Designing and Developing a Parallel Corpus Based on Media Subtitles

Zhu Chong

School of Foreign Languages, UESTC

Abstract: Audiovisual media programs, including movies, television programs and documentaries play an important role in everyday life. They are noteworthy as instruments for both entertainment and cross cultural communications. Studies on their translation, especially subtitle translation (or subtitling), however, have not received the attention they deserve. Reasons for this inadequacy might include an academic bias against the importance of audiovisual media translation, as well as the lack of workable translational or parallel corpora in this field. In this paper, the author proposes building a parallel subtitle corpus, i.e., the Multi-Media Subtitle Corpus (MMSC), and illustrates a way to develop it based on the author's work. Finally, potential applications of such a corpus in translation studies are suggested.

Keywords: parallel corpus, subtitling, translation studies, design

1. Corpus and parallel corpus

Corpus has become one of most important tools in various linguistic researches and parallel corpus of all types is especially useful in translation studies, language teaching, etc. (Sinclair, 1991; Baker, 1995, 1998, 1999; Barlow, 2000; Biber et al. 1998; Bowker, 2001, Hunston, 2002; Kennedy, 1998; Laviosa, 1998; Wang Kefei et al. 2005). However, due to the limitation of bilingual or multilingual language materials and the obstacles in their creation, especially text alignment, parallel corpora are still few in number and small in size compared to other types.

In the meantime, a fast growing and easily available material source has been overlooked by many corpus builders, namely, film and television subtitles.

2. Audiovisual translation and subtitles

2.1 Dubbing vs. subtitling

Most media programs, except the mute or wordless ones, which are few in number compared to other types, when transmitted into a foreign country, such as an American movie brought into China, often need to be transferred linguistically. Two ways of achieving this are most common: dubbing and subtitling (Baker 1998:74, 244; Ma Zhengqi 2005: 6-7).

Dubbing is a form of audiovisual translation that involves 'the replacement of the original speech by a voice track which attempts to follow as closely as possible the timing, phrasing and lip movements of the original dialogue' (Luyken et al. 1991:31 qtd. in Baker 1998:74-75). It is much more laborious and time consuming than any other form of screen translation (Baker, 1998:75; Ma Zhengqi, 2005:3). Moreover, dubbing is significantly more expensive (Baker, 1998:75).

On the other hand, subtitling is a method of screen translation that costs much less time and money. In subtitling, the translated lines, referred to as subtitles or captions, are 'presented simultaneously on the screen' while the original sound and voice tracks are intact and played (Baker, 1998:245). Dialogues of a film or TV program are first translated and then each line is given a time code or in Baker's words, the lines are all 'time-cued' (ibid.).

2.2 Subtitling as an increasing trade

Baker (1998: 248) has pointed out that with the advances of teletext technology, subtitling and especially personal subtitling will become 'standards of language transfer' in the future. Since she made that claim, a decade has passed, and it is not only the improvement of teletext technology but also the prevalence of DVDs and other media that have promoted subtitling as an important standard of audiovisual media translation, including television translation predicted by Baker (ibid).

Entertainment groups and broadcast companies like BBC (British Broadcast Corporation), Paramount Entertainment, etc, publish movies and TV programs using DVD as an important medium in copious amounts each year. These DVDs are usually equipped with one or more sound (or voice) tracks and subtitles. For example, the DVDs of American movies issued for sale in mainland China usually have a Chinese voice track and a Chinese subtitle, beside the original English sound tracks and subtitles.

These subtitles, which are furnished as a part of the standard DVD configuration, greatly facilitate the viewing of foreign films and TV programs for the public. The subtitle translators are often professionals hired by the DVD publishers.

Nowadays, besides the professionals, an increasing number of amateur translators are also practicing screen translation and their products are abundant on the Internet. These amateur works differ greatly in quality. Many of them are poor but some are decent or even better than the official releases.

Subtitling as a trade, done by professionals and amateurs, has grown so fast that the sheer volume of translation work done in it each year is stunning. It would be a waste for corpus and especially parallel corpus builders to overlook this large source of language materials.

3. Overall design and structure

To build the MMSC corpus, like the creation of any corpus, involves overall design and planning, data collecting, data encoding and storing, and data processing, etc. (Sinclair, 1991:14-22; Yu Shiwen, 2003:83-91) In the case of creating a parallel corpus, it is necessary to single out one particular step, namely, text alignment in data processing (Yu Shiwen, 2003: 93-99).

3.1 Aims and guiding principles

The overall design of a corpus centers on its building targets and these targets establish the controlling guidelines of the whole project (Sinclair, 1991:15; Yu Shiwen, 2003:83). Therefore, it is necessary to introduce these concepts before the detailed steps of construction are presented.

