Evolution and present situation of corpus research in China

Zhiwei Feng
Institute of Applied Linguistics, China

In this paper, the author introduces in detail the development and present situation of corpus linguistics in China: earlier corpora, large-scale & authentic text corpora, national corpora, speech corpora, bilingual corpora and corpora of minority languages in China. The various processing techniques for corpora are also introduced: automatic word segmentation of Chinese text, automatic PoS tagging, automatic tagging of phrase structure and automatic alignment of bilingual corpora. This paper is a bird's-eye view of corpus linguistics of China. Finally, the author discusses several problems in present corpus research: standardization of corpus specifications, commonly sharing of language resources, knowledge properties, etc.

Keywords: corpus; large-scale & authentic text; speech corpora; bilingual corpora; corpora of minority languages in China; automatic word segmentation; automatic PoS tagging; automatic tagging of phrase structure; automatic alignment of bilingual corpora

The research of linguistics must be based on real language data and work on a large quantity of detailed material. Only then more reliable theoretical conclusions can be drawn. The traditional collecting, sorting and processing of language material had to be implemented completely manually. This is a kind of dull and time-consuming mission. After the invention of the computer, people can hand over these works to the computer, diminishing human labor. Subsequently, some innovative methods were gradually created for this kind of work and some preliminary theories were initiated. Thus a new discipline — corpus linguistics took shape. Since the corpus is physically stored on the computer, many scholars tend to take the view that corpus linguistics is a branch of natural language processing (NLP).

International Journal of Corpus Linguistics 11:2 (2006), **173–207**.
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

At the present level of research, corpus linguistics is little more than a novel research approach primarily used in certain research linguistic domains working with real language data. Precisely speaking, corpus linguistics still lacks a commonly accepted and fully developed theory. It cannot be mentioned in the same breath as the other more mature disciplines like computational linguistics, sociolinguistics and psycholinguistics.

At present, corpus linguistics is primarily concerned with compiling, storing, grammatical tagging, and the statistical analysis of syntactic and semantic features of machine-readable natural language texts. It also involves the use of corpora within natural language processing for a quantitative analysis of language, for lexicography, for the analysis of style in written language, for the understanding of natural language and for machine translation.

Corpora are the research object of corpus linguistics. The compilation of corpora and the representation of the data on electronic memory must be facilitated. The stored data should be the authentic materials used in reality. However, only after being analyzed and processed, these authentic materials can become useful language resources.

For many years in the research of machine translation and natural language understanding, the primary objective has been the analysis of syntactic and semantic features. Consequently, during a very long period, many systems have been entirely based on rules of syntax and semantics. Nevertheless, given the present level of computer theory and technology, it is an extremely tough mission to fully display the diverse facts of language and the extensive background knowledge needed in understanding language if we restrict ourselves solely to the rule-based approach.

Therefore rule-based machine translation and natural language understanding systems can only be successfully used in some restricted sub-languages. To escape this dilemma, the researchers of natural language processing started to survey large-scale non-restrictive natural language for the purpose of adopting a statistics-based model in order to deal with non-restrictive authentic natural language. It is self-evident that based on the large scale language materials stored in the computer, corpus linguistics will have the possibility to test the traditional theoretic linguistic conclusions drawn by the means of manually gathering very restricted language data. Accordingly, corpus linguistics will give us a more intensive and comprehensive insight of complex phenomena of the authentic natural language.

This article elaborates the evolution and achievement of corpus linguistics in China. The readers will obtain a bird's eye view.

1. A survey of the evolution of corpus linguistics in China

1.1 Earlier Chinese corpora before 1991

1.1.1 *The initial study of corpora in China*

In China, in the beginning of 1920s, some scholars embarked on setting up text corpora, studying the frequency of Chinese characters. They used a statistical approach for the purpose of preparing frequency lists of basic Chinese characters. It is apparent that this kind of corpus was not machine-readable and that its scale was also very small. However, it was the embryonic form of a modern corpus and it can be seen as a pioneer study of corpora in China. It is memorable in the history of the development of corpus linguistics in China. *The Applied Glossary of Modern Chinese* 《语体文应用字汇》was published by the Commercial Press in 1928. This glossary was compiled on the basis of corpus data processed for the sake of pedagogy by the educationist Chen Heqin in 1925. Chen explicated, in the preface of the glossary, that the Applied Glossary had many versions including Pastor Kronz's research on Chinese characters and Chen's glossary, *The Glossary of 4000 Frequently Used Words* 《常用四千字表》.

