Constructing of a Large-Scale Chinese-English Parallel Corpus

Le Sun, Song Xue, Weimin Qu, Xiaofeng Wang,Yufang Sun

Chinese Information Processing Center
Institute of Software, Chinese Academy of Sciences
Beijing 100080, P. R. China
lesun, bradxue, qwm, wxf, yfsun@sonata.iscas.ac.cn

Abstract

This paper describes the constructing of a large-scale (above 500,000 pair sentences) Chinese-English parallel corpus. The current status of Chinese corpora is overviewed with the emphasis on parallel corpus. The XML coding principles for Chinese–English parallel corpus are discussed. The sentence alignment algorithm used in this project is described with a computer-aided checking processing. Finally, we show the design of the concordance of the parallel corpus and the prospect to further development.

Introduction

With the development of the corpus linguistics, more and more language resources have been established and used in language engineering research and applications. As we all know, there are different kinds of corpora for different kinds applications. For example, the Chinese Part-Of-Speech annotation corpus used to train program for Chinese word segmentation and POS tag, the Chinese tree bank used to Chinese syntax study, and so on.

In this paper the constructing of a large-scale Chinese-English parallel corpus, which is totally above 500,000 pair sentences and the first year task is 100,000 pair sentences, is described. The applications of the large-scale Chinese-English parallel corpus put emphasis on the sentence template extracting for EBMT (Example-Based Machine Translation) and translation model training for SBMT (Statistical-Based Machine Translation). The latent applications may include the bilingual lexicon extraction, special term or phase extraction, bilingual teaching, Chinese-English contrastive study, etc.

Numerous corpus data gathering efforts exit all of the world. The rapid multiplication of such efforts has made it critical to create a set of standards for encoding corpora. CES (Corpus Encoding Standard), which is conformant to the TEI Guideline for Electronic Text Encoding and Interchange of the Text Encoding Initiative (TEI 2002), has been adopted by many corpus-based work. The XML Corpus Encoding Standard (XCES) is a part of the Guideline developed by the Expert Advisory Group on Language Engineering Standards (Ide, N., Bonhomme, P., Romary, L. 2000). The coding of our Chinese-English Parallel Corpus is in broad agreement with the TEI Guideline for electronic texts.

In the following section, we first present a brief review of the current status of Chinese corpora with the emphasis on parallel corpus. Then the XML coding principles for Chinese–English parallel corpus are discussed in detail. Following this is the sentence alignment algorithm used in this project with a computer-aided checking processing. Finally, we show the design of the concordance of the parallel corpus and the prospect to further development.

1 Chinese Corpus Project Overview

The Chinese Corpus constructing work started in 1920's, See Zhiwei Feng (2001). The machine-readable corpora established in 1980's are listed as following:

- Chinese Modern Literature Corpus (1979), 5.27 Million Chinese Characters, WuHan University;
- Modern Chinese Corpus (1983), 20 Million Chinese Characters, Beijing University of Aeronautics and Astronautics;
- Middle School Chinese Book Corpus (1983), 1.06 Million Chinese Characters, Beijing Normal University;
- Modern Chinese Word Frequency Corpus (1983), 1.82 million Chinese characters, Beijing Language & Culture University.

The first national large-scale Chinese corpus project is proposed in 1991 by State Language Commission in China. The Chinese texts used in this corpus are selected carefully under the condition of times, genre, and field. Now the corpus is about 20 million Chinese characters.

From 1992, there are several large-scale Chinese corpus constructed by different institutes. The most noticeable in them is the Chinese POS annotation corpus accomplished by Institute of Computational Linguistics, Peking University, with the cooperation with Fujitsu Company. The content of this corpus is people's daily, one of the most popular newspapers in China. The Chinese texts are segmented and added POS tag with high precision. The total Chinese Characters are about 27 million.

There are several Chinese corpora in Tsinghua University also. The corpus, which is used for Chinese segmentation study, includes 100 million Chinese characters. The Hua Yu corpus (2 million Chinese characters) is a POS tagged field-balance corpus. And the 10 percent of this corpus has been used for constructing Chinese tree bank.

These are also other valuable Chinese corpora established in ShanXi University, Harbin technical University, ShangHai Normal University, City University of Hong Kong, Taiwan Academia Sinica, University of Pennsylvania and so on. Please refer to Zhiwei Feng (2001) for detail.