The major aims and guiding principles of building the MMSC corpus are,

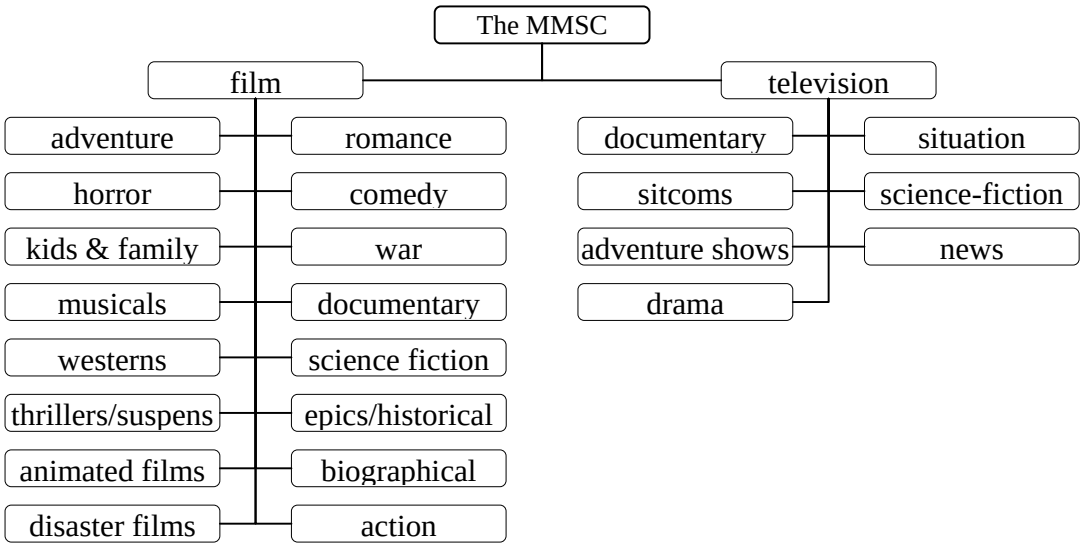
1. Aims: to build a data base for
    a) the theoretical and practical studies on subtitle translation;
    b) facilitating translator training;
    c) facilitating English, especially spoken English teaching and learning;
2. Guiding principles: the subject corpus should:
    a) contain as many texts as possible;
    b) be open and extensible in a way that:
        i. it should supply researchers with transparent and friendly interfaces and standards so that they can submit new data and correct the errors that have been wrongly or carelessly included or input;
        ii. it should be able to be easily embedded into a platform that can be accessed through the World Wide Web;
        iii. it should have interfaces for multi-media data insertion and linking, which enables further efforts to make the corpus a multi-media tool for translation studies and language teaching as well.
    c) only include *suitable* or *eligible* texts in terms of their translation quality.

3.2 Framework

3.2.1 Text categories and sub-categories

As mentioned earlier, the subject corpus mainly contains subtitles from films and television. Thus it should be reasonable to categorize the texts in a way consistent with the classification of these programs.

Therefore, the overall framework of text categorization can be shown by Graph 1:



Graph 1

3.2.2 Database structure and organization

.The adoption of XML as the way of storing data has become more and more popular in the field of corpus creation, such as the parallel corpus designed by researchers from Peking University (Chang Baobao and Bo Xiaojing, 2003). The texts in the MMSC corpus are also stored in XML files. Each subtitle is stored as one XML file following certain rules of the file nomenclature. The query and concordance interface works as a webpage, written in JavaScript, which accesses and operates all the XML files, retrieving and displaying texts in a desired format in the user's web browser.

4. Development

4.1 Text selection and collection

Olohan (2004: 47) argues that the ‘subjectivity of decisions’ concerning which texts to be included into or excluded from a corpus causes a lot of trouble in the creation of a corpus and translation studies based on it. So is the situation in building the MMSC corpus.

Subtitles, as mentioned before, are extremely abundant in number and easily available through different ways, including the Internet, DVD sets, etc. They can be roughly divided into two kinds: those produced and issued by official producers and those construed by amateur subtitle translators or groups. These amateur translators may be students who are movie-fans or professional translators who happen to have the hobby of translating movies or television subtitles for fun or other ends. Many of these amateurs’ works are not very good in quality. However, as mentioned before, some are of very high quality. Thus it would be too hasty to jump to the conclusion that all amateur works are bad and should be neglected entirely.

Therefore, careful selection should be done before the subtitles are taken into the corpus in order to ensure the quality of the corpus itself. The author assumes three ways to decide whether a subtitle should be included into the corpus:

1. whether the subtitle comes from a credible source
2. whether the subtitle enjoys high recommendation on the internet
3. whether the subtitle passes the personal examination of the corpus builders

When a subtitle is agreed qualified and desirable to be included, it then must be retrieved from a certain medium. The author obtains all the subtitles needed from two main sources: DVDs and the Internet.

All the raw materials collected are stored in a folder named ‘Draft’ and need to be further

treated before they can be imported into the corpus.

4.2 Text pre-procession

The texts collected and stored must be further processed before they are ready for alignment. This process includes file type conversion and text formatting for alignment.

The file type conversion step is to convert the subtitle files into plain text files which are stored in a folder named ‘lexical version’.

Then these plain text files are treated according to the following rules:

1. In each subtitle, all the lines with the same time code must be arranged into one line.
2. A time code must be in the same line and followed by the texts under it.
3. All time codes contained in the subtitles should be kept.

