[*Translating and the Computer 25*, November 2003 [London: Aslib, 2003]

Can language technology respond to the subtitler's dilemma? - A preliminary study

Minako O'Hagan
School of Applied Language and Intercultural Studies
Dublin City University
Glasnevin
Dublin 9, Ireland.

minako.ohagan@dcu.ie

1. Background

This study was prompted by three independent developments. One is a widely publicized criticism of the Japanese subtitles of the film *The Lord of the Rings: The Fellowship of the Ring* - the first episode in the J.R.R.Tolkien trilogy. Shortly after the film's release in Japan in March 2002, complaints about the quality of the subtitles from Japanese fans led to petitions to the film's Japanese distributor and the director, Peter Jackson (O'Hagan, 2003a). The fans claimed that the Japanese translation reflected the subtitler's lack of appreciation of the world of the *Lord of the Rings* (LOTR)[1] as created by Tolkien. It was subsequently revealed that the entire subtitling task had been completed in just one week by a single subtitler who had never read the book. This highlighted the fact that in the film industry subtitlers are often at the mercy of the market-driven approach becoming prevalent in the entertainment industry as a whole. For example, even the translation of the title of the film is not determined by the

[1] For a comprehensive background context, see the Japanese site: http://sa.sakura.ne.jp/~straydog/bard/toda.html

subtitler alone, at least in Japan (Bailey, 1999). A similar squeeze on language support seems common in other sectors which increasingly rely on a range of interlingual transfer that has come to be known collectively as GILT (globalization, internationalization, localization and translation). In the case of the computer software industry, the release date of a product is strictly dictated by marketing strategies and least by localization (translation) requirements and the publisher of the software has the final control over the translation issues. Given the increasing time pressure and digitization of film production, it seems inevitable that subtitling may become subject to the localization model that is being applied to a wide range of products today. One of the main differences between software localization and film subtitling as practiced today is the lack of technology application to the translation process in subtitling. Localization is heavily augmented by technology from the beginning to the end whereas subtitling translation largely relies on human talent alone. The question is then if there is scope for applying language technology to subtitling.

Another impetus for this study came from the recent "surprise language" experiments funded by the US Defense Advanced Research Projects Agency (DARPA) in March and June 2003 (Dean, 2003; Oard, et al, 2003). These experiments were designed to test the ability to build automatic translation tools within a very short time-frame in emergency scenarios such as a terrorist attack, leading to intelligence gathering and monitoring of information in a given language. For its dry run session in March, the researchers were asked to build a system in two weeks from the time of the announcement of a surprise language which turned out to be Cebuano. One of the key issues put to the test was how existing technologies and techniques can be deployed to devise a usable system when unforeseen needs arise. In the context of subtitling, could a usable computer-assisted translation (CAT) environment be quickly assembled for a subtitler who was given an assignment with very short notice? In association with the DARPA experiments, this study seeks to address the scope of the currently available technology without a major breakthrough for subtitle translation.

The third factor which has a link to this study is the phenomenon of "fan-subs". This term refers to the subtitles produced unofficially by Japanese animation (anime) fans for non-Japanese speaking viewers outside Japan (Nornes, 1999). Despite their questionable legal status, fan-sub groups have been in existence since the late 1980s, circulating among fans not-for-profit subtitles for anime which are not officially released abroad, or when there are lengthy delays before the release of officially subtitled versions[2]. The quality of fan-subs

[2] A similar reaction manifested recently with J.K. Rolling's Harry Porter books for which numerous pirate translations appeared ahead of the scheduled release of the official translations primarily because

varies since the task is undertaken by anime enthusiasts, generally untrained in subtitling, on the strength of their familiarity with the genre. The fan-sub model relates to the present study in two ways; one is the impact of the genre-knowledge in subtitling and the other unorthodox and yet potentially effective approaches taken by the former in comparison with conventional subtitling (Nornes, ibid). Despite varying levels of linguistic knowledge, fan-subs seem to sometimes display virtuoso solutions to language transfer problems because of their understanding of the genre, including such factors as intertexuality and audience expectations. These issues have a particular relevance to the LOTR case. Fan-subs also provide an insight into possible developments of future subtitling (Nornes, ibid; O'Hagan, 2003b).

