Investigating Repetition and Reusability of Translations in Subtitle Corpora for Use with Example-Based Machine Translation[1]

Marian Flanagan[2] and Dorothy Kenny[2]

1. Introduction

Repetition and reusability are two notions that lie at the heart of a number of current approaches to computer-aided translation (CAT) and machine translation (MT),[3] but are rarely problematized in the literature. In this paper, we discuss these notions in the context of the example-based machine translation (EBMT) of movie subtitles in a project currently underway at Dublin City University. We start by describing, in sections 2 and 3 below, how repetition and reusability have been dealt with by other researchers in CAT and MT, before going on to outline our own approach in section 4. We characterize our approach as combining both 'prospective' and 'retrospective' phases. The prospective phase relies on measurements of levels of repetition in our texts, on the one hand, and human evaluations of the reusability of existing translations, on the other, while the retrospective phase relies on real-user evaluations of automatically translated subtitles after they have been inserted into relevant movie clips. While the focus of this paper is on the prospective evaluation of repetition and reusability in our corpora, we will also make some comments on why prospective measures and judgements might be expected to correlate with retrospective evaluations. We will provide some quantitative analysis results from the corpora used with the EBMT system, as well as an overview of the responses generated in the human evaluation. The paper concludes with comments on the usefulness of a prospective phase, and the benefits this has for the next phase of research.

2. Repetition and reusability in corpus-based machine translation

As already indicated, repetition and reusability are central notions in contemporary machine translation (MT). Current corpus-based approaches to MT rely on parallel corpora, that is, bilingual corpora of source texts and their translations (normally aligned at the sentence level), to discover how parts of the source text are typically translated in the target language. Although the different types of MT are no longer as clearly delineated as they once were (see papers in Carl and Way 2005), we can divide corpus-based MT, in theory at least, into Statistical MT (SMT) and Example-based MT (EBMT). SMT combines knowledge of the probable translations of individual words and phrases as derived from the parallel corpus (the translation model), with a probabilistic model of the target language (the language model), to produce the most likely translation of a given input string. EBMT functions somewhat

---

[1] This research is being funded by the Irish Research Council for the Humanities & Social Sciences (IRCHSS)

[2] School of Applied Language and Intercultural Studies, Dublin City University
e-mail: marian.flanagan23@mail.dcu.ie, dorothy.kenny@dcu.ie

[3] Computer-aided translation refers to the use of software programs to assist human translators in their work, and machine translation refers to the automatic translation of texts, with or without human input.

like a translation memory (TM) system (see below), in that it compares new source-language input (for example, a sentence or a phrase, or fragments of either) against source-language examples already held in a database, and then extracts a preferred (perhaps because it is the most common) corresponding target language translation. The two main differences between EBMT and TM technologies are that firstly, as the above description indicates, EBMT typically functions at subsentential level, whereas TM rarely does, and secondly, translations for successive fragments of a source text are combined *automatically* in EBMT to produce a target language output, whereas human translators working with TMs decide what the target text will ultimately look like.[4] At a very basic level, however, both SMT and EBMT share with TM systems the basic assumption that new source texts will be translatable on the basis of material that has already been translated and stored; and there will be enough repetition of source-language linguistic units already encountered and translated by humans, and enough reusable translations of these units, to make possible either fully automatic translation (SMT and EBMT) or considerably faster human translation (in TM scenarios).

Researchers in EBMT are rarely explicitly interested in repetition levels in texts to be translated however, or in the number of matches they will find between new input and the parallel corpora they work with. As proper evaluation of research systems relies on strict separation of training and test data, there is no value in comparing test data (new input) to training data (the parallel corpus) before runtime. The success of such systems is ascertained after runtime, when the automatically translated output is evaluated either by humans or machines. For much the same reasons, there is no value in examining, prior to runtime, the target-language segments that we can predict will be offered as translations of source-language input, with a view to ascertaining their reusability in the new context. EBMT researchers have, however, been concerned with the impact that the size and internal composition of their parallel corpora can have of the quality of the output of their systems. Thus Armstrong, 2007 and Armstrong et al., 2006 investigate the effects on translation outputs of using smaller, but more homogenous corpora, and larger, but more heterogeneous corpora. In these cases, the quality of outputs is evaluated using a combination of automatic evaluation metrics and human judgements. However, no prospective qualitative analysis was carried out on the corpus.