After being treated through these steps, all the texts are modified in structure and should be in a format as the following example:

#1|000037.203>000039.671^Do you know who Tyler is?

#2|000041.174>000044.166^Yeah. He's a da waar director.

#3|000044.377>000046.777^What's a da waar?

Example 1

The texts with such an internal structure are stored in a folder named ‘Pre-processed Version’ and ready to be aligned.

4.3 Text alignment

The alignment level in the MMSC is not strictly on sentence level but on ‘tods’ level. Tod is a word coined by the author to represent a single unit of texts in a subtitle, which is included in one single time-stamp or time-cue. In effect, it refers to each caption line or lines shown on the screen at one particular time as one unit.

In Example 1, there are 3 separate tods, each with a particular time stamp in the format like ######.###>######.### which indicates the reference beginning time and end time when the subtitle is printed on the screen.

The whole alignment process is centered on the time-cue information contained in each tod.

If a pair of tods from two subtitles of the same program share the same time code, it would be easy to align the two tods. All that has to be done is to extract the time code, and then extract the texts under it and store them in the appropriate format. This situation between the two tods is called exact match situation.

Unfortunately, the exact match situation is extremely rare in most subtitles. The time cues in most subtitles are actually much more complicated and different time code match relations result in different aligning algorithms and a unified algorithm to govern all the patterns, if possible, is preferred.

4.3.1 Tod match relations

The different tod match relations are described here by examples and pictures:

1. Exact match relation:

Few subtitles have the time codes which are ideal for aligning. In other words, parallel lines share the exactly same time codes. For example:

#1|**000033.566>000036.694**^Dwar	#1|**000033.566>000036.694**^因为海
fed by the vast expanse of the open	洋的辽阔而显得渺小的
ocean	#2|**000036.870>000040.397**^是这个
#2|**000036.870>000040.397**^the	星球上最巨大的生物
biggest animal that has ever lived	舌头有一头大象那么重
on our planet.	

Example 2

2. Random match relations:

a) Self-contained

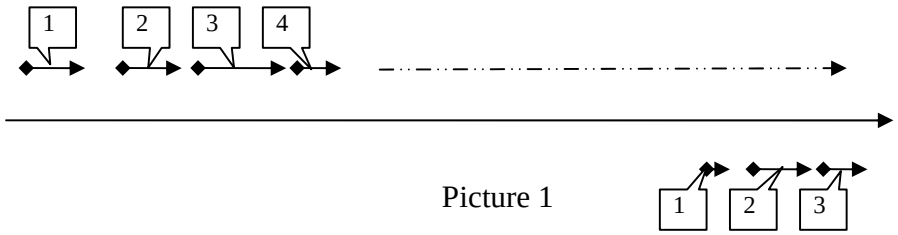
A self-contained line is a line that does not share any part of its time span with any line from the subtitle of the other language.

It is shown by the example and the picture below:

#1|000021.228>000025.927^很久以前在	#1|000159.886>000200.944^- Captain. -
遥远的银河系…	Yes, sir?
#2|000030.571>000034.803^星际大战	#2|000202.521>000204.648^Tell them we
#3|000042.917>000049.413^首部曲：威	wish to board at once.
胁潜伏	#3|000204.757>000207.385^- [Machinery
#4|000051.258>000054.216^银河共和国	Beeping] - With all due respect,
内已爆发动乱	

Example 3

The corresponding relations between these subtitle lines can be roughly reflected in the following picture, where the solid line in the middle stands for the time line and the arrows above the time line are Chinese subtitles and those below are English ones.

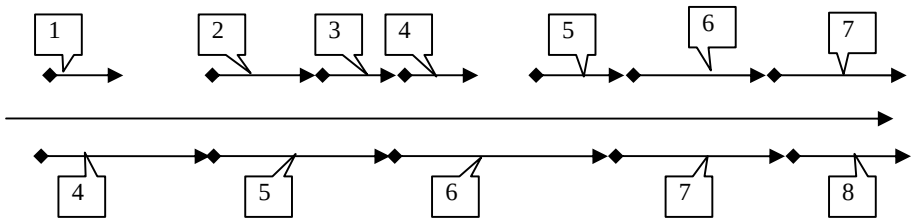


Picture 1

In this particular sample, the first 4 lines from the Chinese subtitle do not share any time span with all the English lines. These solitary lines are defined as self-contained. Self-contained lines do not have parallel counterparts. Therefore they should be separately treated and marked.

b) Contain and Contained

‘Contain’ and ‘contained’ are a pair of complementary relations. When the time span a Chinese line occupies is larger and totally covers the time span of an English line, the Chinese line is said to contain the English line and vice versa. In the following picture, the Chinese line (above the time line in the middle) numbered 1 is contained by the English line (below the time line) numbered 4; the English line numbered 5 contains the Chinese one numbered 2; the English line 6 contains the Chinese 4; the English 7 contains the Chinese 6 and so on.