Chen Heqin (1928) made two statistical calculations. At first, he adopted six corpora containing 554,478 Chinese characters (tokens) and identified 4261 different Chinese characters (types). Then he used a corpus containing 34,818 Chinese characters, and identified 458 Chinese characters (types) which were different from the 4261 Chinese characters found at his first go. However, the results of the second time were lost in the Civil War. What was printed in *The Applied Glossary of Modern Chinese* was only the result of his first analysis.

The six corpora employed by Chen Heqin are respectively classified into six categories:

1. Children's books: 127,293 Chinese characters
2. Newspapers (The majority is popular newspapers): 153,344 Chinese characters
3. Women's magazines: 90,142 Chinese characters
4. After-class works of primary school students: 51,807 Chinese characters
5. Classical and modern fiction: 71,267 Chinese characters
6. Miscellaneous: 60,625 Chinese characters

The glossary came with an attachment: *The Contrastive Table of Characters Frequencies*. This was a table listing characters according to their absolute frequency in the corpora.

The remarkable Chinese educationist Tao Xingzhi wrote the preface for the *The Applied Glossary of Modern Chinese*. In the preface, Tao says that the modern educationists would, depending on the goal, inquire the utility of an assignment, a passage, an arithmetic problem, and a sport in the school. They would contend that what needs to be learned should be what can be used. Later on, they also examined the Chinese characters the students were learning. Was every character the students were learning useful? In order to answer this question, quite a few scholars, many of them working in China, began to probe the use of Chinese characters. Cheng Heqin's research was the most systematic. He and his nine assistants spent two to three years examining modern Chinese texts containing tens of thousands Chinese character tokens and compiling *The Applied Glossary of Modern Chinese*. Before printing this glossary, principles of selecting Chinese characters for *Thousand-Characters Lessons for Civilians* 《平民千字课》 had been drawn up. These principles could certainly be used as guidelines for the selection of Chinese characters in teaching books for the primary schools. Though this glossary was not perfect, I believe it has certainly made a great contribution to the education of adults and civilians in China.

1.1.2 *Earlier machine-readable corpora*

In the year of 1979, China has begun the construction of the machine-readable corpora. The more important machine-readable corpora from that period are:

- Chinese Modern Literature Work Corpus (in 1979), 5.27 million words, Wu Han University
- Modern Chinese Corpus (in 1983), 20 million words, Beijing Aviation University
- High School Chinese Language Teaching Material Corpus (in 1983), 1.068 million words, Beijing Normal University
- Modern Chinese Corpus with Statistics of the Frequency of Words (in 1983): 1.82 million words, Beijing Language College

We take the corpus of Beijing Language College as an example to shed light on the earlier generation of machine-readable corpora.

In 1979, Beijing Language College, currently renamed as Beijing Language University, undertook the key research project *Statistical Survey of Modern Chinese Words* to embark on a large scale statistical survey of the frequency of Chinese words. As there are no spaces between words in Chinese texts, word segmentation is necessary.

In this research project, the words in the Chinese text were (manually) segmented in order to count their frequency and to analyze the data. The size of

this corpus is 1.82 million characters. After the word segmentation, the words appeared with spaces in the text. The total amount of words is 1,315,752 occurrences (tokens), representing 31,159 different words (types). As the result of this project, *The Dictionary of Modern Chinese Word Frequency* 《现代汉语频率词典》was published.

The texts they selected can be classified according to the following four genres:

1. Newspapers and political critics: 440,000 characters, accounting for 24.4% of the total amount of corpus.
2. Science, technology and popular science texts: 290,000 characters, accounting for 19.8%.
3. Spoken corpus: 200,000 characters, accounting for 11.1%.
4. Literature works: 890,000 characters, accounting for 48.7%.

The entire corpus contains 1.82 million characters. At that time, the corpus with such a large volume was seen as a relatively large corpus.

The author of *The Dictionary of Modern Chinese Word Frequency* argues that if the frequency of the commonly used words is not less than 1 per million, i.e. that it occurs on the average at least once in a million words, then the results of the statistical analysis could be considered reliable. Given the fact that *The Dictionary of Modern Chinese Word Frequency* was compiled based on the corpus with 1.31 million word (1.82 million Chinese characters), its extraction of words should therefore be reasonable, economical and appropriate.