In view of the demands of the entertainment industry as a whole, including localization of animation titles and computer games, subtitling is likely to come under even more severe time pressure in the future. New media such as DVD could drive the trend further. Moreover, the simultaneous shipment (simship) model applied to software localization could be the inevitable consequence of the progressive globalization of the entertainment industry whereby the release of the original product and the localized versions coincide. With the simship model, the *Lord of the Rings* film comes out simultaneously in different cities of the world in different languages. Behind this is the impetus coming from the widespread use of digital technology for production and delivery of audio-visual content which seems set to bring the localization industry and subtitling closer together. The mere fact that a single DVD can hold multilingual subtitles and dubbed tracks[3] begs for rapid availability of translations. Will this mean the adoption of the localization model to subtitling, which is likely to lead to increased use of language technology?

These background factors seem to justify the need to investigate the scope of language technology applications in the translation process in subtitling.

2. Subtitling and Language Technology

There is a dearth of information technology applications for subtitling translation as is the case for literary translation, for which any kind of automation is generally considered unlikely in the foreseeable future. Literary translation, as opposed to technical translation, tends to draw on more extensive extra-linguistic knowledge and contextual information. This in turn

non-English speaking fans could not tolerate the delay.

[3] A single DVD has a capacity up to 17 GB (as compared to 700MB of a CD) and is able to embed subtitles in up to 32 languages or up to 4 dubbed versions (Karamitroglou, 1999).

leads to translation based on dynamic, rather than formal equivalence (Nida, 1964), limiting the benefits gained by use of language technology. In addition, subtitlers are constrained by strict limitations on the length of the translation due to the screen space available and also the speed at which the viewer can read the text. This means a radical reduction in text while minimizing the loss of critical information. For example, Japanese subtitles for films in cinemas have to be less than 26 characters based on the 4 character-per-second rule (Nornes, 1999). Subtitling also imposes unique restrictions on translation in terms of the need for cohesion between the text and the image being shown to the viewer, while the conversion from a spoken dialogue to a written text (inter-semiotic transfer) in itself poses a challenge. For Japanese subtitles, there is a further restriction on the use of certain classes of ideographs[4] in consideration for the varying education levels of the viewers, whereas ideographs are useful in economizing the number of characters. These characteristics of subtitle translation seem to mitigate against the application of technologies such as machine translation (MT) or translation memory (TM) where the former benefits text which tolerates literal translation and the latter text with a high proportion of internal and external repetition[5]. On the surface, subtitling does not appear to render itself readily to either technology.

However, unlike other forms of literary translation, subtitles are mostly used for dialogue and tend to contain shorter sentence chunks than the typical prose of literary text. Shorter sentences are considered to be one of the MT-friendly factors. Another factor in favour of language technology applications is the increasing digitization of audio-visual content, which, in theory, facilitates the integration of technology into the subtitle translation process. In addition, the Internet is allowing subtitles of commercial feature films to be widely accessible via the Web. The availability of subtitles in electronic form may motivate further research into corpus-based text analysis of subtitles[6] and into the feasibility of language technology applications in this particular type of language transfer. Furthermore, there are a number of MT systems in existence producing subtitles (captions) on the fly for TV programmes, including an early example by NHK (Vasconcellos, 1993). The recent MUSA (Multilingual Subtitling of multimediA content) project funded by the European Union is aimed at building an MT system for producing captions. The system is designed to convert audio streams into

[4] Japanese Kanji characters (ideographs) are classified from easy to difficult and the Japanese education system follows this classification for Kanji curriculum. When it is necessary to use difficult Kanji, subtitlers may use a linguistic device called "furigana" which allows the phonetic reading of the character to be shown just above it as superscript.

[5] "internal repetition" refers to repetitions within a single document whereas "external repetition" refers to repetitions between separate documents (such as an old and new version of computer manual).

[6] There are as yet no substantial parallel corpora available in the field of screen translation (private communication with Yves Gambier, October, 2003)

text transcriptions, then produce draft translations in two language pairs and finally rephrase them to produce a shorter text as subtitles[7]. Wisdom gained from the existing systems and ongoing projects will also facilitate the drive to language technology applications to subtitle translation.

Set against these backgrounds, this study took a first step into scoping the potential usefulness of language technology applications in producing subtitle translation.

3. Human error: what goes wrong with human subtitles

Subtitle errors can be broadly categorized into mechanical errors e.g. time-coding and translation errors. In this paper, we are only interested in the latter category which in turn can be grouped into:

1. straight translation errors
2. inappropriate reduction of information in the process of shortening the output
3. proofreading errors

Coming back to the highly publicized case of the Japanese subtitles of the first episode of LOTR, the following shows a summary of the fans' contentions based on some of their postings that appeared in various Web sites[8]. It is clear that the Internet facilitated collective criticism in an unprecedented manner by allowing fans to exchange their views and post concrete examples.

