Drawing a Route Map of Making a Small Domain-Specific Parallel Corpus for Translators and Beyond

Xiaotian Guo
New Vision Languages
Llandudno
North Wales, UK
garlickfred@gmail.com

Abstract

After years of development of corpus technologies, it has become obvious that translators can benefit directly from the achievements of this field. However, it seems that corpus advancement has not been deployed accordingly by translators to aid their translation. As a corpus linguist and translator myself, I believe that when corpus technologies are made attractive and simple enough and when they do feel a strong need and burning desire to make their own corpus to assist their translation, then application of such technology will gradually become part of a translator's life, just as other computer-assisted translation (CAT) tools have done over the past ten years or so. This paper attempts to make a demonstration as to how easy it can be to DIY a corpus by building a small domain-specific corpus between English and Chinese in the field of financial services. The making of such a corpus has been summarised into three simple steps: 1) Collection of raw parallel language data; 2) Alignment; 3) Segmentation and Annotation. It is hoped that other users of corpora including translation trainers, language teachers and students will also find this presentation informative and beneficial.

1. Introduction

At the Fourth International Conference on Corpus Use and Learning to Translate held in Alicante in 2015, there was a strongly felt concern that even though corpus technologies are available and abundant for translators to use, surveys had shown that a very small number of translators are currently deploying this new technology to aid their work in translation (Frérot, 2015). Actually, this phenomenon was addressed several year ago by Bernardini and Castagnoli (2008). In my opinion, there might be at least two reasons that have contributed to this undesirable outcome. One reason is that translators do not find it absolutely necessary to take the trouble to build and use a corpus of their own because they are short of time in learning the method of making a corpus and putting their hands on it while they can use the Google search engine instead (also see Bernardini, 2015). And the other reason is that the corpus technology has not been presented as a user-friendly and efficient means of assistance to translators yet, most probably due to its seemingly daunting complexity involved in it. Even though awareness has been raised of the usefulness of corpus technology for translators (i.e. Bowker, 2002; Quah, 2006), it seems that there is a need to simplify and systemise the process of construction with clearer instructions and perhaps more importantly examples while the technology details are presented to translators. As a corpus linguist and translator myself, I believe that when corpus technologies are made attractive and simple enough and when they do feel a strong need to make their own corpus to assist their translation, then they will gradually integrate the application of corpus technology into their translators work, just as other computer-assisted translation (CAT) tools have been in the past few decades. This paper attempts to make a demonstration through DIYing a corpus by building a small domain-specific corpus in the broad field of financial services, with English as the source and Chinese as the translation. The significance of this demonstration should be applicable to other domains. The paper is to introduce the necessary stages to take if a small domain specific corpus is to be build, i.e. a) collection of raw parallel texts, b) alignment of the collected parallel texts, and c) segmentation to the Chinese texts and annotation to both English and

Proceedings of the 38th Conference Translating and the Computer, pages 88–99,
London, UK, November 17-18, 2016. ©2016 AsLing

Chinese. The issues of translation quality control and copyright in building a corpus by using texts collected from the web are addressed in brief. At the centre of my target readers are professional translators who are new to corpora, but I also hope that others such as translation trainees and teachers, language learners and teachers could find this paper informative and relevant to their interests.

2. Collection of raw parallel language data

The making of this corpus is primarily to demonstrate how it can aid translators in a specific domain, therefore, the size of the corpus does not have be very big. This section describes my method in collecting English-Chinese parallel texts in financial services. English is the source language and simplified Chinese (as mainly used in the mainland of China) is the translation. Collection of raw language data is carried out through a combination of pure manual collection and semi-automatic collection assisted by a web-crawling programme called Wget[1].