In October 2001, a national corpus project, that is, national 863 project about Chinese Information Processing Platform, is launched. It's a cooperation project between five institutes in China, including Institute of Software, Chinese Academy of Sciences, Institute of Computational Linguistics, Peking University, Tsinghua University, Nanjing University and Institute of Language, State Language Commission. The content of corpora and intended scale in this project are showed in table 1 in detail. The large-scale Chinese-English parallel corpus described in this paper is one of the scheming corpora in this project.

The multilingual corpus is important for computational linguistics research and contrastive linguistics study. So there are many multilingual corpus have been established or being developed in many institutes in China mainland. The table 2 shows the Chinese-English parallel corpus had been constructed in Mainland China. There are also some bilingual corpora about other language pair, such as Chinese-Japanese, Chinese-German, etc.

Sub-Project Name	*Responsible Institute*	*First-Year Scale*	*Scheming Scale*
Chinese Balance Corpus	State Language Commission	70 MCC	150 MCC
Chinese-English Parallel Corpus	IOS, Chinese Academy of Science	100 TS	500 TS
Chinese POS Annotation Corpus	ICL, Peking University	7 MCC	30 MCC
Chinese Tree Bank	Tsinghua University	15 TS	60 TS
Chinese Concept Dictionary	ICL, Peking University	20 TC	60 TC
Chinese Semantic Knowledge Base	Tsinghua University	8 TW	24TW

Table 1 The 863 Chinese corpus project

MCC: Million Chinese Character TS: Thousand Sentence
TC: Thousand Concept TW: Thousand Word

Institute	Corpus Describing	Scale
ICL, Peking University	Sentence & Phrase Alignment	5 TS
Harbin Institute of Technology	Sentence, Phrase, Word Alignment	Above 5 TS
State Language Commission	Computer Science and Plato	Unknown
Beijing Foreign Studies University	Literary, Science and Civilization in China	Unknown
Northeastern University	Sentence & Phrase Alignment	Unknown
IOS, Chinese Academy of Science	Sentence Alignment	8 TS

Table 2 The Chinese-English parallel corpus in Mainland

It has been noticed by many scholars that we should build a principle for sharing language resource in research work and to avoid the waste in time and effort in repeated construction.

2 Resource Collection

Unlike single linguistic resource, the parallel resource for special language pair is limited no matter what language pair is. Although the Chinese and English both are most popular language in the world, we still encounter much difficult in obtaining parallel corpus resource from Internet for following reasons:

- There are seldom web pages in China provide the same content in English pages and in Chinese pages;
- The English news in web are translated freely other than literally with many content omission;
- Some bilingual texts are restricted and used only to member.

After two years efforts, there are totally about 16,000KB untagged Chinese-English parallel texts in hand. The genres of the resource we collected are showed in table 3.

Chinese Genre	About Percent
News	10%
Literature	30%
Government Report	25%
Sciences & Technology	35%

Table 3 The genre in parallel texts

3 Coding

3.1 General Principles

The coding of the parallel corpus is in broad agreement with the TEI Guideline for electronic texts. The eXtendible Make-up Language (XML) is used for the text coding. Textual features are marked by tags enclosed within angle brackets. For example, a title is marked by start tag <title> and an end tag </title >. Every element has some attributes to identifier of the element.

The document type definition (DTD) for the texts in the corpus may differs in some respects from the TEI model. The general principle for coding are based on following consideration:

- Comply with TEI guide lines on the whole;
- Define the tag with clear meaning used by most people in china;
- Only used the attributes which can be easily and automatically get from source texts, except the alignment link, which is the key attribute in this corpus and several steps are used to keep high precise (See section 4 for detail);
- Try to keep all the interim resource in hand in case information loses, such as, the title tag in HTML files.

The overall structure of a Chinese-English Parallel corpus is shown by this example:

<article id="UH001">
<Header type ="Unix Handbook">
</Header>
<text>
</text>
</article>

There are two main parts in a text: a header and the main text. Every text has an unique identifier that is, article id, in this case UH001 (indicating text 001 of the Unix Handbook)

3.2 The header

Each text is described by a header, which has four parts in accordance with the TEI guidelines: a file description, an encoding description, a profile description, and a revision description. The file description gives bibliographical information on the source text. The elements include title, author, www address (If the text is obtain from Internet), etc. The encoding description in our corpus is very brief, only the project name and the DTD file name are listed.

