneous alignment rate of 10%.

First results of the TM performance when translating the scripts

This step aims at providing preliminary findings on how many complete matches or partial matches were retrieved from the TM. These results were obtained using OmegaT 2.6.3 and the entire corpus of TMs against the subtitle corpus of all the adaptated movies.

Table 4 shows the total result of repetitions, exact matches, fuzzy matches and no matches found when using the translation memories as reference material for the translation of the subtitles.

Table 4: General results and percentages of Match Statistics with OmegaT

Type of segment	Total Occurrences	%
Repetition	2105	5.9%
Exact Matches	1129	3.2%
Segm 75-99%	1960	5.8%
Segm 50-74%	23553	69.8%
No Match	5334	15.2%
Total Segments	34081	

The general results show that of 34,081 analyzed segments, repetitions account for 5.9% (2,015 segments were repeated). The repetition feature of the translation memory means that there is one translation per repeated segment, if the translation memory system finds the exact segment in the translation memory file; this 5.9% can be automatically translated. This also accounts for the Exact Matches, which were 3.2% of the corpus (1,129 segments).

In literary narrative, dialogues are one of the main guides for plot development, but the neighboring description is in charge of directing the references thereby limiting the dialogues to their function of continuity. This is reflected in the matching of subtitles with their extracted dialogues, especially when the character interventions in the subtitles and in the novel are the same. In these cases, there may be a 100% match, however, that match will not occur frequently along the text: we see that 9.1% was found to be repeated, rendering it difficult to rely on automatic translation as a resource to speed up the translation process.

Note that 5.8% of the corpus (1,960 segments) obtained a match between 75% and 99%. This range is considered highly relevant for translation because these segments need minor edition and, as a whole, save the translator a great amount of time (Somers 2003, Bowker 2005).

In total, 69.8% (23,552 segments) of the corpus found a match in the 50%-74% group. Although it can be assumed that matches below 70% are not very useful for translation memory systems, many TM software products state the minimum match value between 60% and 75%, and recommend starting with a low percentage such as 50%. With these values in mind, the hits found in the fifth group could be considered partial matches (O'Brien 1998). As shown by the alignment and preliminary fuzzy match analysis, the tendency is that most of the dialogues stored in the TM will serve as contextual information for the subtitles.

The remaining 15.2% of the corpus (5334 segments) obtained no match with the subtitle TM file. It can be assumed that these dialogue segments either belong to films with a high degree of free adaptation (change of plot, change of characters, etc.) or to novels whose dialogues tend to be especially longer in the original and have been significantly cut.

Conclusions and future work

The aim of this paper was to design a method to automatically create translation memory files for subtitling. The method searches for dialogue marks in the novels and sets the boundaries to extract each character intervention. These character interventions are then aligned to create a TMX file that is used as translation memory when translating the subtitles.

It is possible to automatically recognize and extract dialogues from the English and Spanish novels. This process is language-dependent, requiring a defined set of pre-processing steps and a specific set of scripts per language. The alignment results show that the books and their translations followed similar organization patterns and speech continuity.

The preliminary translation memory test showed that a high percentage of partial matches were in the 50% to 74% group. Probably these low percentages were due to the transformations from written novel to cinematographic script, but an initial analysis showed that they contain enough contextual information to help in the translation process.

In regard to future work, it is necessary to improve the features of the dialogue extractor, taking into account the special needs of Spanish narrative, by marking the end of dialogues to allow full automatic recognition. It is also necessary to test the resulting TM files using different translation memory systems. Further ways of automatically shortening the longer dialogues might be useful to obtain higher matching scores in the TM.

References

Bakhtin, Mikhail. 1986. *Speech Genres and Other Late Essays*. Translated by Vern W. McGee. Austin: University of Texas Press.

Bowker, Lynne. 2005. “Productivity vs Quality: A pilot study on the impact of translation memory systems”. *Localisation Focus* 4(1): 13-20.

Carroll, Mary, and Jan Ivarsson. 1998. *Code of Good Subtitling Practice*. Approved at the meeting of the European Association for Studies in Screen Translation in Berlin.

Díaz Cintas, Jorge, and Aline Remael. 2007. *Audiovisual translation: subtitling*. Manchester: St. Jerome.

Elson, David, and Kathleen McKeown. 2010 “Automatic Attribution of Quoted Speech in Literary Narrative”. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*, Atlanta.1013-1019.

Harrington, John. 1977. *Film and/as literature*. Englewood Cliffs NJ: Prentice-Hall

Maranhão, Tullio. 1990. *The Interpretation of Dialogue*. Chicago: University of Chicago Press.

Mejías Moreno, Raquel. 2010. *Herramienta para la traducción asistida en la industria audiovisual*. Proyecto de Fin de Carrera. Escuela Politécnica Superior. Universidad Carlos III de Madrid.

O’Brien, Sharon. 1998. *Practical experience of computer-aided translation tools in the localization industry*. Lynne Bowker, Michael Cronin, Dorothy Kenny and Jennifer Pearson (eds) Unity in Diversity: Current Trends in Translation Studies. Manchester: St. Jerome. 115-122.

Online Film School. *Script adaptation*. http://filmschoolonline.com/sample_lessons/sample_lesson_adaptation.htm Visited September 2014.

Portolés, Elisa. 2009. “Cómo escribir diálogos”. *Online didactic resource* http://www.alquimistasdelapalabra.com/dialogos/teoria_dialogo/index.html. Visited September 2014.

Rauma, Sara I. 2004. *Cinematic dialogue, literary dialogue and the art of adaptation*. Pro Gradu Thesis. University of Jyväskyla. https://jyx.jyu.fi/dspace/bitstream/handle/123456789/7349/G0000703.pdf?sequence=1Visited. September 2014.

Rimmon-Kenan, Shlomith. 1989. *Narrative fiction: Contemporary poetics.* London: Methuen.

Sherlock, Karl. 2011. “Some practical advice about using punctuation and markers in creative writing”. *English 126 Creative Writing, Punctuation and Prose* http://www.grossmont.edu/karl.sherlock/English126/punctuation.html. Visited September 2014.

Somers, Harold. 2003. “Translation memory systems”. In Harold Somers (ed.) *Computers and Translation: A Translator’s Guide*. Amsterdam and Philadelphia: Benjamins. 31-47

SUMAT An Online Service for Subtitling by Machine Translation. 2011 *Annual Public Report 2011*.

http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=270919. Visited September 2014.

Sumiyoshi, Hideki, Hideki Tanaka, Nobuko Hatada and Terumasa Ehara. 1995 “Translation workbench for generating subtitles for English TV news”. Tokyo: NHK Science and Technical Research Laboratories.

Writing Studio. 2001. *The art of adaptation*. http://www.writingstudio.co.za/page62.html.Visited September 2014.