2.1 Pure manual collection

Unlike collecting monolingual texts, starting to collect parallel texts can be somewhat tricky. One way to be adopted is to find out some websites containing texts of the relevant languages in a searching engine such as Google. In order for Google to search from the internet some candidate websites, a few key words can be tried in both the source language English (in this case) and the translation language Chinese (in this case) in Google. For example, some of the English terms and their translation in Chinese can be typed into the Google search engine for a preliminary search such as *financial services*, *foreign exchange*, *trading*, *platform*, *risks*, *terms and conditions*, and the equivalent Chinese translation of these terms. When you have the retrieved websites by Google, you may select and open some potential websites for detailed look for possible parallel texts. Sometimes a few different sets of key words need to be fed into Google before serious candidate websites can be captured. Double quotation marks can be used to search multiple key words in a string so that Google concentrates on the exact phrasing instead of a combination of the individual words in the string. Sometimes you need to try different sets of key words for several times before there appear some websites containing the right information needed. Some other detailed advice for this purpose is available online for users reference[2]. Translators who find it necessary to build a domain specific corpus of their own normally would have known some websites containing potential parallel texts while they are working on their translation tasks. Therefore, these websites could serve as a starting point for the collection of raw parallel texts.

Manually collected parallel texts from the internet through keying bilingual key words or terms in a search engine are normally mixed in languages in one document and cannot be directly passed for alignment programmes to carry out alignment because most current alignment programmes only accept the input of the parallel data in the way that the source language is in one file and the translation in another (see Section 3 for alignment). Separating Chinese from English and saving the two texts originally in one document into two individual files takes two stages. The first stage is to separate the mixed texts with a uniform marker called delimiter to facilitate the recognition of the boundary of the two differently coded languages by the next software. In this process, regular expressions can be used to separate the two languages properly. It may take several rounds to conduct the separation completely due to the various situations of the mixture of the Chinese language. Other programmes using

[1] The manual is available at https://www.gnu.org/software/wget/manual/wget.html, last accessed on September 30, 2016.

[2] http://www.multilingual.ch/Help_find_parallel_texts_using_google_extended_help.htm

regular expressions such as Replace Pioneer can also be used for this purpose[3]. The second stage is to use Excel to input the delimited document with parallel texts and separate English and Chinese in two different columns which enables us to save the source file in one document and the translation in another by copy and paste, once the boundary of each string of source language is defined in relation to its translation.

2.2 The semi-automatic approach

Apart from manually collecting parallel texts, there are programmes available to assist this purpose. Since there must be human involvement to some degree and at some point, I would call this type of collection semi-automatic approach. There are many programmes used in the acquisition of parallel texts such as Parallel Text Miner (Nie, 1999), STRAND (Resnik, 2003), Bilingual Internet Text Search (Ma and Liberman, 1999), the Parallel Text Identification System (Chen et al. 2004), and Wget. This research is going to use Wget[4], a well-known and widely used freeware web-crawler used for downloading data from the internet. Once the programme is downloaded and installed onto your computer, you will need to read the manual before you test it for the first time. The programme works in command prompt, you will need some basic knowledge in handling file paths and commands to work with it. The Manual contains all the various and possible commands for different tasks but you do not have to know all of them in order to assign a task to it. Due to the importance of collecting raw data for the construction of the corpus, a detailed and clear guidance is provided. However, due to the space it takes, this part is included in Appendix A.

Knowing the names of as many websites as you need to search is crucial to the semi-automatic approach of data collection. You might need to start from the websites you know from the translation tasks you complete and gradually build up more websites relevant to your purpose of corpus construction. Some users prefer to make a base word list to feed the programme to maximise the output of the search in the internet (see Wang and Su, 2009), especially when the corpus to be constructed is meant to be big for purposes such as machine translation (for example, Koehn, 2005), dictionary compilation (for example, Héja, 2010), and term extraction (for example, Baisa et al, 2015). Since parallel texts are stored in different directories, normally one file of text contains one language only rather than two languages mixed together. This means two files of different languages can be directly passed for the aligner programme for alignment. But in cases when the two languages are mixed in one file, they would need to be separated by a uniform delimiter to enter the next stage (refer to Section 2.1 for details). A research by Zhang et al (2006) in their automatic acquisition of Chinese-English parallel corpus from the website uses a pre-defined strings indicating English and Chinese versions in several possibilities used in a website having several language versions (see Appendix B for details). This first-hand experience in parallel data collection could lend some ideas for the collection of parallel texts as well.