3. Repetition and reusability in computer-aided translation

In contemporary CAT, the above-mentioned translation memory tool holds pride of place. A translation memory (TM) is basically a collection of previously translated source texts alongside their respective translations. These source and target texts are aligned; that is, explicit links are made between corresponding segments in the

[4] There are, of course, problems associated with both SMT and EBMT. In the former, for example, the language model might override the translation model and yield a perfectly natural, well-formed target-language string that is not an 'accurate' translation of the source string. In EBMT, the combination of previously unconnected target-language chunks into a new sentence or text may result in 'boundary friction', where unsightly 'seams' appear between chunks that do not agree, for example, in person or number. For detailed discussions of issues and techniques in EBMT in particular, see Carl and Way (2003 and 2005).

original and the translation.[5] Such segments are usually identified at the sentence level, but other structural units, for example, cells in a table, or headings, can also be identified as segments. TM tools are designed to automatically check an existing TM to see whether a new source language segment (or something like it) has already been translated by the translator (or a colleague). If this is the case, the existing translation that has been aligned with the source-text segment already in the TM is offered as a possible solution to the human translator. The benefits of this technology are said to include enhanced translator productivity and increased consistency in translation.

It is not our intention here to give a complete overview of TM tools, or to explore the potential disadvantages of the technology. (The reader is referred here to Bowker 2002, Reinke 2004, and Somers 2003, on the one hand, and to Bowker 2005 and Kenny 2007, on the other, for fuller discussions.) Rather, we are interested in how the notions of repetition and reusability are dealt with in the context of TMs: most TM software comes with an analysis tool specifically designed to count the number of repetitions in source texts presented for translation. **Repetitions** here are understood as segments that recur in exactly the same form in the source text. Thus the heading 'Further Reading' might occur at the end of each of seven chapters in a training manual. The first instance is novel, but the other six will constitute repetitions. Repetitions may all be contained within a single text (internal repetitions), or they may be spread across a 'family' (Heyn 1998) of related texts (external repetitions). Whatever the case, the basic thinking is that the translator will have to translate only the first instance of a segment; his/her translation can then be recycled when all the repetitions of that segment are encountered. This kind of analysis can also be conducted against the background of a TM that already contains relevant source segments and their associated target segments. In this case, segments in the new source text will be checked against the TM to see if there are any existing **matches** in memory. 'Further Reading' might have an **exact match** in a source-language segment that has already been translated into German as 'Weiterführende Literatur' for example. Alternatively, the analysis tool might uncover source segments in the TM that are merely similar to a new source segment. These are known as **fuzzy matches**, and their recognition will depend on how high or low the human user sets the similarity threshold in the analysis software. When presented with either an exact or a fuzzy match from memory, the idea is that the human translator will evaluate the match and decide whether he/she should accept or reject the existing translation, or perhaps adapt it on the basis of current needs. The basic assumption behind the use of TM tools to enhance productivity, however, is that reuse of just-created or previously existing translations will be maximised rather than minimised. In commercial TM environments source-text analysis is used to ascertain repetition and match levels, and thus to calculate how much 'leverage' the translator will get out of existing or emerging translations, given a new job.[6] But even this source text analysis cannot directly reflect target text reusability, a point that is made in the more critical literature on TM.

The analysis tools associated with TMs are thus designed to measure levels of repetition in source texts, and the extent to which segments in the source text match exactly or partially segments already in the TM. The matching algorithms used, although proprietary, are generally thought to rely on similarity measures such as edit

[5] A TM is thus a particular instantiation of a parallel corpus, albeit one that is used specifically in the context of CAT, and is usually accessed, updated and maintained using dedicated TM tools (see Bowker and Barlow 2004, Kenny 2007).
[6] And hence how long the job will take, and how much the translator will be paid.