The country or region use the language is indicated in the profile description. The description under <language> used in our corpus is in terms of labels like: Mainland Chinese (MaC), Hong Kong Chinese (HKC), Taiwan Chinese (TwC), Singapore Chinese (SiC), American English (AmE), British English (BrE), Canadian English (CaE), etc.

Another tag used in the profile description is <textclass>. According to the parallel resource in hand, the texts are grouped into 4 genres (as show in table 3), such as, News , Literature, Science & Technical，Government Report.

A series of changes are listed in the revise description and specified the change, the date of the change, the person responsible for the change, and the nature of the change.

3.3 Text Units and Alignment Unit

The corpus texts are segmented according to the natural units, such as: chapter, paragraph, sentence (S-unit), and word. The English words are simply marked by spacing as in ordinary written text. The Chinese words are not indicated by space in order to avoid the segment error.

An ID is given to every paragraph to indicate the relative position in whole chapter. The sentence is called S-unit, the same as Johansson, Ebeling and Oksefjell (1999) to underline that they are not necessarily sentences in a grammatical sense.

The sentence alignment type between Chinese S-unit and English S-unit maybe 1:1, 2:1, 3:1, 1:2, 1:3,2:2, 3:2, 2:3. Links between parallel texts are showed by attributes of S-Alignment. One of the Chinese alignment unit (it may beyond one S-unit) are linked with the correspondence English alignment unit.

3.4 Sample Text

A sample text of our Chinese-English parallel corpus is showed in figure 1.

```
<?xml version="1.0" encoding="gb2312" ?>
- <Article id="GR23">
  - <Header type="人民日报白皮书">
    - <fileDesc>
        <title>关于中美贸易平衡问题</title>
        <date>1997</date>
      </fileDesc>
      <encodingDesc>CEPC.dtd (See Chinese-English Parallel Corpus manual)</encodingDesc>
    - <profileDesc>
        <language>MaC-to-Eng</language>
        <text_class>Government Report</text_class>
      </profileDesc>
    - <revisionDesc>
        <date>2002-03-25</date>
        <person>安阳</person>
      </revisionDesc>
    </Header>
  - <Text>
    + <S_Alignment num="918">
    - <S_Alignment num="919">
        <CH_sentence p_num="2" s_num="2">中国方面一直非常重视并采取积极措施扩大自美国进口。
          </CH_sentence>
        <EN_sentence p_num="2" s_num="2">The Chinese side has always paid great attention to
          the need and taken active measures to increase imports from the United
          States.</EN_sentence>
      </S_Alignment>
    </Text>
</Article>
```

Figure1 Sample Text

4 Sentence Alignment

4.1 Algorithm Overview

The key attribute in this corpus is alignment link, which connect the one or more Chinese sentence with one or more correspond English sentence. In order to keep high precise in sentence alignment, several steps are used with the human and computer cooperation.

The first step to extract structural information for parallel corpus is paragraph alignment and sentence alignment, that is noting which paragraph and sentence in one language correspond to which paragraph and sentence in another language.

This problem has been studied by many researchers and a number of quite encouraging results have been reported. However, almost all bilingual corpora used in research are clear (nearly without sentence omission or insertion) and literal translation bilingual texts. The performance tends to deteriorate significantly when these approaches are applied to noisy complex corpora (with sentence omission or insertion, less literal translation).

There are basically three kinds of approaches on sentence alignment: the length-based approach (Gale & Church 1991 and Brown et al. 1991), the lexical approach (key & Roscheisen 1993), and the combination of them (Chen 1993, Wu 1994 and Langlais 1998, etc.).

The first published algorithms for aligning sentences in parallel texts are length-based approach proposed by Gale & Church (1991) and Brow et al (1991). Based on the observation that short sentences tend to be translated as short sentences and long sentences as long sentences, they calculate the most likely sentence correspondences as a function of the relative length of the candidates. The basic approach of Brow et al. is similar to Gale and Church, but works by comparing sentence length in words rather than characters. While the idea is simple, the models can still be quite effective when used to clear and literal translated corpora. Once the algorithm had accidentally mis-aligned a pair sentence, it tends to be unable to correct itself and get back on track before the end of the paragraph. Use alone, length-based alignment algorithms are therefore neither very robust nor reliable.

Kay & Roscheisen (1993) use a partial alignment of lexical items induce a maximum likelihood at sentence level. The method is reliable but time consuming.