No matter the data collection is through a pure manual approach or a semi-automatic approach, it is necessary and worthwhile to have a scan of the texts to have a preliminary idea about the quality of the translation. Texts found with too many errors in translations even at a glance should be abandoned at this stage and not allowed to enter the next stage of corpus construction. The manual method sounds slow but can be pragmatic for building a small corpus. What is more important, the quality of the translation can be supervised more easily. However, the semi-automatic approach is faster downloading raw data for scrutiny obviously. A combination of the two approaches might suit most professional translators if the quality of translation is essential and overrides the quantity of the corpus.

[3] Visit http://www.mind-pioneer.com/ to download the programme, last accessed on September 29, 2016.

[4] https://www.gnu.org/software/wget/

3. Alignment of parallel texts

At the end of Section 2, collected parallel texts of the source language English and its translation Chinese of a particular title (or theme or topic) are stored in two separate files. For parallel texts to be converted to translation memories, each and every language unit (segment) has to be matched perfectly well, be it a full sentence or a clause or a phrase or even an individual word such as those in titles, headings, list items, and table cells. The processing of putting each segment in one line and different segments in different lines is called alignment and a programme conducting such a task is called an aligner. Normally, aligners use various parameters (also called anchors) for aligning segment pairs such as segment length and punctuation marks, which work well to certain language pairs, especially languages in a close family, and certain text types. Ideally, all the segments in the source should be matched by the same number of sentences in the translation, and all the punctuation marks in the two languages are exclusively identical. Actually, due to the differences between languages and cultures, it is difficult or even impossible for a translation to reach that level of equivalence. Take the English and Chinese language pair for example, a translator may have to use several simpler and shorter sentences in the Chinese translation to match a long and complex sentence in the source language English. And a question mark at the end of the English greeting "How are you?" would need a different punctuation mark in the Chinese translation "你好!"[5]. As a result, it is not surprising that an aligner would be able to do part of the job of automatic alignment but not the entire job. Human involvement would be a must to ensure each and every segment is correctly matched. There are many programmes that can carry out the task of alignment, for example, the Uplug by Tiedemann (2000) and of course the various versions of CAT tools including SDL Trados. In this study SDL Trados Studio 2014 will be used to demonstrate how this could be done. It takes a series of stages to complete the process of alignment, i.e. a) creating a translation memory (TM), b) introducing the two separate parallel text files into the Studio, c) correcting the alignment carried out by the aligner, and d) importing the segment pairs into the TM. To save space these details are included in Appendix C. Users who are familiar with the Studio need no further and more detailed explanations but for those who are new to the Studio and for those who have never used the function of aligning documents, they can refer to the manual of the Studio for other details. There are also tutorial videos on youtube introducing tips and tricks in alignment and TM import[6]. In the same manner, you can have your finely collected and selected parallel texts aligned and converted into your own TMs for reference. Apart from Studio 2014, users are recommended to try other aligners and see which one works better in a particular aspect and which one works better in other aspects. In my comparative investigation into Studio 2014 and another CAT tool developed in China called Xue-Ren (literally Snowman) CAT, I found that Xue-Ren CAT does a much better job in alignment than Studio 2014. For example, it is easier to split and merge both the translation and the source segments. And above all, it is possible to edit the source segments in Xue-Ren CAT which is a huge advantage if the user finds it necessary to do some editing on the source, for example to delete a serial number at the beginning of a segment which is not in the translation segment.

Till now, translators shall be able to benefit from corpus technology because the aligned segment pairs can be technically converted to a TM and aid translation in CAT programmes.