Picture 2

c) Overlap

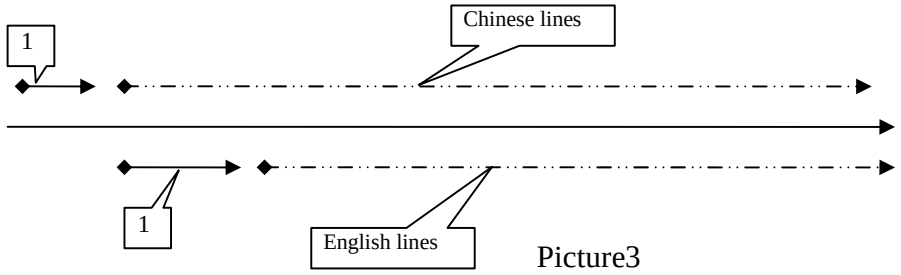
Overlap is the situation where one line does not contain and is not contained by any other lines (of the other language) while it is not self-contained either. In Picture 2, Chinese line 3 overlaps with English line 5; Chinese line 5 overlaps with English line 6; Chinese line 7 overlaps with English lines 7 and 8.

The three kinds of basic random match relations can result in a multiplicity of tod match situations, which must be dealt with in the alignment process.

4.3.2 Tod match situations in alignment

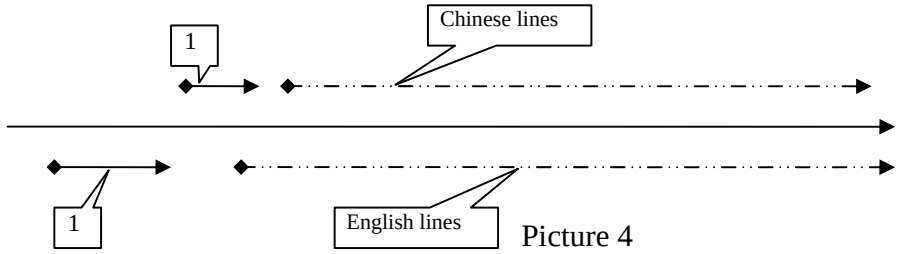
At the beginning of the aligning process, the first Chinese line (the leftmost above the time line) is taken out and examined with the first English line (it does not matter which language is taken first in this discussion). There may be five possible situations:

1. The Chinese line is self-contained as shown in the following picture:



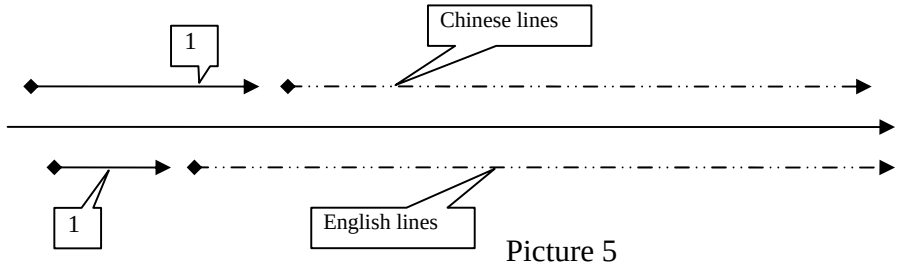
Picture3

2. The English line is self-contained as shown in the following picture



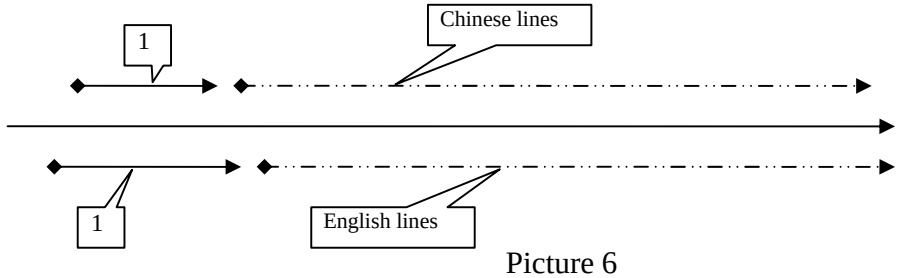
Picture 4

3. The Chinese line contains the English line as shown in the following picture:



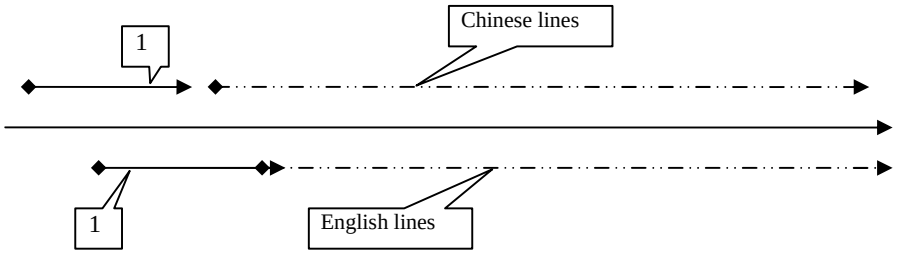
Picture 5

4. The Chinese line is contained by the English line as shown in the following picture:

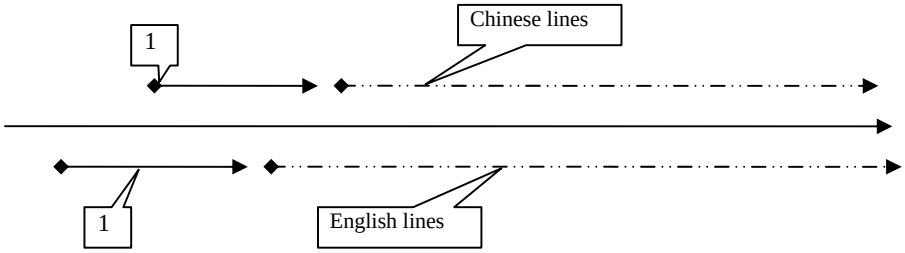


Picture 6

5. The Chinese line overlaps with the English line as shown in the following two pictures:



Picture 7



Picture 8

Actually these are the only possible situations that can happen through the perspective of the three basic match relations. A governing strategy to deal with all these situations is fundamental to text alignment.

4.3.3 The aligning strategies

Several points need to be made clear before the explanation of the strategies:

In the description of the strategies, those words in capitals like ‘MARKED WITH A FLAG’ and ‘DROPPED’ are macro-processes that are to be realized in a certain computer language.

‘The first line’ in this context refers to the current Chinese line or English line which is being examined or processed; ‘the second line’ refers to the line immediately to the right of ‘the first line’ on the time axis and so on.

The strategies are described as follows:

The Governing Strategy: DECIDE the relation between the first Chinese line and the first English line for further treatments.

Strategy 1. When the first Chinese line is self-contained:

The Chinese line is MARKED WITH A FLAG ‘self-contained’ and DROPPED from the aligning process; the second Chinese line is SET as the first (current) Chinese line and examined with the first English line; repeat the Governing Strategy.

Strategy 2. When the first English line is self-contained:

The English line is MARKED WITH A FLAG ‘self-contained’ and DROPPED from the aligning process; the second English line is SET as the first (current) English line and examined with the first Chinese line; repeat the Governing Strategy.

Strategy 3. When the first Chinese line contains the first English line:

DECIDE whether the first Chinese line contains the second English line and:

- If yes, COMBINE the first and the second English lines and SET the combination as the first English line and repeat this strategy (Strategy 3).
- If not, DECIDE whether the first Chinese line overlaps with the second English line:
  - If yes, COMBINE the first and the second English lines and SET the combination as the first English line; repeat the Governing Strategy or go directly to Strategy 5. (The situation has virtually turned into Situation 5 where the Chinese line is ahead of the English line, and thus can be handled by Strategy 5.)
  - If not, consider the first Chinese line and the first English line as parallel and MARK them as ALIGNED; SET the second Chinese and the second English lines as the first Chinese and the first English lines respectively; repeat the Governing Strategy.

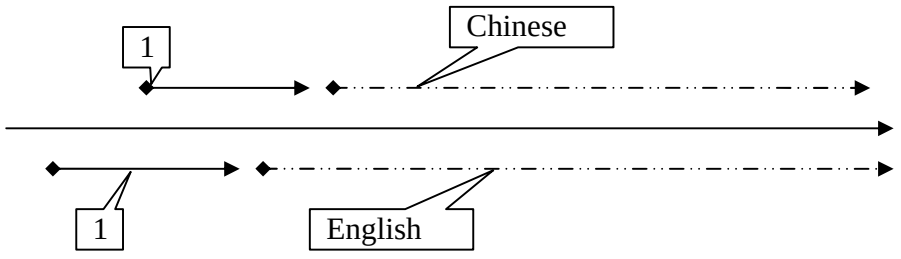
Strategy 4. When the Chinese line is contained by the English line:

DECIDE whether the second Chinese line is contained by the first English line and:

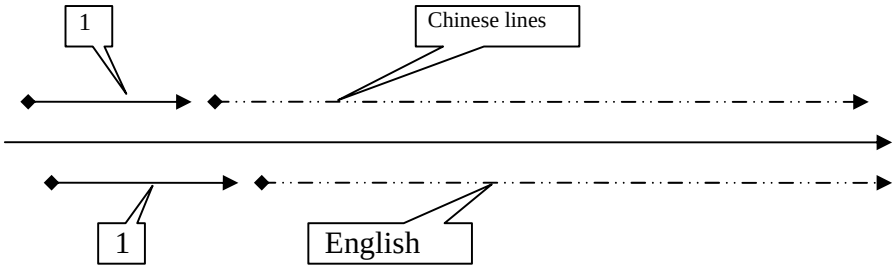
- If yes, COMBINE the first and the second Chinese lines and SET the combination as the first Chinese line and repeat this strategy (Strategy 4).
- If not, DECIDE whether the second Chinese line overlaps with the first English line:
  - If yes, COMBINE the first and the second Chinese lines and SET the combination as the first Chinese line; repeat the Governing Strategy or go directly to Strategy 5. (The situation has virtually turned into Situation 5 where the Chinese line is behind the English line, and thus can be handled by Strategy 5.)
  - If not, consider the first English line and the first Chinese line as parallel and MARK them as ALIGNED; SET the second English and the second Chinese lines as the first English and the first Chinese lines respectively; repeat the Governing Strategy.