Chen (1993) combines the length-based approach and lexicon-based approach together. A translation model is used to estimate the cost of a certain alignment, and the best alignment is found by using dynamic programming as the length-based method. The method is robust, fast enough to be practical and more accurate than previous methods.

The first sentence alignment model used to align English-Chinese bilingual texts is proposed by Wu (1994). For lack of cognates in English-Chinese, he used lexical cues to add the robust of his model.

All of these works are test on nearly clear and literal translation bilingual corpora.

There are seldom papers related to paragraph alignment. It's believed by most of the researchers that the paragraph alignment is an easier task than sentence alignment. Gale & Church (1991) suggest that the same length-based algorithm can be used to align paragraph also.

4.2 The Alignment Steps

Sentence alignment algorithm of our system can be outlined as follows:

Step 1: Align sentence by the improved length-based algorithm.(Desicibed in Sun etc. 1999)

Step 2: A lexicon checking process is added to judge all the alignment results in step 1. A score is given to every alignment pair (A Chinese word segmentation system is used in this process to find Chinese word).

Step 3: The alignments whose score above a threshold C_1 are judged as correct alignment. Remove these correct alignments from bilingual texts temporally.

Step 4: The rest parts are aligned again by length based approach.

Step 5: Repeat step 2, the score of every lignment is showed as a reference to human checking.

4.3 Computer-Aided Checking

It's obviously difficult to increase greatly the accuracy and robust of sentence alignment only by length based approach. So a lexicon checking process is added to our system. The alignment results obtained by length based approach are checked by an English-Chinese lexicon. A score S_A is given to every alignment sentence pair. The score S_A is calculated by following idea, that is, the twice number of correctly matched English words and Chinese words to the sum of number of English and Chinese words. In figure 2, the interface for human checking is showed in order to processes the noise Chinese-English parallel resource.

4.4 Experiment Results

We tested our alignment algorithm with part of a computer handbook (Sco Unix handbook). There are about 4681 English sentences and 4430 Chinese sentences in this computer handbook after filter noisy figures and tables. The detail experiment result of automatic sentence alignment is show in table 4. The total precision is about 95%.