distance, to compute the distance between a new source-language input and segments already in memory. How we can measure the success of these matching algorithms has been the subject of a handful of studies in recent years, two of which we review below, and this is where the notion of **reusability** comes into play. Whyman and Somers (1999), for example, propose an evaluation metric that takes into account the ‘usefulness’ to a translator of matches proposed by a TM tool. They argue that ‘what is required from a translator’s point of view is a direct indication of the practical usefulness in the translation process of the proposed translations, rather than a theoretical measure of the similarity…of these to the input segment’ (1999:1274). They go on to argue that such ‘practical usefulness’ can be captured objectively by measuring ‘the effort required to convert the proposed match into the correct translation’ and that ‘this effort can be quantified in terms of the number of *key-strokes* needed’ (ibid). Whyman and Somers are interested in the number of key-strokes required to transform the target-language segment retrieved from memory into the ‘ideal’ translation (itself a kind of edit distance, albeit one that is independent of the systems evaluated, and can accommodate mouse clicks, etc), but they acknowledge that determining what the ‘ideal’ translation might be introduces an element of subjectivity. They thus opt, somewhat counter-intuitively, to measure the number of keystrokes required to transform the source-language hit from TM into the new source-language query (as if they were translating from English to English), and suggest that this measure will be sufficient ‘for most practical purposes’ (ibid). Whyman and Somers do not manage to eliminate all subjectivity from their evaluation, however, as the metric itself relies on a distinction they make between ‘hits’, that is matches deemed to be ‘relevant’, and the more general category of ‘matches’, or ‘any proposed retrieval from the database’ (1999:1272). The authors themselves acknowledge that recognition of ‘hits’ requires subjective judgements, which are based on previous translation experience, but add that ‘the match which the system ranks as best is indeed the match most likely to be a hit’ (ibid: 1277). Another aspect of Whyman and Somers’s treatment that merits comment here, is that they assume that (the target-language half of) exact matches can simply be pasted into the target document (ibid: 1266, 1268) without further ado. ‘Exact matching of strings of characters is such a straightforward problem’, they add, ‘that this aspect of TM software is of no interest to us whatsoever’ (ibid: 1268), and they thus exclude hits that correspond to exact matches from their analysis. Finally, Whyman and Somers are primarily interested in quantifying the usefulness of hits, and although qualitative phases are implied by the need to divide matches into hits and non-hits, and the need to use a database that is ‘appropriate’ for their test data, they do not comment in any detail on these phases. On the latter point, however, they indicate that ‘The text contains a number of text segments for which matches in the database will be found’ (ibid:1272), thus suggesting some kind of prior familiarisation with their data.

Reinke (2004) conducts a detailed evaluation of the retrieval performance of three TM tools. Like Whyman and Somers (ibid), Reinke notes that the ‘relevance’ of matches in a TM can be judged either on the basis of formal similarity between stored and new source-language segments, or according to the extent to which the retrieved target-language segment fits into the new, as yet emerging, target text. Unlike Whyman and Somers, however, Reinke does not assume that the target-language part of an exact match can be simply pasted into the emerging target text. Orthographically identical sentences may have different meanings, and hence different translations. And even orthographically identical sentences with the *same* meaning might have different translations on different occasions for reasons of text cohesion and/or coherence, amongst others (ibid: 154ff, 237ff; see also Bowker 2005, López Circuelos 2003, and Nedoma and Nedoma 2004). It is also worth noting, that even the producers of translation memory software acknowledge that not all exact matches are of equal value. An exact match that occurs between two other exact matches, for example, is recognised by SDL Trados 2007 as a 'PerfectMatch'™; its enhanced usefulness stemming from the fact that it shares exactly the same local co-text as the current query.

In his evaluation, Reinke (ibid: 161ff) first ascertains repetition levels in new texts to be translated, and exact and fuzzy match levels between these new texts and existing translation memories. The new texts are, in fact, subsequent versions of texts already translated, and Reinke analyses each new text using a smaller memory consisting of the precursor of that new text, and a larger memory, consisting of the precursors of all five new texts. Having seeded three different translation memories, accessed by three different translation memory tools, with relevant segments from the first version of each text ('relevance' here is based on the analyst's judgement and depends on the extent to which the meanings of initial and revised segments coincide), Reinke then records the retrieval performance of each tool, given segments from the new texts as input queries. Although this evaluation yields 'objective' numerical measures for each system's retrieval performance (in terms of recall and precision scores), Reinke ultimately argues for a more qualitative approach to evaluation, and his analysis is actually characterized by detailed discussions of the actual contents of segments.