- discrepancies of proper nouns between the film's subtitles and the translation versions from the original book (e.g. names such as Strider and Gollum) with which the fans were familiar.
- excessive dynamic translation allegedly due to the limitation on the length of translation, some of which the fans considered had deviated too far from the original critical meaning (assumed to be due to the subtitler's ignorance of the key point of the story)
- straight translation errors which were considered avoidable if the subtitler had been familiar with the Tolkien's story (or possibly had more time)
- wrong and inconsistent register - (e.g. indication of the subordinate relationship of Sam to Frodo)
- the title of the film being different from the title of the book in its translations (as explained

[7] See www.esat.kuleuven.ac.be/-spch/cgi-bin/projects.cgi?MUSA
[8] see http://herbs.tsukaeru.jp/english top.html [The Fellowship of the Rings Distorted in Japanese subtitles]

earlier, this was not solely the subtitler's decision)

The subtitler in question was quite unperturbed by these criticisms, pointing out a number of specific requirements of subtitling; (1) subtitles are different from print media (such as books) and some words or phrases which may look good in print do not necessarily come across well on screen and (2) subtitling is not the same as translation, in that extracting essential information conveyed within the limited number of words becomes critical. Despite the local film distributor's initial noncommittal response and the original subtitler's lack of remorse, the campaign against the LOTR Japanese subtitles seems to have had a positive impact on the DVD version as some of the errors in the cinema release were apparently corrected. Furthermore, an unprecedented measure for quality assurance was implemented by the local distributor for the second episode of LOTR in that the Japanese subtitles were to be back translated into English for approval from the film's global distributor (O'Hagan, 2003a).

It has to be pointed out that the film did extremely well in the Japanese market and that there were many viewers who enjoyed the film and did not agree with some of the claims made by the fans. This reflects the fragmented nature of cinema audiences (Jackel, 2001). However, the controversy pointed to a number of problematic factors in subtitling translation in Japan:
(1) the major feature films are almost exclusively handled by a very few high-profile subtitlers; (2) the translation process seems entirely dependent on these human talents; and (3) the timeframe is becoming increasingly shorter. Further, there are a number of emerging issues such as (1) the durability and flexibility of DVD will mean that subtitles will be exposed to a greater number of viewers on a longer time span who could scrutinize the translation and (2) with advancing computer technologies subtitling may become increasingly accessible to fans of the film who in turn could, with their own PC, make their own subtitles. In relation to the latter point, one of the noteworthy developments of the LOTR subtitle claims was the fact that some fans had suggested revised subtitles of their own. This makes a rather overt link to the fan-sub model where the genre experts undertake subtitling. Continuously improving technology seems to allow amateur subtitlers to more readily undertake their own subtitling just as fan-subs have quickly shifted from the cumbersome analogue environment of video tapes to the digital world and the Internet both of which made their work easier.

In summary, human subtitle translation is increasingly exposed to public scrutiny facilitated by new modes of communication such as the Internet while the amateur fans may be motivated to create their own version if not satisfied with the official subtitles. In the meantime, the market reality makes the deadlines shorter and shorter for professional

subtitlers to complete the project. The obvious answer seems to lie in exploring technology applications. Two specific aspects can be considered: (1) how can technology speed up the process of professional subtitling and possibly improve quality? and (2) to what extent can technology be used to make usable the subtitles by amateurs, who are genre-experts without formal linguistic or subtitle training? A series of experiments were conducted to probe these aspects.

4. Experiment design

This study set out to investigate whether today's off-the-shelf language technology has a scope to facilitate the translation process for subtitling. It is the first phase of a longer-term research project. The primary objective of this phase was to ascertain if today's language technology has prospects for this type of application. Two scenarios were created in conducting experiments: one scenario explored how a Japanese subtitler with a pressing assignment could quickly assemble useful translation-support data to speed up the translation time and also retain or improve quality, using Translation Memory (TM). This scenario was in analogy with the DARPA experiments where timeliness was of paramount importance while making best use of available technology. The second scenario was to test to what extent amateur translators could succeed in producing usable subtitle translation from English into Japanese by using a free on-line MT system. This was based on the "fan-sub" model where an amateur subtitler whose linguistic knowledge may be weak, tackles the challenge by drawing on genre knowledge and MT output. Another overall question this study intended to address was to respond to the conference theme of user experience of language software applicable to a wide range of user groups. During the course of the research, a range of software was used and brief user comments are included in the conclusion section. Table 1 summarises the software used for this study.