[5] Sadly, in the current day Mandarin teaching overseas, the daily greeting of Chinese equivalent to "How are you?" tends to be taught as "你好吗？ " - a literal translation copy of "How are you?" Actually, in real Chinese conversations, this is seldom used.

[6] Refer to https://www.youtube.com/watch?v=EKlkZEjkLL8E for details of the process, last accessed on September 24, 2016.

However, for those who wish to make a full use of the corpus (for those beyond translators such as translator trainers, language teachers and students) there is one step further to take to benefit more from the construction of the corpus.

4. Segmentation and annotation

After the data of your parallel texts has been properly collected and finely selected, there are two options to use it. As illustrated in Section 3, it could be aligned and then converted into a TM. It could also enter a different stage to be explored for different purposes. For example, how a parallel corpus can be used to examine the features of the languages under study in a particular syntactic structure, how a parallel corpus can be used for translator training for the awareness of language differences in different aspects of linguistic parameters, and how it can benefit terminologists in term retrieval, and even how it can aid language teaching and learning.

As is well known, the Chinese language is a distant language from English and differs from it in many different ways. Unlike the English language, Chinese characters are not separated by spaces and therefore the boundaries between Chinese characters are not defined. Unless the characters are separated semantically (no matter it is one character or two characters or multiple characters), there is no way for a programme to identify and process a given order. The process of separating Chinese characters is called segmentation. For the collected parallel texts to become a usable corpus, the Chinese files need to be processed by segmentation.

4.1 Segmentation

Segmentation technology is an important part of natural language processing to the Chinese language. There are programmes available for public use, among which ICTCLAS is probably the most often used for the segmentation of Chinese characters. The ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) (Zhang et al., 2003) can be downloaded free and installed in your computer.

By now the entirety of the parallel texts collected and selected into the final folder on your computer could be treated as a corpus, but it is a raw corpus. Simple queries such as individual words or strings of characters can be searched by a programme which is able to process two files at the same time to display the source language segment with the translation segment side by side. Such kind of programme is called a concordancer in corpus linguistics. Since processing a parallel corpus (two files and two languages at the same time) is different from processing a monolingual corpus, it takes a different concordancer to work on a parallel corpus. Currently, the most often cited concordancer for a parallel corpus is probably ParaConc by Barlow (2003). Of course, professional translators may also use the TM function of their CAT tools to examine the parallel texts in them. In a raw and untagged corpus, it is possible to examine part of the performance of individual words or characters, but it is not possible to examine the languages from a broader view such as how many verbs often occur in a particular syntactic structure in one of the languages (or even in the two languages) and what they are. For more complicated and specific queries, a corpus would need to be marked by certain tags to enable bulk identification and processing by a concordancer. Currently, the most often used and available tagging is part-of-speech (POS) tagging (also called annotation), which means each and every word would be tagged by the POS it belongs to in the context.

4.2 Annotation

Apart from the function of segmentation, ICTCLAS can also carry out POS tagging to the Chinese language. For the English language, there are many POS taggers available, but the most often cited and used one is probably CLAWS by developers at Lancaster. Created in the 1980's and developed over the last few decades, this programme requires a license to use which means a cost would be involved. After POS tagging, each word will be attached with a special coding of POS. Take the sentence "This is a black cat." for example, after POS tagging by CLAWS, the original sentence would be something like "This_DD1 is_VBZ a_AT1 black_JJ cat_NN1 ._." in which the POS of each word is encoded accordingly. Due to the differences of the two languages, the two tagging systems would have different parts of speech. However, this should not stop users from searching those parts-of-speech shared by the two languages. After the parallel texts are POS tagged, users may consult them for a much wider range of queries. For example, the belief held by some scholars can be tested that the Chinese language prefers the use of verbs whereas the English language favours the use of nouns. Supported by some analysis of the data generated, translation trainers, language teachers and students could have a better understanding of the two languages.