Strategy 5. When the Chinese line overlaps with the English line:

There are two kinds of overlap situations: the Chinese line is ahead of the English line or the Chinese line is behind the English line. To say one line is behind the other means that the end of one line is behind the end of the other one. This is shown by Pictures 9 and 10:



The first Chinese line is ahead of the first English line Picture 9



The first Chinese line is behind the first English one

Picture 10

Hence it is necessary to deal with them respectively:

- ❖ Chinese line ahead situation, namely, the first Chinese line is ahead of the first English line:

DECIDE whether the first Chinese line contains the second English line and:

- ➢ If yes, COMBINE the first and the second English lines and SET the combination as the first English line; repeat this strategy (Strategy 5).
- ➢ If not, DECIDE whether the first Chinese line overlaps the second English line and:
  - ✧ If yes, COMBINE the first and the second English lines and SET the combination as the first English line; repeat the Governing Strategy or go directly to Strategy 4. (The situation has virtually turned into Situation 4 and thus can be handled by Strategy 4.)
  - ✧ If not, consider the first Chinese line and the first English line as parallel and MARK them as ALIGNED; SET the second Chinese and the second English lines as the first Chinese and the first English lines respectively; repeat the Governing Strategy.

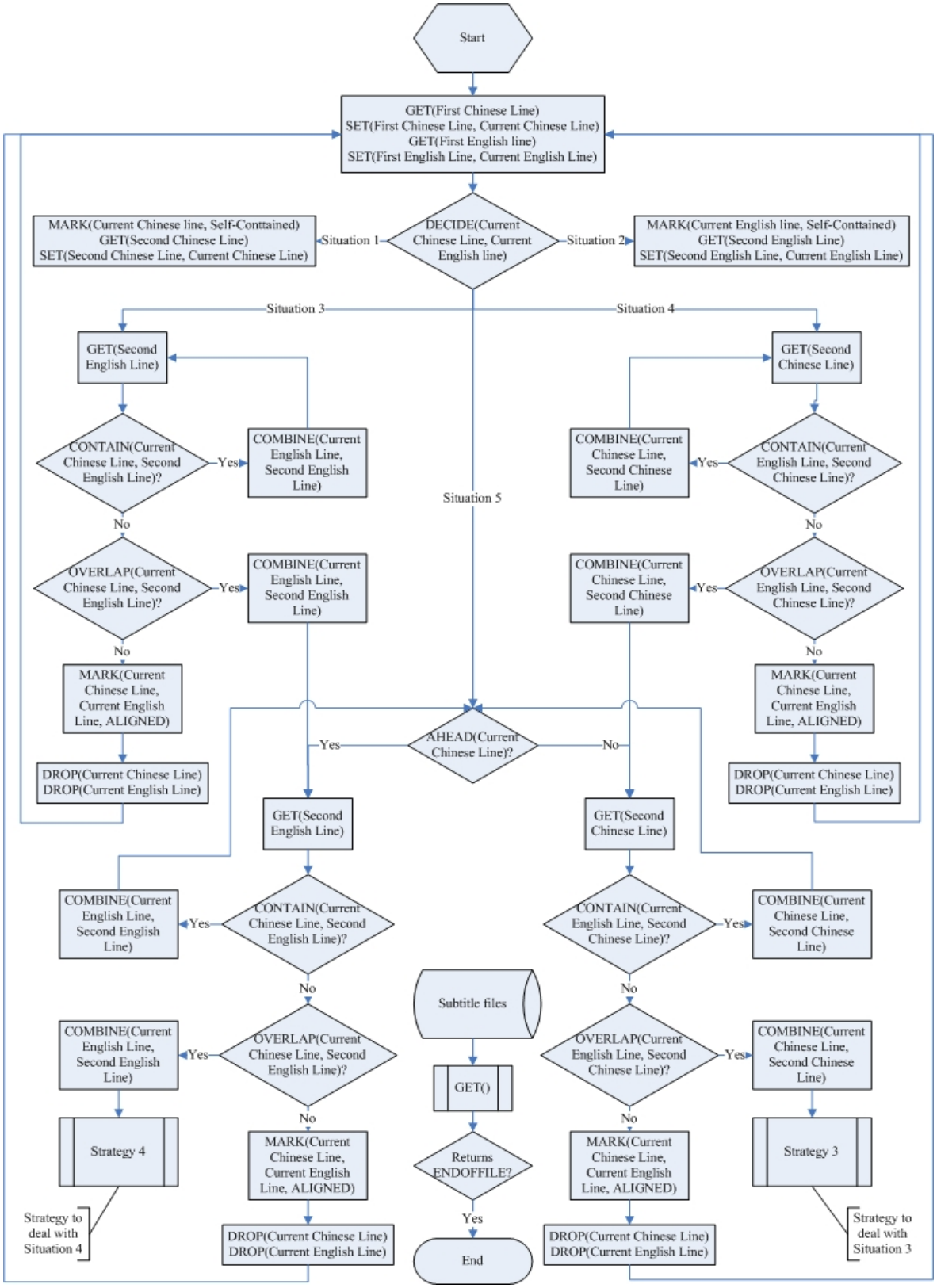
- ❖ Chinese line behind situation, namely, the first Chinese line is behind the first English line:

DECIDE whether the second Chinese line is contained by the first English line and:

- ➢ If yes, COMBINE the first and the second Chinese lines and SET the combination as the first Chinese line; repeat this strategy (Strategy 5).
- ➢ If not, DECIDE whether the second Chinese line overlaps the first English line and:
  - ✧ If yes, COMBINE the first and the second Chinese lines and SET the combination as the first Chinese line; repeat the Governing Strategy. (The situation has virtually turned into Situation 3 and thus can be handled by Strategy 3.)
  - ✧ If not, consider the first Chinese line and the first English line as parallel and MARK them as ALIGNED; SET the second Chinese and the second English lines as the first Chinese and the first English lines respectively; repeat the Governing Strategy.

4.3.4 The aligning algorithm

The whole procedure of text alignment is shown by an algorithm graph as is shown below. In the graph, all squares refer to certain functions or macro-functions defined in the aligning program; the lozenges refer to certain judgment processes which lead to processes in different directions. When the program reaches the end of the texts, the whole process comes to an end.



The Aligning Algorithm

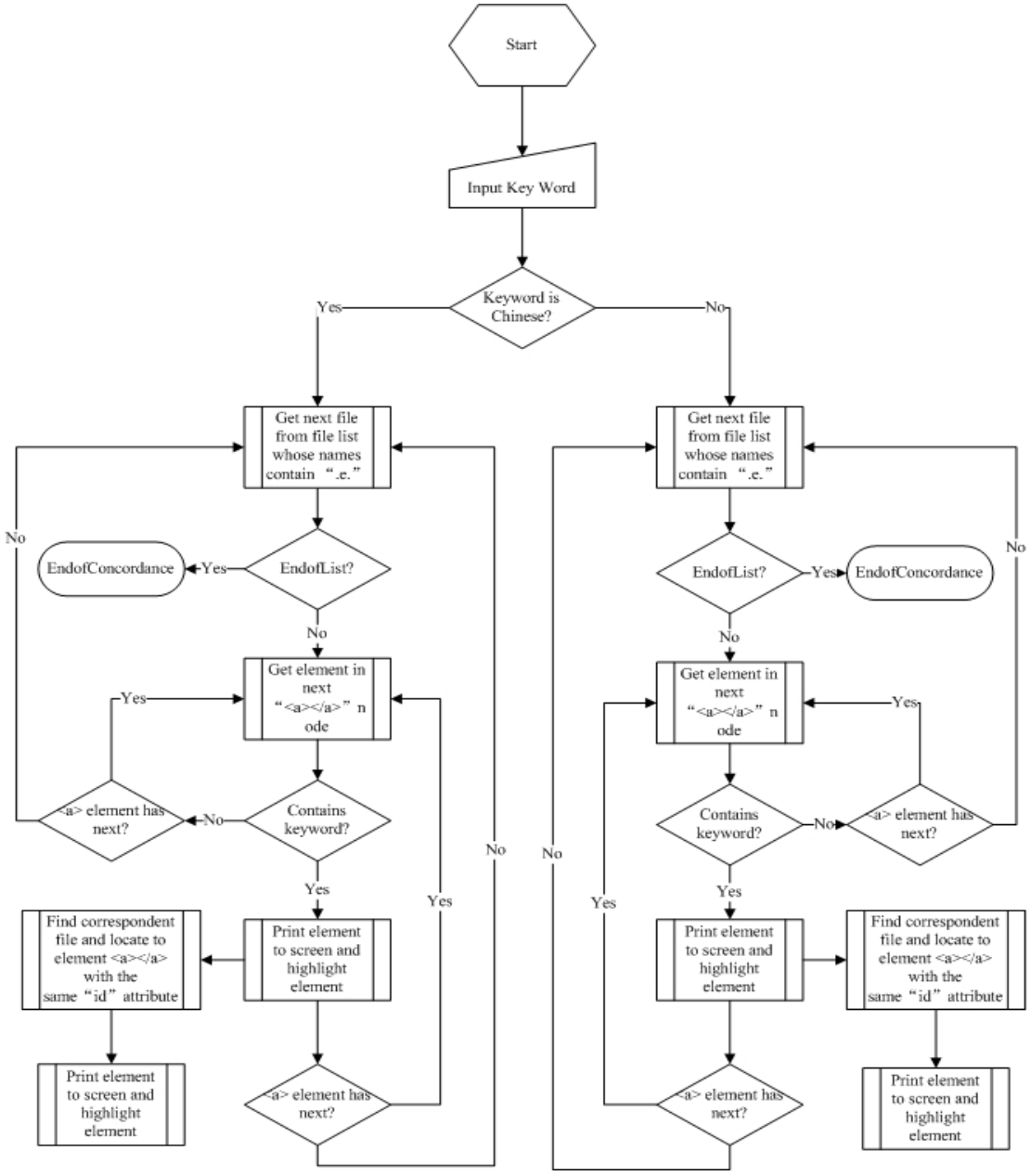
4.4 Text annotation

Corpus annotation means to add informative or explanatory linguistic information to a corpus as Leech (1997:2) states. He also claims that annotation can greatly add value to a corpus and widen the range of studies the corpus can benefit. The author intends to annotate the MMSC corpus with POS tagging and several softwares are employed. However, this has not yet been completed by the time this paper is finished.

4.5 Concordancer

The concordancer is one of the most import tools in corpus based studies and usually is, as Kennedy (1998:251) says, ‘easily available’. There are softwares like Wordsmith, ACWT, etc, which can perform well-formed concordances in many types of corpora, including the MMSC corpus. However, as in the case of the MMSC, which is designed to be accessed through the Internet, a user friendly concordance interface is of great importance to the project. Therefore, the author designs and develops a concordancer for MMSC which can perform KWIC (Key Word in Context Concordance) in it.

The concordancer is written in JavaScript, which works with XML database and runs as a webpage portal. At the completion of the paper, the author develops a draft and simple version of the concordancer (beta version 1.1). The working algorithm of the concordancer is shown by the graph below.



The Concordance Algorithm

4.6 Online subtitle processor

The MMSC corpus is designed to be a dynamic corpus whose size increases as more subtitles are processed and included into it. At the completion of the paper, the MMSC contains a simple version of online subtitle processor (beta version 1.3) in order to automatically align and

include subtitles donated by users.

5. Potential applications

As stated in the aims of building the corpus, the MMSC corpus can serve as a data bank for media translation studies, especially subtitling studies. It is also supposed to be of use to English learning activities and teaching practices. Translator trainees and trainers may also find the corpus helpful.