Table 1: Software Products used for the experiments

Function	Name of the product/site	Cost

Subtitle sources • Commercial DVDs • Software to locate available subtitle texts in electronic form on the Internet	• DVD versions of: - Lord of the Rings: The fellowship of the ring; - Lord of the Rings: The Two Towers; - Harry Potter: The Chamber of Secret • Subtitle finder V 1.2 http://www.softpile.com/Utilities/Miscellaneous/Review_16617_index.html	10,000 yen Free on the Internet
Subtitle extractions • DVD VOB file extractor • Subtitle extractor from VOB files with their time codes as a text file	• SmartRipper 2.4.1 http://www.dawnload.net/video_software/dvd_rippers/smartripper.cfm • SubRip 1.17 http://www.divx-digest.eom/software/subrip.html	Free on the Internet
OCR (Japanese & English) • To turn Japanese and English text into machine-readable form	• One Touch OCR for Excel and Word http://softplaza.biglobe.ne.jp/shop/aisoft/ocr	Free for one month trial
Translation Memory	• Trados TM 5	
Online MT • Online translation site powered by Amikai translation engine	• Excite Honyaku http://www.excite.co.jp/world/text	Free

The following sections describe each step of the experimental process in more detail.

4.1 Scenario 1: Could TM help subtitling?

In this experiment a scenario was set whereby a Japanese subtitler undertook to translate the second episode of LOTR within one week. The main objectives were to see what constitutes a useful set of data in relation to language technology application and to test the effectiveness of a Translation Memory (TM) system. Originally the possibility of creating a translation memory based on the entire set of Tolkien's LOTR books in English and their Japanese translation had been contemplated. However, the non-availability of electronic text for the Japanese version (while one version of the original English text was located online) and the implications of the text alignment effort between the English and Japanese text, made this impractical within the timeframe available. Instead, a translation memory was created based on the English and Japanese subtitles of the first episode of LOTR *(The Fellowship of the Ring)* available on DVD. The subtitles were extracted from DVD, aligned and made into a translation memory with the expectation that the memory will contain some matches when compared with the text of the second episode of LOTR *(The Two Towers)* which the subtitler was to translate.

The process started with downloading the necessary software to extract (rip) the subtitles

from DVD[9]. *SmartRipper* was used to extract VOB files[10] from DVD, which in turn was processed by another software *(SubRip)* specifically to isolate subtitles. This turned out to be a treacherous exercise particularly with Japanese scripts as subtitles are encoded as image, thus requiring OCR (Optical Character Recognition) process. The OCR component which came with *SubRip* was not optimized for Japanese characters. This meant that the character recognition component needed to be done manually by initially re-typing nearly every single Japanese character (although the system started to learn as it went along). To extract and obtain the 15,007-character Japanese subtitles (see Table 2), it took over 6 hours' effort even excluding the proofreading and the post-OCR correction process. So, to prepare a clean text this part represented roughly a day's work in total. For Roman alphabet-based scripts this should have been a near automatic process, although it was noted that even with the English subtitles OCR was not perfect with certain characters such as the letter "I" [as in I am] and "l" [as in lost]. The subtitle extraction software had come with an editing tool to assist the correction task. The total time spent on editing and cleaning after the extraction of the 8,606-word English subtitles (see Table 2) was just over 2 hours - a fraction of the time needed for the Japanese subtitles. Once bilingual subtitle texts were extracted, they were aligned using Trados *WinAlign* and a translation memory was created. The aligning process also proved extremely time-consuming, taking up to 5 hours as the automatic alignment results needed to be substantially corrected as is explained below. The next step was to extract the English subtitles from the second episode of LOTR to use this as the source text to produce Japanese subtitles. It was noted that as the familiarity with the software increased, the time it took me to complete this process became shorter.

4.2 Scenario 2: To what extent can MT be useful to provide the basis for an amateur subtitler to produce usable subtitles?

In this scenario, an amateur subtitler who is not a trained professional but is familiar with the story was to translate the first episode of LOTR *(The Fellowship of the Ring)* using MT output as a basis to boost the translation process. Here, cost and speed considerations were of paramount importance as in the case of fan-subs and the quality of the resulting subtitles. A popular English-Japanese translation site *Excite Honyaku* which is freely available online was chosen because the Amikai translation engine used by this site is often judged best among similar English/Japanese systems (Amikai, 2003; Takahashi, 2000). The English

[9] The author was concerned about the legality of this exercise, but the subtitle ripping software had a clause stating that this activity is legal so long as the original DVD was privately owned and the resulting data are for private use.