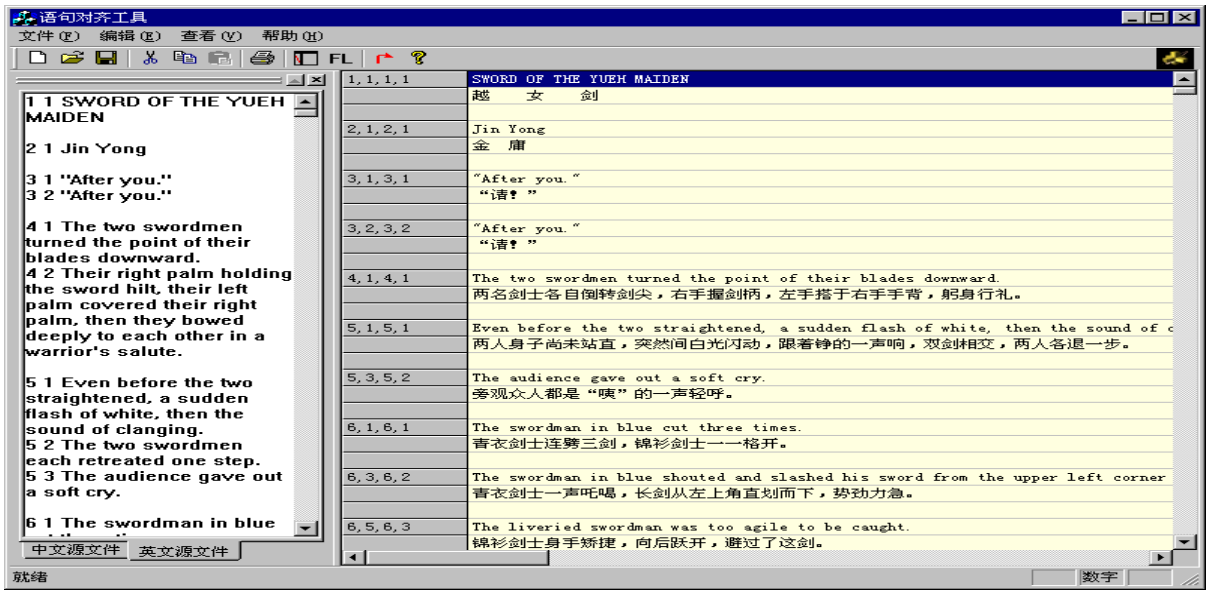

Figure 2 Interface for Human Checking

Class of Alignment	No. of Aligned Sentence Pair	No. of Correct Sentence Pair	No. of Error Sentence Pair	Precision
1:1	2992	2957	35	98.83%
1:2	238	211	27	88.66%
2:1	414	352	62	85.02%
2:2	113	97	16	85.84%
1:3	35	24	11	68.57%
3:1	75	49	26	65.33%
2:3	13	6	7	46.15%
3:2	22	16	6	72.72%
3:3	6	3	3	50.00%
0:1	3	2	1	66.67%
1:0	7	4	3	75.00%
Total	3918	3721	197	94.97%

Table 4 The detail experiment result of automatic sentence alignment

5 Bilingual Concordance Design

We also designed a bilingual concordance tool used for discovering facts during the translation between Chinese and English. Besides a listing of the keywords with the contexts in which they appear, the correspondence translation sentence also be presented in this tool. The options may include bilingual concordances, sorting in a variety of orders, and producing basic text statistics. The intended interface is showed in figure 3.