5. Using parallel corpora for other purposes

Apart from the values and potentials to professional translators, a parallel corpus can be very useful for other purposes such as natural language processing and machine translation, translator training, term construction and dictionary compilation, and language teaching and learning, and translation studies, to name only a few. Due to the space of this paper, I will concentrate on the use of parallel corpora in translator training and language teaching and learning.

5.1 Use of parallel corpora in translator training

Together with other types of corpora such as monolingual corpora, comparable corpora and learner corpora, parallel corpora can be used for translator training thanks to the advantages of bilingual and aligned parallel texts in a parallel corpus. How certain lexical items and syntactic structures are represented can always be examined in the other language through a bilingual concordancer. In translator training, the teacher can assign tasks to translator trainees to reflect some of the difficulties in the process of translation for this specific language pair. For example, due to the differences of the English and Chinese languages, long English sentences, especially with clauses are very often split into shorter clauses in the Chinese translation. Translation teachers could show theirs students how professional translators deal with this type of syntactic complexity through real translation examples. To further the training in this aspect, teachers could ask their students to translate long English sentences into shorter Chinese sentences in the translation. Even a further step, as a supplement, teachers may also ask their students to combine shorter Chinese sentences into suitable long English sentences for some practice of translation from Chinese to English. Of course, instead of the teacher dominating the classroom, translator trainees could have an active role to play in making use of a parallel corpus. While they are doing a translation exercise, they may consult a parallel corpus for a particular expression in the other language when they do not know how to express, or when they simply wish to confirm something about which they are not absolutely certain (see Yepes, 2011, for a few examples of using parallel corpora in translator training and see Zanettin et al, 2003 for some papers focusing on the topic "Corpora in Translator Education"). It can be envisaged that translator trainees would have a better chance of developing not only their translation skills, but also their translation strategies and methodology, if they could observe the data carefully and sum up the implicit rules from individual examples.

5.2 Use of parallel corpora in and language teaching and learning

Language students and translator trainees have something in common in using a parallel corpus for study, although their study purposes are slightly different (the purpose of the former group is for general language acquisition whereas the purpose of the latter group is to get trained to become translators in the future). Therefore language teachers and students could use the strategies and methodology used by translator training in their language classroom. In addtion, there should be more values to explore and cultivate from a parallel corpus for language teachers and students. If we look at the literature, it would be easy to find a massive reservoir of theories and explorations in this aspect (for example, Barlow, 2000; Wang, 2001; Hunston, 2002, to name only a few). However, while we appreciate the usefulness and beauties of corpora, it is important that appropriate corpora should be selected to the right level of students because language study via examining corpora suits students of intermediate and advanced level more than beginners (see St. John, 2001). Using corpora (including parallel corpora) for language teaching and learning is probably the most discussed topic in the field of present day foreign language teaching and learning[7].

As shown above, a parallel corpus can be useful not only to professional translators but also translator trainees and language teachers and students. However, just as roses have thorns, a parallel corpus would have its own problems to be aware of. The next section talks about the issue of translation quality control and copyright awareness.

6. A few tips of caution and some advice

A corpus can be very useful after the construction is complete. However, there are at least problems users need to be cautious about: the translation quality control and the awareness of copyright. They should not assume that all translated texts would be of good quality and high standards simply because the texts have been published online or simply because they are from the website of a well-known company or organisation. Some translations, especially those of commercial agreements, terms and conditions, and guidance to products and services, are provided online for the reference of potential users only. They do not possess the degree of authority and integrity as the source documents have. That is why there are reminding clauses at the end of many legal documents alerting users that where there are conflicts between the original language and the translation the original language would prevail. Therefore, it is recommended that before any serious use of the corpus a check on the quality of the translation is carried out to ensure that the translation has reached the expected standards. Apart from the problem of translation qualities, builders and users of a corpus would face a thorny issue - the copyright issue sooner or later. Some owners of websites have stated explicitly that the information published to the public online is copyright protected and prohibits the use for commercial purposes without permission. However, there are some websites that do not prohibit the use of the information on them as long as users are aware of and responsible for any unexpected outcome that arises from the use of the information on the websites. On the safe side, it is recommended that a certain degree of caution be exerted for a selection of suitable websites and the copyright issue be cleared

[7] For a bibliography on the use of corpora and corpus-based methods in the language learning and teaching context, refer to http://www.corpora4learning.net/resources/bibliography.html, last accessed on September 28, 2016.

before use. Having said this, however, all these problems should not stop corpus users from exploring language use and translation skills in the relevant data strategically and carefully as a methodology.