[10] VOB (DVD Video Object) files contain multiplexed MPEG-2 video streams, audio streams and subtitle streams.

subtitles for LOTR extracted from the DVD version were put through this site to obtain a Japanese translation. The whole process was over within 1 hour even though the MT site had restrictions on the total number of input words (a maximum of 4000 English words at a time), thus having to break the text into smaller chunks for processing. It was noted that when the text volume was closer to the limit, the translation speed seemed to markedly go down. Once the translation was done, the main task was to consider the quality in view of producing the final translations suitable for subtitles. Due to the fact that I could not find a suitable experiment participant assuming the role of a typical fan-sub creator, the final part of formulating fan-subs on the basis of MT output was not possible to carry out. The experiment therefore resorted to quantifying clearly unusable translations produced by MT.

4.3 Comparative data

Although scanning of entire parallel texts in English and Japanese from the LOTR book and its translation was impractical for the time frame available for this experiment, in order to obtain supplementary data for comparative purposes, the first chapter of book one of LOTR and its published Japanese translation were scanned. In order to do this, Japanese OCR (this software included English OCR as well) was downloaded from the Internet. The particular product was chosen as it was freely available for one-month trial period and also the product was fully integrated into *Microsoft Word* which I was using for this experiment. It was noted that direct scanning from the pages of the Japanese book (which used a thin paper, thus showing the reflection of the print of the reverse side of the page and also with a relatively small font) did not perform very well. The book pages first required to be photocopied and enlarged before they could be scanned and go through the OCR process. It turned out that the English book had similar problems in terms of thinness of the page and the size of the font, thus the same procedure was needed.

The scanning and OCR process for Japanese text of 31,850 characters took over 5 hours with the 9,821 word English text taking about 2 hours[11]. By the time the text was proofread and cleaned, the whole process took a day's work. This clearly demonstrated the impracticality of scanning the whole parallel texts from the English and the Japanese books. Other comparative data were drawn from the bilingual subtitles of Harry Potter's second film *The Chamber of Secrets.* In this case, the English subtitles were available via a Web site, located by the software *Subtitle Finder,* where as the Japanese subtitles could not be found

[11] Kenny (1999) quotes, based on her experience of scanning and OCR English text, a rate of 50,000 words per hour. The significant difference may have been partly due to the quality of the text resolution and/or the performance of the particular OCR.

online and were extracted using the same method as described above from the DVD version, again taking a similar amount of time as the case of LOTR Japanese subtitles.

5. Analysis

5.1 Scenario with Translation Memory

With the first scenario using TM, the data capture process obviously took too long for the one week deadline. As explained in 4.1 it took nearly 2 days to create a translation memory based on the first episode of LOTR, including the extraction of the subtitles and their clean-up as well as alignment process. The creation of a translation memory based on the original LOTR book and its Japanese translation would have used up the best part of the time available (one week) although the resultant memory may have been very useful. The experience clearly indicated that the fulfilment of this scenario so critically hinged on the availability of electronic text. The issue concerning obtaining text in machine readable form is well documented by researchers conducting corpus-based studies as in Kenny (2001), including the copyright issue of using the entire book in electronic form.

The experiment clearly demonstrated that the translation memory created based on the bilingual subtitles from the first episode was of little use for the translation of the second episode of LOTR. As illustrated in Fig. 1, the analysis given by Trados TM indicated 113 repetitions in terms of segments (191 repetitions in terms of words) and varying levels of matches (fuzzy matches) above 50%, together making up 4% in relation to the whole text. Apart from certain recurring proper names (Frodo, Gandalf, etc) and short phrases, such as "come on", "go", etc the translation recycling idea did not work. Even where TM recognized a matching segment, the translation recalled from the translation memory was sometimes totally useless. This was due to the fact that some target language subtitles extracted from the memory were dynamic translations and were not applicable in a different context even though the source sentence may have been exactly the same. It was also noteworthy that there were 210 "unconnected source segments" as indicated in the alignment statistics (see Figure 2), reflecting the nature of subtitling work where substantial reduction of the information is often inevitable. The lack of symmetry between the source and the target segments was conspicuous and the automatic alignment process had left many source segments unconnected to the target, proving that the re-alignment process was extremely time-consuming. With the statistics indicating only 2% for the matches above 50% and 2% repetitions, the limited applicability of TM was clear and the actual translation attempt of the second episode using TM was abandoned half-way through.