Our study shares some ground with both Whyman and Somers (ibid) and Reinke (ibid). Like the former, we intend to investigate, initially at least, a single solution for each segment we translate, although the EBMT system we use can potentially produce several translations for a single input. Unlike Whyman and Somers, however, we will focus on the target language. Like Reinke, we have a very deliberately separate source-text analysis phase, and we make informal (at this stage) predictions about how levels of internal and external repetition in our source texts might be expected to influence levels of reusability of translations in our corpus. Again like Reinke, we place heavy emphasis on the qualitative evaluation of translations proposed by our system, although we do not eschew quantitative analysis. We also divide our analysis of reusability into a prospective and a retrospective phase. In the former informants give their opinions on whether or not automatically generated translations of exact matches with our example base will sit comfortably in their new target text. As this suggests, we are thus interested in the EBMT equivalent of exact matches with a translation memory, another point of divergence from, especially, Whyman and Somers. In the retrospective phase, real viewers of movies to which our automatically produced translations are added as translated subtitles give their opinions on the success or otherwise of those subtitles. The prospective phase of our evaluation is elaborated upon in the next section, along with issues of corpus design and creation, and the quantitative analysis of our corpora.

4. Repetition and reusability: the evaluation phase

4.1. Corpus design and creation

Three corpora are used for our research. Firstly there is a Harry Potter corpus (HPC), which is a bilingually aligned corpus of subtitles taken from the four Harry Potter films on DVD. This corpus belongs to the fantasy genre. We call this **corpus A**. We then decided to compile a more general fantasy corpus, so our second corpus contains subtitles from the three Lord of the Rings (LOTR) films, also taken from DVD (LOTRC), and combined this with the HPC, namely **corpus B** (HPC + LOTRC). The third corpus is a combination of the HPC, the LOTRC, and a more general corpus of subtitles taken from 25 films on DVD, the Mixed General corpus (MGC), namely **corpus C** (HPC + LOTRC + MGC). The genre of the films in the MGC ranges from action/adventure to romance and period dramas. The idea behind creating three different corpora relates to work mentioned above (Armstrong et al., 2006). On this occasion, we compiled corpora using only DVD subtitles, and therefore all the corpora could be considered homogenous data; however, we distinguish between: subtitles which are very subject specific and from the fantasy film genre (HPC); those which are slightly less subject specific, but still remain within the same film genre (LOTRC); and those which are not considered subject specific, come from an array of film genres, and which mostly contain everyday spoken language (MGC). We continue to investigate the impact corpus size and homogeneity have on the quality of the output, hence the structure of the three corpora.

4.2. Repetition analysis

The first step in looking for repetitions is to carry out some simple repetition and match analysis using Trados Translator's Workbench 'Analyse' function with the three basic corpora, HPC, LOTRC and MGC. We plan to subtitle movie clips from the first Harry Potter film, and therefore want to get a rough idea of (a) how many repetitions exist within the entire HPC, (**internal and external repetitions**), and (b) the extent to which segments in HPC recur in exactly the same form in the LOTRC on the one hand, and the MGC on the other. To find out (a) we simply analyse the HPC against an empty TM. To find out (b) we analyse the HPC first against a TM containing all the source and target segments from the LOTRC, and then against a TM containing all the source and target segments from the MGC. By comparing the HPC against already seeded TMs, we get a score for the number of **exact** and **fuzzy matches** with that TM. In this, the first stage of our analysis, we are interested in exact matches only. As can be seen from Table 1, the HPC has 920 repetitions. The table also shows 38 100% matches in the LOTRC and 261 100% matches in the MGC. This is an indication that there are segments in both the LOTRC and MGC which are the same as segments in the HPC, and therefore could potentially provide good translations for the corresponding segments, as well as possibly being considered as reusable segments in different contexts.