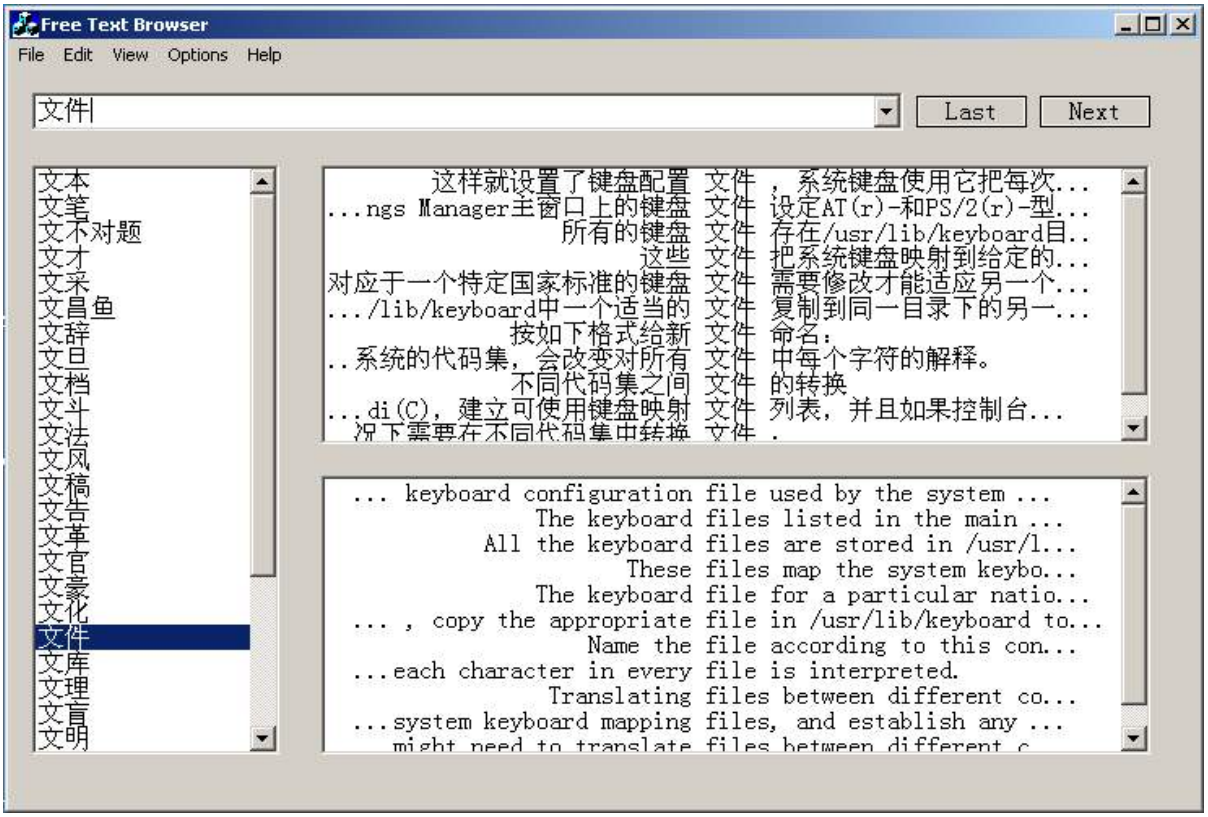

Figure 3 The Interface for bilingual Concordance

6 Conclusion & Further Prospects

In this paper, we introduce the developing project, that is, the constructing of a large-scale (above 500,000 pair sentences) Chinese-English parallel corpus. The current status of Chinese corpora is overviewed with the emphasis on parallel corpus. The XML coding principles for Chinese–English parallel corpus are discussed. The sentence alignment algorithm used in this project is described with a computer-aided checking processing in order to processes the noise Chinese-English parallel resource.. We show the design of the bilingual concordance for the parallel corpus, also.

As a beginning project, there is still much room for further development. The parallel resource is relative rare, so the new ways, such as, data exchange with other researcher institute and translation company, should be launched to obtain more parallel resource which can be used to research society. The coding principle should be adjusted in real work. A coding rule in more detail should form in near future. We also intend to add the option for recommendation the correspondence translation word for input keywords in concordance tool.

Acknowledgements

This work is supported by China 863 project (Grant No. 2001AA114040) and the National Science Fund of China under contact 69983009. Our thanks go to all the project members from

five institutes for discussion and the anonymous reviewers for kind suggestions. .

References

Catherine N. Bal (1997), Tutorial: Concordances and Corpora, http://www.georgetown.edu/cball /corpora/tutorial.html

D. Wu, (1994) *Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria*, In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94), pp.80-87

I. D. Melamed. (1996) *Automatic Detection of Omissions in Transaltions,* In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark

ISLE, International Standards for Language Engineering http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

J.S. Chang and M. H. Chen (1997) *An alignment method for noisy parallel corpora based on image processing techniques,* In Proceedings of the 35th Meeting of the Association for Computational Linguistics, Madrid, pp. 297-304

Kay M., and Roscheisen M. (1993). *Text-Translation Alignment*, Computational Linguistics, 19/1,pp.121-142

Le Sun , Lin Du, Yufang Sun, Jin Youbin (1999) *Sentence Alignment of English-Chinese Complex Bilingual Corpora.* Proceeding of the workshop MAL'99, 135-139

N. Ide, L. Romary (2001). A Common Framework for Syntactic Annotation *Proceedings of ACL'2001,* Toulouse, 298-305

N. Ide, L. Romary, (2000) XML Support for Annotated Language Resources. *Proceedings of the Workshop on Web-based Language Documentation and Description,* Philadelphia, 148-153.

N. Ide, P. Bonhomme,, L. Romary (2000). XCES: An XML-based Standard for Linguistic Corpora.. *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece, 825-30.

P. F. Brown, J. C. Lai, and R. L. Mercer (1991) *Aligning Sentences in Parallel Corpora*, In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pp.169-176.

P. Fung, and K. W. Church (1994) *K-vec: A New Approach for Aligning Parallel Texts*, In Proceedings of the 15th International Conference on Computational Linguistics (COLING'94), Tokyo, Japan, pp. 1096-1102,

Ph. Langlais, M. Simard, J. Veronis, S.Armstong, P. Bonhomme, F. Debili, P. Isabelle, E. Souissi, and P. Theron. (1998) *Arcade: A cooperative research project on parallel text alignment evaluation*. In First International Conference on Language Resources and Evaluation, Granada, Spain.

S. F. Chen, (1993) *Aligning Sentences in Bilingual Corpora Using Lexical Information*. In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics, pp. 9-16

Shiwen Yu, Xuefeng Zhu, Hui Wang, Yunyun Zhang (1998), *The Grammatical Knowledge-base of Contemporary Chinese: A complete Specification.* Tsinghua University Publishers

Stig Johansson, Jarle Ebeling, Signe Oksefjell (1999), English-Norwegian Parallel Corpus:Manual, http://www.hf.uio.no/iba/prosjekt/

TEI (2002) The XML Version of the TEI Guidelines http://www.hcu.ox.ac.uk/TEI/Guidelines/

W. A. Gale, and K. W. Church (1991) *A Program for Aligning Sentences in Bilingual Corpora*, In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pp. 177-184

Zhiwei Feng (2001), *The History and Current status of Chinese Corpus Research*, International Conference on Chinese Computing ICCC2001, pp. 1-15 (In Chinese)