If properly annotated with POS information, the MMSC can greatly enhance its usage in the above mentioned areas, and may be expected to be of some use to the practice in more academic disciplines such as example based machine translation (EBMT), lexicography, subtitle quality control, automatic subtitling, etc. These applications, of course, may require the MMSC corpus to be upgraded and improved in some aspects to better suit their ends.

It is also worth mentioning that the MMSC corpus keeps all the time-cue information intact and thus can be extended and added with new functions to combine audio-visual materials from movies and television with the subtitles. In this way, the MMSC corpus shall become a multi-media corpus.

6. Conclusion

In this paper, the author proposes the idea of using film and television subtitles as the source for creating a parallel corpus and discusses the building process based on the author's work, which is still quite immature and crude due to the author's lack of some necessary knowledge and skills, especially computer programming skills, and all the limitations of sources and software. This pilot study and the simple corpus product, however, as the author wishes, might be of some use to scholars and researchers interested in the area of parallel corpus building and subtitling studies.

Bibliography

Baker, M. (1995) ‘Corpora in translation studies: an overview and some suggestions for future research’. *Target* 7(2): 223-243.

Baker, M. (1998) *Routledge Encyclopedia of Translation Studies*. London: Routledge: 74-76, 244-248.

Baker, M. (1999) ‘The role of corpora in investigating the linguistic behaviour of professional translators’. *International Journal of Corpus Linguistics*: 4.

Barlow, M. (2000), ‘Parallel texts and language teaching’, in Botley, McEnery & Wilson (eds.) *Multilingual Corpora in Teaching and Researchin,* 106-115. Amsterdam: Rodopi.

Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics*. Cambridge: CUP.

Bowker, L. (2001) ‘Towards a methodology for a corpus-based approach to translation evaluation’. *Meta*, 46(2): 345-364.

Chang, Baobao and Bo, Xiaojing (2003) ‘The markup guidelines for the Chinese-English parallel corpus of Peking University’. *Journal of Chinese Language and Computing* 2:195-214.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. New York: Addison Wesley Longman Ltd.: 251.

Laviosa, S. (1998) ‘The corpus-based approach: a new paradigm in translation studies’. *Meta,* 43(4).

Leech, G. (1997), ‘Introducing corpus annotation’, in Garside, R., Leech, G., and Tony McEnery (eds.) *Corpus Annotation*, 1-16. London: Longman.: 2.

Ma, Zhengqi (2005) *An Introduction to Film and Television Translation*. Beijing: CUCP:1-10,15-50.

Olohan, M. (2004) *Introducing Corpora in Translation Studies*. London: Routledge, 40-55.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford Univ. Press, 13-15, 41-42.

Wang, Kefei (2004) *Bilingual Parallel Corpus: Design and Applications*. Beijing: FLTRP: 1-50, 105-199.

Yu, Shiwen (2003) *An Introduction to Computational Linguistics*. Beijing: Commercial Press: 83-108.