Corpus	**Translation Memory**	**Repetitions/100% Matches**
Harry Potter	Empty	920 (repetitions)
Harry Potter	LOTRC	38 (100% matches)
Harry Potter	MGC	261 (100% matches)

Table 1: Repetition rates and 100% matches between the Harry Potter corpus and various TMs

After looking at general statistics for the different corpora, we wanted to choose ten movie clips from the first Harry Potter film, *Harry Potter and the Philosopher's Stone*.[7] Given the fact that we wanted to focus on repetitions and the reusability of their translations, it was decided to locate clips which showed high levels of repetition. This way there was a larger number of translation examples to show the subjects during the prospective phase, and more data to work with when trying to establish reusability of the subtitles in different contexts. Even though technology such as the Trados Translator's Workbench can provide quantitative data on the contents of a corpus very quickly, it is necessary to manually go through the data, in order to find out exactly where the repetitions occur (relative to the clips), and we did this using a colour-coding scheme. In the first Harry Potter film we marked in yellow repetitions that occurred only within this film (internal repetitions). We marked in red those segments that represented external repetitions only. And we marked in green, those segments that were repeated both internally to the first Harry Potter film, as well as externally in the rest of the HPC, and/or the LOTRC and/or the MGC (identified as 100% matches in Table 1 above). We used Microsoft Word as our editing environment. This allowed us to group together all repeated segments within a corpus (using Word's Sort function), and thus identify exactly which segments accounted for the repetitions counted by the Trados's Analyse tool. Microsoft Word also gave us a convenient way of colour-coding segments. Once the coding was done, if we selected a clip which had only yellow markings, there would be a chance that the internal repetitions were actually only in the selected clip, and therefore the repetition information gathered from the data would be redundant, as the test data (current clip) is never included with the training data (the current TM), and there would be no match saved in the training corpus, so the EBMT system would be unable to provide a previously saved translation. If, however, the clip chosen had green markings, it would not matter if the repetition occurred in the clip itself, as the green marking indicates that there are more occurrences of this segment elsewhere in the other corpora, and hence in the training data, allowing for the repetition to be found by the EBMT system. If a segment is colour-coded red, it means it does not appear in the first Harry Potter film, and therefore it would never be the case that a red segment would appear twice in the same clip.

Based on this information we selected the ten most 'colourful' clips which provided us with various examples of internal and external repetitions as well as 100% matches. For the ten movie clips internal repetitions are calculated by comparing each clip with an empty TM; external repetitions are calculated by comparing each clip with a TM compiled of the other nine clips; and the 100% matches are calculated by comparing each input clip with three different TMs, slightly modified versions of corpora A, B and C: corpus A minus the Harry Potter input clip (**corpus AM**), corpus B minus the Harry Potter input clip (**corpus BM**), and corpus C minus the Harry Potter input clip (**corpus CM**) (the input clip is removed from the TMs given that test data is never included with training data). The HPC minus the Harry Potter input clip in these corpora is represented as HPCM. Tables 2 & 3 give results for the internal repetitions and external repetitions in each clip. Table 4 provides results for the 100% matches between each clip and the TM used in the analyses. When automatically generating the subtitles to be put on movie clips for the evaluation sessions, we ran our EBMT system three times, once using each of the

[7] See Armstrong et al., 2006 for reasons why we decided to choose movie clips from a Harry Potter film.

three corpora, to see if both increased corpus size as well as a good mix of film specific subtitles (HPC), fantasy genre specific subtitles (LOTRC), as well as general language subtitles (MGC) would improve output.

Clip No.	Internal Repetitions
1	0
2	3
3	0
4	1
5	4
6	0
7	0
8	0
9	1
10	0

Table 2: Number of internal repetitions per movie clip

Clip No.	External Repetitions
1	0
2	1
3	2
4	1
5	1
6	1
7	0
8	1
9	0
10	1

Table 3: Number of external repetitions spread across all ten movie clips

	100% Matches		
Clip No.	**HPCM (corpus AM)**	**HPCM + LOTRC (corpus BM)**	**HPCM + LOTRC + MGC (corpus CM)**
1	6	6	7
2	7	11	18
3	7	8	11
4	8	11	15
5	13	17	25
6	3	3	5
7	9	10	17
8	3	4	6
9	6	10	12
10	4	4	10

Table 4: Number of 100% matches between each movie clip and the three TMs used for the study

Table 4 shows the 100% matches between the movie clips and each of the corpora used with the EBMT system. We carried out this analysis to see if we would detect more repetitions and matches if we increased the corpus size, and also introduced subtitles from genres other than fantasy. In each case, the number of 100% matches did increase, highlighting the fact that if you increase your training data, you are more likely to increase your repetition rates. This quantitative data, however, says nothing about the linguistic content of the corpus, and the increase in repetitions may have no bearing on the kind of quality we require to increase the acceptability levels of the subtitles. The next section describes the qualitative study we carried out to take this analysis a step further. We locate each of the repetitions from the three corpora, and present these subtitles to three subjects, who subjectively rate each subtitle.