However, in 1971, when frequency statistics of words were carried out in the developed countries, the corpora used for these studies were significantly larger than that of *The Dictionary of Modern Chinese Word Frequency*. With the increasing capability of the corpus linguistics, we can today find corpora with billions of words. Compared with these corpora, the corpus used for *The Dictionary of Modern Chinese Word Frequency* is obviously small. In spite of its small scale, *The Dictionary of the Frequency of Modern Chinese Word* was a milestone in the statistical analysis of word frequency in China.

The statistics of word frequency resulted in the following table of words:

1. The table of word frequency, arranged according to the alphabetic sequence of Chinese pinyin, listed 16,593 commonly used words and displayed
   - There are a great number of words with the initials Z, S, J, Y: 1457 word with initial Z accounting for 8.78%, 1327 words with the initial S accounting for 7.99%, 1243 words with the initial J accounting for 7.49%, 1205 words with the initial Y accounting for 7.26%.

- There are few words with the initials E and O: only 64 words with the initial E, accounting for 0.38%, and only 13 words with the initial O accounting for 0.07%.

2. The table of words arranged according to frequency order: The first 100 words take 40% of the total amount of the corpora. The first 500 words take 70% of the total amount of the corpora. The first 2562 words take 85% of the total amount of the corpora. The table of words contains 31,159 words, accounting for 100% of the total amount of corpora. From the first 100 words to the first 500 words, 400 different words have been added and correspondingly the coverage has increased by 30%. However, from the first 2562 words to the total of 31,159 words, 30,597 different words have been added, but the coverage has increased by only 15%. Thus, a conclusion can be drawn that the highly frequent words can exert a great impact on the increase of coverage, and the words of low frequency have only a slight impact on the increase of coverage.
3. The table of words arranged according to the usage rate: The usage rate is a new notion proposed by Juilland and Chang-Rodriguez for calculating the frequency of Spanish words in 1954. They also supplied the mathematical formula for calculating the usage rate of word. The usage rate of word, calculated according to the mathematical formula, can comprehensively describe the occurrence frequency and the distribution rate of words over texts in the corpus.

   Based on the mathematical formula of the usage rate of word, they calculated the usage rate for every word in corpus and worked out the table of words according to the usage rate. This table is further divided into two tables: the table of the first 8000 words with high usage rate and the table of words with low usage rate.

   In the table of the first 8000 words with higher usage rate, there are altogether 4186 words with usage rate above 20 (usage rate > 20), words accounting for 90.1% coverage of the entire corpus (314,404 word tokens). This suggests that in the original corpus used to compile *The Dictionary of Modern Chinese Word Frequency*, nine out of ten texts are composed of these 4186 words, which thus become the candidates of the "commonly used words".

   In the table of words with low usage rate, there are 22,446 words with usage rate 5 (the usage rate = 5) and usage rate below 5 (usage rate < 5), which are generally words of low frequency. If there are some words with this usage rate that match the required frequency, then these words are

evenly distributed and will be selected as the candidates of the "generally used words".

4. The table of words of high frequency classified according to the genres is sub-divided into four tables:
   a. The table of the first 4000 words in the genre of political editorials in newspapers and magazines: This table is based on calculated 34 texts containing 290,000 word occurrences (representing 440,000 characters) and 12,107 different lexical items. The accumulative frequency of the first 4000 words is 94.77%, among which some political terms like "唯心" (spiritualism), "党派" (clan) all have a higher frequency indicating the characteristics of the genre of political editorials.
   b. The table of the first 4000 words in the genre of popular science: This table altogether based on 21 texts containing 200,000 word occurrences (290,000 characters) and 12,364 different lexical items. The accumulative frequency of the first 4000 words covers 92.27% of the sub-corpus, among which some scientific and technical words like "纤维" (fiber), "合成" (synthetic) etc have a higher frequency, indicating the characteristics of the genre of popular science.
   c. The table of the 4000 most frequent words in the spoken language genre: This table is based on 18 texts, containing 160,000 word occurrences (200,000 characters) and 8263 different lexical items. The accumulative frequency of the first 4000 words covers 96.65% of the corpus. Statistics shows that the total amount of words of spoken genre is one third less than that of the first two sorts. However, there is a considerable amount of high frequency words. The frequency of the first 1000 most frequent words is 6% higher than that of Table (a) and 12% higher than that of Table (b). This illustrates that although there is not such a large amount of words in the spoken genre, they cover a significantly large amount of the sub-corpus.
   d. The table of the first 4000 high frequent words in works of fiction: This table is based on 106 texts, containing 660,000 words (890,000 characters), and 23,622 different lexical items. The accumulative frequency of the 4000 most frequent words covers 90.63% of the sub-corpus. This indicates that despite the fact that the literature works contain a large quantity of words, the occurrence rate of high frequency words is relatively low for the sake of diversity. This suggests that the works of fiction possess a multitude of diverse and colorful words.