7. Conclusion

I have chosen to make a parallel corpus because such a methodology can become immediately and directly usable by and useful to translators and others, and therefore more convincing to them, if they have a need or desire to DIY one for themselves. The purpose of this paper is to show that the task of making a small domain specific corpus is not as daunting as some people might think. However, it would be wrong to fall into the belief that the process of completing the task is extremely simple. As has been demonstrated in this paper, making a corpus involves several stages including preliminary planning, collection of raw materials, processing of parallel texts, converting parallel texts to translation memories and even necessary post-editing before use. It would take a considerable amount of time to complete the process, especially for those who have never put their hands on this kind of tasks before. If you wish to make a corpus on one day and to start using it the next day, you would need to have tried quite a few times using the methodology (or a similar one) and to have been quite familiar with the general routine and process shown above as your task requires. The application of corpora is almost ubiquitous thanks to the development of corpus linguistics over the past few decades. The fast growth of computer technologies and the capacity of computer processing and storage means a job taking several days in the past only requires a few minutes even seconds to complete now. Professionals in the broad area of translation should make the fullest use of this invaluable asset to aid their work. It can be envisaged that more and more professional translators would set up their own corpus to aid their work in translation, not because it is useful, not because it looks posh, not because other professionals are using it just like other CAT tools such as translation memories and term bases, but because it is gradually becoming a sort of advantage in getting jobs in the increasingly competitive translation market. It is hoped that this paper has attracted professional translators and the like one step closer to the decision to give it a go to the making of their own corpus after having heard about it for a long time.

Acknowledgements

Thanks are given to Dr. Huaqing Hong at Nanyang Technological University (Singapore) for sharing with me some of the tools used in this research.

References

Baisa, Vít, Barbora Ulipová, and Michal Cukr. 2015. Bilingual Terminology Extraction in Sketch Engine. In *Ninth Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 61–67.

Barlow, Michael. 2000. Parallel Texts in Language Teaching. In Botley, Simon, Anthony McEnery, and Andrew Wilson (eds.) *Mutiligual Corproa in Teaching and Research*. Rodopi, Amsterdam, pages, 106-115.

Barlow, Michael. 2003. *Paraconc: A Concordancer for Parallel Texts*. Athelstan, Houston.

Bernardini, Silvia. 2015. Exploratory Learning in the Translation/Language Classroom: Corpora as Learning Aids. Paper presented in the CULT Conference, Alicante.

Bernardini, Silvia and Sara Castagnoli. 2008. Corpora for Translator Education and Translation Practice. In *Topics in Language Resources for Translation and Localisation*. John Benjamins, Amsterdam, pages 39-55.

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

Chen, Jisong, Rowena Chau, and Chung-Hsing Yeh. 2004. Discovering Parallel Text from the World Wide Web. In *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*, pages 157–161.

Frérot, Cécile. 2015. Corpora and Corpus Technology for Translation Purposes in Professional and Academic Environments. Major Achievements and New Perspectives. Paper presented in the CULT Conference, Alicante.

Héja, Enikö. 2010. Dictionary Building Based on Parallel Corpora and Word Alignment. In Dykstra, Anne and Tanneke Schoonheim (eds): *Proceedings of the XIV. EURALEX International Congress*, pages 341-352.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge University Press, Cambridge.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79-86.