4.3. Qualitative human evaluation

For this phase, three native German speakers with similar backgrounds in language and linguistics were asked to give their opinions on the reusability of translations of repeated subtitles in our selected movie clips. We used the ten movie clips from *Harry Potter and the Philosopher's Stone* described above. On average there are 25 subtitles per clip, with a mixture of one and two-line subtitles. Of these subtitles approximately a quarter are repeated segments within each clip (colour-coded yellow, red or green). Each participant was presented with two sets of booklets: the first set of ten booklets (Clips 1-10) contained information relating to each of the ten clips, including the context in which the clip is set, the original English subtitle, and the speaker of the subtitle. Where there were repeated source text subtitles, three translations were provided, each chosen by the EBMT system depending on which of the three corpora was used. Subjects were asked to indicate whether or not the subtitle chosen by the EBMT system was deemed acceptable or not. If the subtitle was not repeated within the corpora, no EBMT subtitle was provided. The second set of ten booklets (Options 1-10) contained alternative translations for the repeated segments, where these alternatives, although extracted from the corpora had not been chosen by the EBMT system.

The session was structured in the following way. Firstly the participants were asked to read a pre-experiment briefing and to sign the briefing if they were in agreement with the structure. By doing so they also agreed to the session being recorded on cassette tape, to capture any additional comments they may have made during the session. Next the participants were asked to look at the Clip 1 booklet, read the context given for the clip, and to go through each of the EBMT chosen subtitles in order, noting whether or not the subtitle is acceptable in this context. The three options in response to whether a segment is acceptable were 'yes', 'no' and 'don't know'. After each subtitle the options booklet relating to the clip booklet was referred to if there were any additional translations offered in the corpora. In some cases there were no alternative translations offered. In the cases where there were other options to choose from, the subjects were once again asked to indicate whether or not they thought the subtitle was acceptable, and once again three choices were offered. The session continued in this manner for all ten clip booklets and option booklets.

4.4. Preliminary results

The results presented in this paper will focus only on the ten movie clips and the translations selected by the EBMT system, and not mention any results from the options booklets, due to space constraints. After looking at all the results from each subject in relation to the clips we can make some observations. The results given below correspond to each set of 3 subtitles:

- There are 62 sets of EBMT subtitles (186 subtitles in total).
- There were 4 sets (6%) where there was no agreement between the three subjects. In the rest of the cases at least two subjects agreed on the same acceptability response (either yes, no or don't know).
- There were 24 sets (39%) where all subjects agreed that all the translations offered were acceptable ('yes' response).
- There were 5 sets (8%) where all subjects agreed that all the translations offered were unacceptable ('no' response).

Table 5 below shows the distribution of responses across the three corpora per subtitle. These results demonstrate that even though the data generated by Trados presented in table 4 indicated the highest level of repetition occurred in corpus CM, the qualitative data shows the subtitles generated using corpus BM received the highest number of 'yes' responses and the lowest number of 'no' responses in relation to the acceptability of the subtitle. The subtitles generated by this corpus also received a slightly higher number of 'don't know' responses.

Response	**HPCM (corpus AM)**	**HPCM +LOTR (corpus BM)**	**HPCM + LOTR + MGC (corpus CM)**
Yes	124	126	120
No	53	49	57
Don't know	9	11	9

Table 5: Subject responses: numbers of yes, no, don't know according to the corpus

We need to mention here that in many of the cases, the same translation was chosen by the EBMT system from all three corpora. Therefore, in the cases where the subjects all gave a 'yes' response for all three translated subtitles chosen by the EBMT system, they were, in fact, approving of the same translation three times. The same applies to a number of the 'no' responses.