The earlier corpora have the following characteristics:

(1) Most of them were set up using the approach of manual keyboard input. These were time-consuming and exhausting tasks. Furthermore, their composition was less principled, they were smaller and had a low rate of repetition. Building this kind of corpora was a tough and demanding task. The well-known expert Liu Yuan, a professor in computer department of the Beijing Aviation University, collapsed from too much work and died in the end. We pay tribute to the dedicated spirit of the early pioneers of corpus research in China.

(2) Beijing Aviation University did the research of automatic segmentation of words and detected two different sorts of Ambiguous Segmentation Strings (ASSs): Overlapping ambiguous segmentation strings and Combinative ambiguous segmentation strings.
   - Overlapping Ambiguous Segmentation Strings: for instance, "地面积" (/dimianji/) can be segmented as "地面" (/dimian/, surface) or "面积" (/mianji/, area), "面" (/mian/) turning out to be the overlapping section and generating ambiguous meanings.
   - Combinative Ambiguous Segmentation Strings: for instance, "马上" (/mashang/) in itself is a word, meaning "at once". However, it can also be segmented as two words: "马" (/ma/, horse) plus "上" (/shang/, on), and it means: "on the horse back". Thus "马上" (at once) has a different meaning from 马" plus "上".

   Liang Nanyuan (1987) made a statistical analysis of a 48,092 word sample of natural science and social science texts. He found 518 overlapping ambiguous segmentation strings and 42 combinative ambiguous segmentation strings. This shows that in Chinese text segmentation, the frequency of ambiguous segmentation is 1.2 time per 100 words: the ratio of the overlapping ambiguous segmentation and combinative ambiguous segmentation is 12:1.

(3) In October in 1990, under the joint endeavors of researchers from the domains of computer and linguistics, a preliminary segmentation standard was implemented: the national standard, GB-13715 *The Segmentation Criterion of Modern Chinese Words Used for Information Processing*. This national standard advances the principles and specification for the segmentation of Chinese words, and it is the guideline for the automatic segmentation of Chinese written language.

1.2 The construction of National Chinese Corpus

In 1991, the State Language Commission of China, which has now been merged into the China National Ministry of Education (MOE), began to construct a large scale Chinese corpus on the national level which aimed to be the driving force of the research of morphology, syntax, semantics and pragmatics of Chinese linguistics and to provide a reliable language resource for the research of Chinese information processing. The size was planned to be 70 million Chinese characters and was to become the largest Chinese corpus in the world. This corpus was a balanced corpus with a sophisticated composition and applying the following constraints:

a. *The constraint of time*: The corpus should have a diachronic dimension with an emphasis on the synchronic dimension. The corpus, divided into 5 periods, should be composed of language materials since 1919 to the present time, mainly from the Chinese materials after 1977.
b. *The constraint on the cultural background*: It should be primarily composed of texts understandable to common people with secondary education.
c. *The constraint on domains*: The corpus is composed of three portions: a humanity & social science portion, a natural science portion and an integrative portion. The humanities & social science portion is further classified into eight large categories and 29 smaller categories. The natural science portion is subdivided into six large categories. The integrative portion is also subdivided into two large categories. The corpus is mainly composed of general language resources but the priority is given to texts in the humanities and the social sciences.

So far, only the 20 million character corpus has been accomplished in terms of inputting and proofreading. It has been further processed in terms of segmentation and PoS-tagging. The construction of this corpus cost 2 million Yuan (Renminbi) mainly due to the manual input and the high cost of human labor.

In order to process this National Chinese Corpus, the Chinese National Foundation of Social Science established the key project *The Research of the Modern Chinese Vocabulary for Information Processing* anticipating possible applications of the results. This project is divided into 10 categories:

1. A table of modern Chinese words for segmentation purposes in the information processing
2. The development of software for disambiguation of ambiguous segmentation and for recognizing proper names (personal names and place names)

3. Research of Chinese PoS tag-sets
4. Research of the structure of word formation
5. Research of modern Chinese words having two or more parts of speech
6. Descriptive research of the grammatical attributes of modern Chinese
7. Research for a modern Chinese predicate-verb machine dictionary and slot relation of verb
8. Research for compiling a modern Chinese knowledge dictionary and the description of semantic net inside the lexicons
9. Research of the semantic features of frequently used verbs and collocations
10. Manual tagging of Chinese text phrasal structure

Currently, this project is being completed. The Institute of Applied Linguistics (attached to China National Ministry of Education) has established a research team for the project, *The In-depth Processing of the Chinese Corpus*, leading towards the in-depth processing of the 20 million-word core corpus within the National Chinese Corpus.