5.2 Scenario with Machine Translation

For the second scenario, a comparison was made between the human-produced LOTR Japanese subtitles and those by MT firstly to obtain quantitative data on mechanical differences such as sentence length. As a comparison, a similar analysis was made using the data from *Harry Potter: the Chamber of Secrets* (HP) to check the representativeness of the above results obtained from LOTR.

As illustrated in Table 2, the length of the human subtitles for LOTR (15,007) was 55% that of the MT output (27,211) in terms of number of characters. This compared with 68% for HP (Human translation 23,341 versus MT's 34,333 characters). These clearly pointed to the curtailing process conducted by human subtitlers when they formulate subtitles. By comparison, with the translation of the book, the result was reversed in that the length of human translation was 113% of the MT output, showing that human translation was longer than the translation produced by MT.

Table 2: Summary of MT Experiment Results

	LOTR (The Fellowship of the Ring)	**LOTR book: Chapter 1 of Book 1)**	**HP (The Chamber of Secrets)**
English Text No. of English words	8606	9821	10383
Human Translation into Japanese No. of Japanese characters	15007	31850	23341
Machine Translation into Japanese No of Japanese characters	27211	28024	34333
MT to HT wordage ratio	0.55	1.13	0.68
Unintelligible MT sentences	317(20%)	637 (63%)	961 (51%)
Average length of an HT sentence	11 characters	31 characters	13 characters
Average length of an MT sentence	17 characters	28 characters	18 characters

In this exercise, the quality of MT outputs needed to be judged from the point of view of an amateur subtitler (who was likely to have less linguistic knowledge than a professional subtitler) trying to create subtitles on the basis of MT. As explained earlier, a rather crude method was resorted to by counting the number of unintelligible sentences. The criterion for the judgement was simply whether or not the given MT sentence made any sense on its own. Intelligibility was thus primarily gauged subjectively by the author. While MT output of the

LOTR subtitles looked promising (80% were intelligible), it was not the case with HP subtitles where just under 50% were intelligible. This in part seemed to be caused by the difficulty for MT in its treatment of lexical items in Harry Potter where Latin and French derived words were used. This contradicted the expectation in that Harry Potter which targeted a younger audience turned out to be more complex than the Lord of the Rings in the context of MT.

In terms of quality, MT outputs of the text from the LOTR book were worse than the other two cases, with unintelligible sentences making up 63%. One of the most notable differences between the inputs from the book and the subtitles was the shorter sentence length of the latter, which appeared to support the common wisdom that the shorter the sentence, the better the MT output. This was, in fact, one of the user guide suggestions provided by the translation site used[12].

The next question sought with this scenario was whether MT outputs provide any useful basis on which the non-translator genre-expert could produce high quality subtitles. This became a question of post-editing as well as matching of text against the image and was something the present study could not test as explained earlier. However, it is hoped that the next phase of the study will accommodate this dimension.

6. Conclusions

This study was prompted by the emerging need to meet increasing market demand on subtitling by using language technology. It was also motivated by the possibility of non-professional subtitling undertaken by fans, as exemplified by the fan-sub phenomenon for the Japanese animation genre, together with the negative reaction by some Japanese viewers to the LOTR subtitles. Assuming that professional subtitling will face ever reducing turnaround times subtitles will probably have to incorporate some form of translation aid to prevent severe declines in quality. The aim of this study was to investigate the suitability of language technology to subtitle translation and to suggest which tools might be useful and under which circumstances.

The first and foremost problem identified was the lack of availability of digital versions of published books and subtitles in the source and the translation. While new software tools such as *Subtitle Finder* seem to yield a surprising range of subtitles available in electronic form in a wide variety of languages, it was noticeable that Japanese language subtitles were

[12] see http://www.excite.co.jp/world/help/text

hard to come by. Whether this is due to particular problems with Japan is unclear, but the encoding issue with double-byte characters may have played a part as well as the attitudes by copyright holders. The same applies to the availability of digital text of books - while a version of the English text of LOTR book was found on the Internet, its Japanese translations could not be found. This experience brought home the pleas often made by researchers conducting corpus-based studies who seek the commitment of players in the book publishing to allow the basic resources for their research to be available. These constitute peripheral and yet critical research issues[13].