Ma, Xiao-Yi and Mark Liberman. 1999. BITS: A Method for Bilingual Text Search over the Web. In *Proceedings of Machine Translation Summit VII*, pages 538–542.

Nie, Jian-Yun, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74-81.

Quah, Chiew Kin. 2006. *Translation and Technology*. Palgrave Macmillan, Hampshire and New York.

Resnik, Philip and Noah A. Smith. 2003. The Web as a Parallel. In *Corpus Computational Linguistics*, Volume 29, Issue 3, pages 349-380.

St John, Elke. 2001. A Case for Using a Parallel Corpus and Concordancer for Beginners of a Foreign Language. In *Language Learning and Technology*. Volume 5, Number 3, pages 185-203.

Tiedemann, Jörg. 2000. Extracting Phrasal Terms Using Bitext. In *Proceedings of the Workshop on Terminology Resources and Computation*, pages 57-63.

Wang, Dong-Bo, Xin-Ning Su. 2009. Automatic Building of Sentence Level English-Chinese Parallel Corpus. In *New Technology of Library and Information Service*. Issue No. 12, pages 47-51.

Wang, Li-Xun. 2001. Exploring Parallel Concordancing in English and Chinese. In *Language Learning and Technology*, 5(3), pages 174-184.

Yepes, Guadalupe Ruiz. 2011. *Parallel Corpora in Translator Education*. http://www.redit.uma.es/archiv/n7/4.pdf [last accessed September 30, 2016].

Zanettin, Federico, Silvia Bernandini, and Dominic Stewart (eds). 2003. *Corpora in Translation Education*, Routledge, London and New York.

Zhang, Hua-Ping, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 184–187.

Zhang, Yi, Ke Wu, Jian-Feng Gao and Philip Vines. 2006. Automatic Acquisition of Chinese-English Parallel Corpus from the Web. In *Proceedings of ECIR-06*, pages 420-431.

Appendix A: Using Wget to download parallel texts

Suppose you wish for Wget to download the website www.example1.com onto your own computer, the command you need to type onto the command prompt[8] would be C:\wget wget -r www.example1.com in which C:\wget shows the path of wget, wget activates the executable programme called wget, the letter r in -r (meaning recursive) shows how the recursive downloading is going to take place, and the www.example1.com is the website to be downloaded. Wget is able to display the directories of a website in the way the directories are stored in the original website, enabling users to have an idea how the directories and sub-directories are related to each other.

Some websites contain several versions of different languages in different directories and even sub-directories under the same root domain. For example, the English text from www.example2.com/en/ and the Chinese text of translation from www.example2.com/cn share the same root domain www.example2.com. However, some organisations use totally different domains for their different versions of websites in which case you will need to type in two domains to acquire parallel texts.

The raw data withdrawn by Wget may be downloaded into different directories but may also be into the same directory. As a clue, the name of folders such as en (for English) or cn

[8] Assuming that you have saved the Wget directly under the C disk.

(for Chinese) indicates the information under the directory en would be related to the English version of the website whereas the directory cn would be related to the Chinese version. Sometimes, when you open a directory, you find it empty. This might be because the depth of downloading has been reached and Wget cannot dig any deeper than that level of directory. The default depth of downloading is five. Wget is able to process multiple websites in one command. Therefore, it is worthwhile to make a list of website domains as a file and save it in the same directory as the Wget application programme and give a relevant command to download all these websites in one go. Downloading files from the servers of documents takes time depending on how fast your internet connection speed is and how responsive the servers are. It is the advantage of Wget that you may leave the computer working on the task in the background while you work on something else.