1.3 Large scale authentic text corpora

Since 1992, many large-scaled authentic text corpora have been established among Chinese information processing institutions (including research institutes and universities) in China. These corpora have become the fundamental resource for the research of Chinese information processing. Without the corpora, this research of Chinese information processing would not have been conducted. The following institutions have compiled large-scale authentic text corpora:

- *People's Daily* Compact Disc Corpus Section
- Institute of Computational Linguistics at Peking University
- Beijing Language University
- Tsinghua University
- Shanxi University
- Shanghai Normal University
- Beijing Posts and Telecommunication University
- The Hong Kong City University
- North-east University
- Harbin Polytechnic University
- Institute of Software at Chinese Academy of Science
- Institute of Automation at Chinese Academy of Science

- Beijing Centre for Japanese Studies at Beijing Foreign Language University
- The Preparatory Office of the Institute of Linguistics at Academia Sinica (Taiwan)

These corpora are described as follows:

1.3.1 *People's Daily* Compact Disc Corpus

48-year text and picture material of this newspaper are all collected and issued publicly.

1.3.2 *Institute of Computational Linguistics at Peking University*

This institute has compiled a tagged modern Chinese corpus. The institute is in cooperation with Fujitsu Company (Japan), and has been processing 27 million Chinese characters of *People's Daily* Corpus, including word segmentation and part-of-speech tagging, proper noun (proper noun phrase) tagging, and phonetic notation tagging of multi-pronunciation word.

Paradigm 1:

古城/n 虽/c 遭/v 破坏/v,/w 但/c 它/r 留下/v 了[le5]/u 契丹族/nz 和[he2]各[ge4]/r 民族/n,/w 特别/d 是/v 汉族/nz 劳动/vn 人民/n 共同/d 开拓/v 祖国/n 北疆/s,/w 创造/v 我国/r 历史/n 文明/n 的[de5]/u 足迹/n 。/w

[Despite the fact that the ancient city experienced the destruction, it maintained the tracks that Khitan nationality (契丹族) and other nationalities, especially the Han nationality working people, collaboratively explored the northern frontier areas of our motherland and created the history and civilization of our country.]

Paradigm 2:

19970310-01-002-0020/m [全国/n 人大/j]nt 代表/n 、/w [陕西/ns 西安/ns 美术/n 学院/n]nt 名誉/n 院长/n 刘/nr 文西/nr 利用/v 会议/n 休息/vn 时间/n 创作/v 了/u 邓/nr 小平/nr 画像/n 《/w 与/p 人民/n 同/d 在/v 》/w 。/w 画像/n 表现/v 了/u 邓/nr 小平/nr 同志/n 祝愿/v 祖国/n 繁荣/a 、/w 人民/n 幸福/a 的/u 伟大/a 胸怀/n 。/w （/w 新华社/nt 记者/n 齐/nr 铁砚/nr 摄/v ）/w

19970310-01-002-0020 [Mr. Liu Wenxi, the representative of National People's Congress, the honorary president of Xi'an Art College at Shaanxi Province, took the advantage of the interval of the conference to create a painting of Deng Xiaoping: ***Amidst the People****. The painting displays Comrade Deng Xiaoping cherishing the great hope that our motherland would prosper and people would be wealthy. (photographed by Qi Tieyan, Xinhua News Agency)]*