Coming back to the specific results of the two scenarios, a number of aspects were highlighted by the experiments. The first scenario tested the applicability of TM in a context of serialized films where the subtitler could leverage the subtitles produced for the previous episode. While the result indicated that the actual matches between the translation memory and the new source text were minimal and the repetition within the new source text was also negligible, the experiment pointed to what could be useful. As mentioned earlier, the exact same sentence may occur but due to different contexts, different renditions are often necessary. This seemed to suggest the usefulness of the concordance function whereby the subtitler could see how the same word or phrase may have been used in what context. Kenny (2001) discusses the relevance of concordances to literary translation for the similar reason. This in turn points to the very limitation of today's TM tools which restrict the type of documents which is suitable for TM (i.e. repetitive text internally or externally). Whereas a memory tool with sophisticated bilingual concordance functions which can draw on a wide selection of user-specified parallel corpora (in the context of this study, subtitles of the films from the same genre) could prove to be useful. If this can be combined with a translation memory created from the original book on which the film is based and its translation the effectiveness of TM tool will multiply. This experiment thus provided the direction for the next phase of the study.

The second scenario attempted to test the usability of freely available MT for creating subtitles mainly by non-professional subtitlers. The experiment demonstrated that one aspect of suitability of subtitle text to MT processing - the shorter length of average sentences. My observation was that a large proportion of the raw MT outputs of the LOTR English subtitles could be usable as a pure aid to non-English speaking viewers under certain

[13] Comparative Subtitling Project by the European Association for Studies in Screen Translation called upon subtitling companies and broadcasters to participate in the collection of subtitling data and managed to involve 48 institutions from 25 countries. This exemplifies the importance of cooperation from the players in the field (see http://www.esist.org/currentprojects.html).

circumstances. Given that the MT used for this particular experiment was a low-end product (in terms of cost) and it had not been in any way customised to the specific purpose of the experiment, the results provided a clear scope for potential. In particular, between more closely related languages the results could be markedly better. However, the comparative data from the Harry Potter film showed that the MT output was significantly worse than LOTR. This in turn pointed to the direction of further research in that many other types of films should be tested to compare MT performance and also a different range of MT should be used.

The third component of this study was to make brief comments on the user experience of the range of software used in this experiment. The key resource which made this study possible in the first place was the new and increasingly popular medium DVD to deliver audio-visual content. Its capacity and flexibility allowed the possibility of extracting subtitles, using specialized, but relatively easily available software such as *SmartRipper* and *SubRip.* As with many software products, new software does require the user to climb a fairly steep learning curve, as I experienced myself. However, the most difficult part in this experiment was dealing with Japanese language scripts for the OCR component. By comparison, the Japanese/English OCR product *One Touch OCR* used to scan the text from the books was relatively straightforward with the only aspect I noticed being the importance of maintaining straight orientation of the page for more accurate first-time scanning results. This was applicable to both Japanese and English language OCR. As regards to processing text via online MT, there were no issues either in terms of character encoding or procedures.

7. Future plans

On the basis of the preliminary case study presented above, a number of further studies are being considered which will explore the following aspects:

- Similar tests for both scenarios involving different genre of audio-visual content such as computer games and animation titles
- The use of terminology management/dictionary tools in Translation Memory also in Machine Translation systems
- The use of a range of concordances to test the usefulness combined with Translation Memory
- Audience-based tests with Machine Translation raw outputs as subtitles for end-user feedback
- Experiments involving a fan-sub practitioner to test how MT output could be effectively used as the basis of formulating the final subtitles

Although the following item is not directly feeding on the above experiment results, a further extension to the study could investigate:

- Effective forms of subtitle presentations, using different fonts, colours or various line positions on the screen to place subtitles, which are not considered as the norm.

The author would like to acknowledge the valuable comments on an earlier draft received from Dr Dorothy Kenny.

Figure 1: Trados Analysis of the Second Episode of LOTR against a memory created based on the English/Japanese subtitles of First Episode of LOTR