Appendix B: A list of pre-defined strings

english
chinese
simplified chinese
chineses implified
traditional chinese
chinese traditional
english version
simplified chinese version
traditional chinese version
英文 英文首页
简体 中文首页
繁體 中文简体
英文版 中文繁體
中文版 简体中文
简体版 简体中文版
+繁體版 繁體中文
英文网站 繁體中文版
中文网站

Appendix C: Using SDL Trados Studio 2014 for parallel text alignment and converting the aligned parallel texts into a TM

1. Creating a translation memory

For the parallel texts to be aligned and afterwards to become a part of your translation memory, open SDL Studio 2014 (henceforth called the Studio) and enter the TM database from the left panel where the user can create a new translation memory. If you wish to add these parallel texts to one of your existing TM, you can skip this process and add it to your alignment project when you are asked by the Studio to do so. It is recommended that a new TM is created at this moment because you may like to do some post-editing for quality assurance purpose. For professional translators who are familiar with the Studio, they can wait for later to create a new TM at a later stage while the parallel files are selected from your computer.

2. Introducing the two separate parallel text files into the Studio

The two separate files can be introduced to the programme from the "align documents" function under the Home menu. In the drop down menu of "Align Documents" choose "Align Single File Pair" and browse to pick your source language file first and then your translation file second so that the source language comes on the left column of the alignment window. If you have created a new TM just now, then click "Add" on the top of the small window to pick the TM. If you have already set up a TM before this new and empty TM , click "add" to pick it in the same way. For those skillful users of Studio who have not created a new TM at the beginning, this is the time to do it. Once the parallel files have been selected and the TM created or added, click "Finish" to complete this process.

3. Correcting the alignment carried out by the aligner

Ideally, the aligner could "understand" the segments of the two languages and put them in the cells in which the right column is exactly the translation of the left column, no more and no less. But in fact this is seldom the case especially when the files are big, due to various reasons including those mentioned above. This is where the human involvement must come in. At this moment, there are four columns in the alignment window. In the left two columns are the segments of the source language and their correspondent serial numbers. In the right columns are the serial numbers of all the segments of the translation, matching the serial numbers of the source segments, and the segments of the translation. There are links between the source segments and the translation segments with different colours (green, yellow and red). The green colour shows that there is a better chance that the segments in the right column is the translation of the segment in the left column. The yellow colour shows less confidence and the red colour even less. Experienced users of the alignment function of the Studio would normally disconnect all the pre-connected links. This is because it is easier to "Disconnect All" first and then find the right pairs than disconnecting and finding individual matching ones. To select one pair of segments, hover your cursor on the serial number of one of the segments and click (the background colour will change as a response) and move cursor to the serial number of the other segment and right click your mouse to disconnect or connect. Sometimes more than one cell needs to be executed and this is can be realised by clicking one of the cells first and then click the next line whilst pressing the control key.

4. Saving the alignment and importing the segment pairs into the TM

When each and every segment is correctly connected to its pair, the alignment project can be saved for future use and what is more important the well aligned segment pairs can be saved into the TM through the function "Import Into Translation Memory" under the Home menu. The TM can be tested like other TMs in translation tasks. It is expected that with the TM added into the translation project, once the source file is introduced to the Studio for translation, the newly created TM will be triggered automatically and each and every source segment will be matched in the translation column by their correspondent segment in the TM.

Appendix D: Splitting English sentence and Chinese sentence into different lines[9]

Download and install "Replace Pioneer" on windows platform to finish following steps.

1. ctrl-o open text file

2. ctrl-h open 'replace' dialogue

* set 'search for pattern' to:

  [A-Za-z][a-zA-Z\W]{15,}

* set 'replace with pattern' to:

  \n$match\n

3. click 'replace', done.

Note1: if you only want to add a # before a Chinese sentence, you can set 'replace with pattern' to: $match# in step2.

Note2: we allow Chinese sentences to contain English word less than 15 letters. The user can change 15 to other number in [A-Za-z][a-zA-Z\W]{15,} in step 2.

To see a screenshot of the Replace Pioneer window, visit http://www.mind-pioneer.com/services/1351_Advanced_search_and_replace.html to see the page.

[9] This content of this appendix is based on the content from http://www.mind-pioneer.com/services/1351_Advanced_search_and_replace.html, last accessed on September 29, 2016.