Paradigm 3:
19970310-01-003-0020/m 世纪/n 之/u 交/Ng,/w 中华/nz 民族/n 正/d 迎来/v 前所未有/i 的/u 发展/vn 机遇/n 。/w 十几/m 年/q 来/f,/w 改革/v 开放/v 的/u 不断/d 深入/v,/w 党/n 的/u 民族/n 政策/n 的/u 贯彻/vn 落实/vn,/w 全国/n 人民/n 的/u 大力/d 支援/v,/w 使/v 我国/r 民族/n 地区/n 经济/n 和/c 社会/n 发展/vn 步伐/n 大大/d 加快/v 。/w 加倍/d 珍视/v 和/c 继续/vd 发展/v 这种/r 好/a 的/u 局面/n,/w 巩固/v 发展/v 各/r 民族/n 大/a 团结/an,/w 成为/v 全国/n 各族/r 人民/n 的/u 共同/b 愿望/n,/w 也是/v 在/p 京/j 参加/v “/w 两会/j ”/w 的/u 代表/n 和/c 委员/n 的/u 一致/a 心愿/n 。/w
19970310-01-003-0020 [At the turning point of the century, Our Chinese Nation is encountering the unprecedented opportunity for development. For over ten years, renovation and opening the door are increasingly promoted. The paces of our country's minority areas' development of economic and society have been greatly accelerated due to the implementation of the Chinese Communist Party's nationality policy and the great assistance from the people all over the country. Both the whole country's people and the representatives and commissioners who are participating in Two Congresses in Beijing cherish the common hope of doubly prizing and bettering the excellent existing situation and consolidating and developing the unification of all nationalities.]

After being tested by Fujitsu Company, the corpus proved to have a high precision rate of PoS-tagging.

The team also designed *The Handbook for Modern Chinese Corpus Processing — Word Segmentation and Part of Speech Tagging* (Yu Shiwen & Zhu Xuefeng 2000). For the goal of automatic segmentation, the handbook primarily prescribes the principle for word segmentation of modern Chinese text, in other words, what kind of Chinese combinations can be prescribed as a unit of segmentation. The criteria are based on the principle of combining segmentation and tagging. In Chinese, "a double syllable verb plus a single syllable noun" often constitute a new noun. Yet this kind of new noun has not been entered into dictionaries, even though it should be treated as a unit of segmentation. Therefore, in these guidelines, some PoS-based word-formation regulations were given. The rules for prescribing/defining one unit of segmentation were also provided. Furthermore, the newly coined words were given part-of-speech tagging. In the tagging rules, the part-of-speech tags for general words and for proper nouns were also prescribed.

In addition, in collaboration with the Department of Computer Science at the National University of Singapore, the team also compiled a small scale Chinese treebank: the data used are Singapore high school Chinese teaching materials (1995); most of the sentences are parsed as tree graphs.

Paradigm:
[zj [dj 富士山/n [vp 是/v [np 日本/n 的/u [np [mp 一/m 座/q] 活火山/n]]]] 。/w]
[zj [fj [fj [dj 山峰/n [vp 终年/d 积雪/v]],/w [dj 云雾/n 围绕/v]],/w [vp 只有/d [vp [pp 在/p [np [dj 空气/n 干燥/a] 的/u [np [np 秋/n 冬/n] [np 两/m 季/Ng]]]],/w [vp 才/d [vp 能/v [vp [vbar 看/v 清/a] [np 它/r 的/u 全貌/n]]]]]] 。/w]
[zj [fj [dj [np [vbar 多/d 变/v] 的/u 气候/n],/w [vp 更/d [vp [pp 为/p 它/r] [vp [vbar 增添/v 了/u] [np 神秘/a 的/u 色彩/n]]]]],/w [vp 甚至/d [vp 使/v 它/r [vp [vbar 孕育/v 了/u] [np 许多/m [np 美丽/a 的/u 神话/n]]]]]] 。/w]
[zj [dj [np 富士山/n 的/u 景色/n],/w [dj 四季/t 不同/a]] 。/w]
[zj [fj [fj [fj [fj 春天/t,/w [fj [dj 山顶/s [vp 还/d [vp [vbar 戴/v 着/u] [np 雪/n 帽子/n]]]],/w [fj [dj [dj [np 山腰/n 的/u 雪/n] [vp 却/d 溶化/v]] 了/y],/w [fj [dj [np 细碎/a 的/u [np 小/a 花/n]] [vp 开遍/v 山坡/n]],/w [vp [vbar 远/a 看/v] [vp 象/v [np [mp 一/m 片/q] [np 紫色/n 的/u 海洋/n]]]]]]],/w [fj 夏天/t,/w [fj [dj [np [np 残/Vg 雪/n] 与/c [np 山/n 花/n]] [vp 倒映/v [sp 湖/n 中/f]]],/w [vp 充满/v 诗情画意/n]]]],/w [fj 秋天/t,/w [fj [dj [np [np [np 满/a 山/n] 红叶/n] 与/c [np 雪/n 影/Ng]] 辉映/v],/w [vp 象/v [np 个/q [np 娇羞/a 的/u 姑娘/n]]]]]],/w [fj [dj 冬天/t [dj 则/c [vp 是/v [np [ap 纯/a 白/a] 的/u [mp 一/m 片/q]]]]],/w [ap 庄严/a 而/c 圣洁/a]]] 。/w]
[Mount Fuji (Fuji Yama) is an active volcano in Japan. Its summit, covered with snow, is surrounded by clouds all year round. Only in two dry seasons, fall and winter, can the whole clear mountain be peered. The constantly changing weather covers the mountain with a mysterious veil and has even brought about many beautiful mythologies. Different seasons in Mount Fuji meet different landscapes. In spring, the summit is being capped with snow while snow over the mountainside has already melted. Small and sparse flowers are blossoming all over the slope of the mountain, like a body of purple ocean in a distant view. In summer, the un-melted snow and mountain flowers are being mirrored in the lake to lure people into a fairyland. In fall, the mountain full of red leaves and the landscape of white snow mutually reflected in each other, are looking like a little shy girl. In winter, a vast piece of snow-white makes the mountain a solemn and sacred land.]