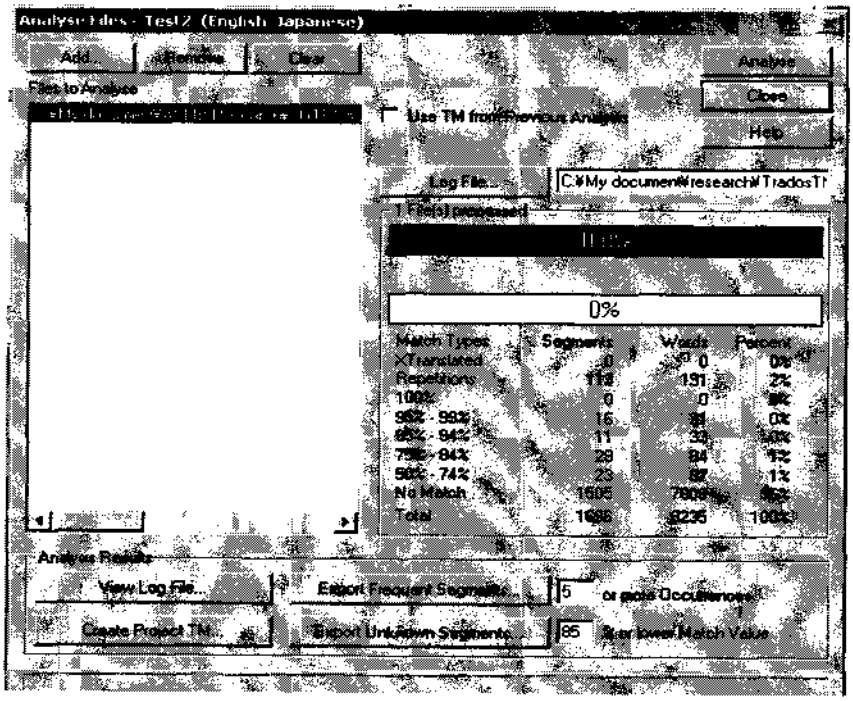
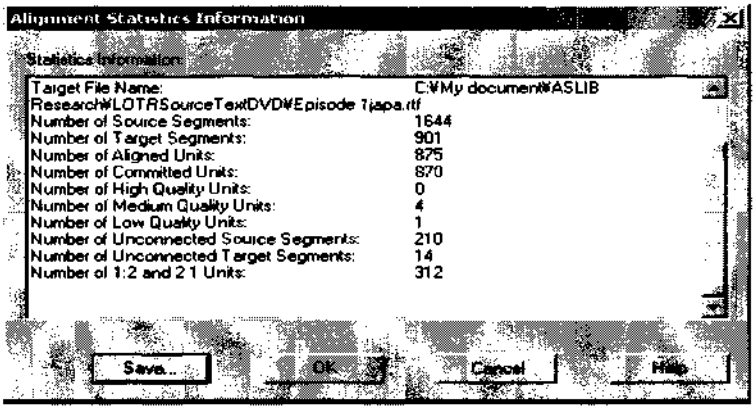


Figure 2: Trados Alignment statistics between English and Japanese subtitles



Bibliography:

Amikai: Best of Breed MT Engine (2003). *LISA Newsletter,* 1.5. Available at: www.lisa.org/archive_domain/newsletter/2003/1.5/index.html.

Bailey, J. (January, 1999). Retitling: Guddo and Baddo: why some movie names sound like gibberish. *Asia Week magazine.* Available at: www.asiaweek.com/asiaweek/99/0122/feat13.html

Dean, K. (2003). Pick a Language, Any Language. Wired News. Available at www.wired.com/new/print/0.1294.59093.00.html.

Jackel,A. (2001). Shooting in English? Myth or Necessity? In Gambier, Y & Gottlieb, H. (eds), *(Multi) Media Translation.* Amsterdam, Philadelphia: John Benjamins Publishing Company.

Karamitroglou, F. (1999), Audiovisual Translation at the Dawn of the Digital Age: Prospects and Potentials. In: *Translation Journal* 3 (3). Online at www.accurapid.com/iournal/09av.htm

Kenny, D. (2001). *Lexis and Creativity in Translation: A corpus-based study.* Manchester: St. Jerome Publishing.

Nida, E. (1964). *Towards a Science of Translating.* Leiden: E.J.Brill.

Nornes, A.M. (1999). For An Abusive Subtitling. In *Film Quarterly,* 52,3:17-34.

Oard, D. et al (2003). Desperately seeking Cebuano. In *Proceedings of Human Language Technology North American Chapter of the Association for Computational Linguistics (HLT-NAACL).*

O'Hagan, M (2003a). Middle Earth Poses Challenges to Japanese Subtitling. *LISA Newsletter 1.5.*
Available at: www.lisa.org/archive_domain/newsletter/2003/1.5/index.html.

O'Hagan, M (2003b). "Into Hypertranslation: Abusive Translation or Something for the Future?" In *International Japanese English Translators Conference IJET-14 Proceedings.* Tokyo: Japan Association of Translators.

Takahashi, N (2000). Jissen Hikaku [Practical Trials] - Online Translation Service. *Internet Watch.* Available at:
www.watch.impress.co.jp/internet/www/article/2000/1030/trans.htm

Vasconcellos, M. (1993). The present state of machine translation usage technology or: How do I use thee? Let me count the ways. In *MT Summit IV Proceedings.* Tokyo: MT Summit IV Secretariat, Asia-Pacific Association for Machine Translation.