5. Discussion and Conclusion

The first aim of the prospective phase of the study reported on here was to get an initial impression of **repetition** levels within our corpora, in order to identify video clips whose subtitles contained enough instances of (internally or externally) repeated segments to make them useful as test cases in a study of the reusability of translations. We were also interested here in producing translated subtitles that drew on both smaller, more homogeneous corpora, and larger, more heterogeneous corpora, in order to be able to relate these corpus variables to the acceptability of subtitles that would eventually be evaluated by subjects (our three 'prospective' evaluators, mentioned above, and the subjects who would eventually view the videoclips 'retrospectively'). The Analyse tool in Trados Translator's Workbench proved very useful in this regard, as it allowed us to ascertain quickly the extent to which source-language segments recurred (in the exact same form) in individual clips, and in the three corpora investigated here. The Sort function in Word also proved useful in allowing us to move on to an analysis of *which* segments were behind the repetition statistics generated by the Analyse tool. The second aim of this phase was to see if we could achieve inter-annotator agreement among three subjects regarding the *reusability* of translations of repeated subtitles within the corpus. Where there are high levels of inter-subjective agreement that translations chosen by the EBMT system are reusable in their new context, we would expect that the translated subtitles in question would also ultimately be deemed acceptable by viewers in our second 'retrospective' evaluation phase. As indicated under 4.4. above, in the vast majority of cases at least two of our annotators give the same response on the acceptability of a translated subtitle in its new context, and in a further 72 cases of translated subtitles (or 24 sets), all three annotators agree. Although our sample of three analysts is very small, the level of inter-subjective agreement among them is high enough for us to make informal predictions about how the subtitles chosen by the EBMT system will ultimately be judged by viewers.

There are, of course, limitations to our research. We focus initially, for example, only on exactly repeated subtitles within our corpus, and have not yet considered levels of repetition below the level of subtitle. Given that we are conducting our research within an EBMT environment such sub-subtitle (analogous to sub-sentential) repetition and reusability is, of course, of interest. Our more focused analysis is also limited to just ten short movie clips, but this can be justified by the detail of our qualitative analysis, and the fact that we are establishing a methodology for the analysis of such clips. In any event, we see our research as making a contribution to the literature on human-oriented qualitative evaluations of the reusability of existing translation.

References

Armstrong, S. (2007) Using EBMT to Produce Foreign Language Subtitles. MSc Thesis, Dublin City University, Dublin, Ireland.

Armstrong, S., C. Caffrey, M. Flanagan, D. Kenny, M. O'Hagan and A. Way (2006)'Leading by Example: Automatic Translation of Subtitles via EBMT?' *Perspectives*, 14(3):163–84.

Bowker, L. (2002) *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Bowker, L. (2005) 'Productivity vs Quality? A pilot study on the impact of translation memory systems', *Localisation Focus*, 4(1): 13–20.

Carl, M. and A. Way, (eds) (2003) *Recent Advances in Example-based Machine Translation* Dordrecht: Kluwer Academic.

Carl, M. and A. Way, (guest eds) (2005) *Machine Translation. Special Issue: Example-based Machine Translation* volume 19, nos. 3–4.

Heyn, M. (1998) 'Translation Memories: Insights and Prospects' in L. Bowker , M. Cronin, D. Kenny and J. Pearson (eds) (1998) *Unity in Diversity? Current Trends in Translation Studies* Manchester: St. Jerome, 123–36.

Kenny, D. (2007) 'Translation memories and parallel corpora: challenges for the translation trainer' in D. Kenny and K. Ryou (eds) *Across Boundaries: International Perspectives on Translation* Newcastle-upon-Tyne: Cambridge Scholars Publishing, 192–208.

López Circuelos, A. (2003) 'Una defensa crítica de las memorias de traducción', *Panace* Vol. IV, n.º 12: 180–82.

Nedoma, A. and J. Nedoma (2004) 'Problems with CAT tools related to translations into Central and Eastern European Languages' in *Translating and the Computer 26*, Aslib Proceedings, London: IMI/Aslib.

Reinke, U. (2004) *Translation Memories. Systeme – Konzepte – Linguistische Optimierung* Frankfurt am Main: Peter Lang. Europäischer Verlag der Wissenschaften.

Somers, H. (2003) 'Translation memory systems' in Harold Somers (ed) *Computers and Translation. A translator's guide* Amsterdam/Philadelphia: John Benjamins Publishing Company, 31–47.

Whyman, E. and H. Somers (1999) 'Evaluation Metrics for a Translation Memory System' *Software-Practice and Experience* 29(14): 1265–84.