The characteristics of corpus research program of Peking University:

1. Large scale: The processed corpora are now up to 20 million words and will reach 27 million words in the near future. There is no precedent of this sort yet in China.
2. In-depth processing: Not only have the segmentation and part-of-speech tagging been implemented but also the phrasal structure of some parts of the corpora has been parsed and annotated and a treebank has also been

compiled. In the large-scale corpus, the place names and other proper names have all been labeled with the tags of their phrasal structure.
3. Wide range of coverage: The corpus of *People's Daily* not only includes news, but also many other domains, styles and genres, including various areas of social science and natural science.
4. High precision rate: Using a combination of manual and automated procedures, its precision rate has reached the highest level in China and compares to the best practice globally.
5. Solution of copyright problems: Peking University, Fujitsu Company and *People's Daily* have reached an agreement regarding copyright. No copyright issues will be raised.

1.3.3 *Beijing Language University*

Since 1996, Song Rou, a professor, in the computing department at this university, has been using large-scale language data for developing dependable statistical tools for the analysis of diverse language phenomena leading to the development of the proofreading system *YUANJING* (*distant view*). His corpus has been compiled from a variety of electronic sources obtained from newspapers and publishing houses and comprised 500 million words. This corpus was sorted using automatic and manual processing. Ten subcorpora were respectively set up. These corpora are:

- *China Today* (Series): 150 volumes (about 60 million Chinese characters)
- *People's Republic China Almanac*: corpus in 1997 (approximately 2 million characters)
- *Newspaper for Publication*: corpus (in 1988) (approximately 2.6 million characters)
- *Glorious Fifty Years — Hunan Volume*: corpus from 1949–1999 (approximately 700,000 characters)
- *People's Daily*: seven-year corpus from 1993 to 2000 (approximately 200 million characters)
- *People's Daily & Market Newspaper*: corpus in 2000 (approximately 14 million characters)
- *People's Daily — South China News*: corpus in 2000 (approximately 6 million characters)
- *People's Daily — East China News*: corpus in 2000 (approximately 5 million characters)
- *Economic Daily*: corpus in 1992 (18.2 million characters)

– *The Xinhua News Agency*: three year corpus from 1994 to 1996 (about 37.93 million characters)

Professor Song Rou also built a retrieval software tool called *Chinese Corpus Retrieval for Linguistic Research* (CCRL), which enables the users to use their own raw corpora and dictionaries to generate concordances for information retrieval.

In addition, Beijing Language University has compiled the following corpora:

– Contemporary Beijing Spoken Corpus (in 1992)
– Modern Chinese Grammar Research Corpus (in 1995)
– Modern Chinese Sentence Pattern Corpus (1995)
– Modern Chinese Corpus (established collaboratively with the Chinese and Bilingual Department at Hong Kong Polytechnic University in 1998)
– Modern Chinese Corpus (established collaboratively with Tsinghua University for the key project of the Natural Science Foundation of China, *The Theory, Method and Instrument of the Research of Corpus Linguistics*, 1998)
– Modern Chinese Circulation Corpus (2002)

1.3.4 *Tsinghua University*

This university has also compiled a modern Chinese corpus. In 1998, a corpus with 100 million Chinese characters was compiled to focus the research on the segmentation of ambiguous sequences. At present, the raw corpus has the volume of 700–800 million Chinese characters.

The team developed software to deal with 'false ambiguity', thus raising the precision rate of word segmentation.

Sun Maosong and Zuo Zhengping (1998) in the Department of Computer Science and Technology went beyond 'true ambiguous segmentation' and 'false ambiguous segmentation'. For instance: both "地面积" (/dimianji/) and "和软件" (/heruanjian/) belonging to the overlapping ambiguous segmentation strings, "地面积" is 'truly ambiguous', whereas ("这几块 | 地 | 面积 | 还真不小" [This land surface area is not so small] "地面 | 积 | 了厚厚的雪" [The deep snow covered the land surface]) and "和软件" should be considered to be 'falsely ambiguous'. Despite of the fact that there are two different sorts of segmentation: "和软 | 件" (/heruan|jian/, 'gentle piece') and "和 | 软件" (/he | ruanjian/, 'and the software') in authentic texts, these cases should be unexceptionally segmented as "和 | 软件" (/he | ruanjian/), so that it only can be read as 'and the software'.

"把手" (/bashou/, it means 'handle' or 'with the hand') and "平淡" (/ping-dan/, it means always 'wateriness') both belong to combinative ambiguous segmentation string. "把手" is considered to be truly ambiguous, while "平淡" is thought to be falsely ambiguous.

The team has also compiled the *Modern Chinese Lexicon for Segmentation* (现代汉语分词词表) for information processing which is becoming the most important language resource for word segmentation. Professor Luo Zhensheng in the Chinese Language Department has compiled the *Modern Chinese Sentence Pattern Research Corpus* (现代汉语句型研究语料库) from which 209 kinds of Chinese sentence patterns are generalized. The National Key Laboratory for Intellectual Technology and System (LITS) at Tsinghua University in cooperation with the Language Information Processing Institute at Beijing Language University designed and compiled the corpus *Hua Yu*. This corpus is different from other corpora in the following features: a balanced distribution with a wide range of coverage reaching far beyond the newspaper texts. The distributions of corpus *Hua Yu* can be seen from the table.

Table 1. The distribution of *Hua Yu* Corpus

Category	Number of Passages	Number of Chinese characters	Percentage	Punctuation marks	Number of Words (tokens)	Percentage
Literature	295	880,057	44%	148,453	760,337	48%
News	376	600,490	30%	86,163	438,095	28%
Academic articles	29	402,623	20%	52,823	278,728	18%
Practical writing	258	119,488	6%	28,727	91,929	6%
Total	958	2,002,658	100%	316,116	1,569,089	100%

Table 2. Distribution of literature corpus

Category	Number of Passages	Number of Chinese characters	Percentage	Punctuation marks	Number of Words (tokens)
Fiction	199	648,796	32.5%	112,749	566,730
Prose	37	80,067	4%	10,347	65,453
Memoir	29	50,401	2.5%	6,908	38,338
Reportage	13	50,019	2.5%	8,225	40,386
Drama	17	50,774	2.5%	10,224	49,430
Total	295	880,057	44%	148,453	760,337

This corpus has been segmented and annotated:

Paradigm:
我|rn 认识|vgn 王眉|npc 的|usd 时候|ng,|，她|rn 十|mw 三|mx 岁|qnm,|，我|rn 二|mx 十|mw 岁|qnm 。|。那时|t 我|rn 正|dr 在|pza 海军|ng 服役|vgi,|，是|vi 一|mx 条|qns 扫雷舰|ng 上|f 的|usd 三七|ng 炮手|ng 。|。她|rn 呢|y,|，是|vi 个|qng 来|vgn 姥姥|ng 家|ng 度假|vgi 的|usd 中学生|ng 。|。那|rn 年|qt 初夏|t,|，我们|rn 载|vgn 着|utz 海军|ng 学校|ng 的|usd 学员|ng 沿|pg 漫长|a 海岸线|ng 进行|vf 了|utl 一|mx 次|qv 远航|vgx 。|。到达|vgn 了|utl 北方|s 著名|a 良港|ng 兼|vgn 避暑|vgp 胜地|ng,|，在|pza 港|ng 外|f 和|pg 一|mx 条|qns 从|pg 南方|s 驶来|vgi 满载|vgn 度假者|ng 的|usd 白色|ng 客轮|ng 并行|vgi 了|utl 一|mx 段|qns 时间|ng 。|。进|vgn 港|ng 时|ng 我|rn 舰|ng 超越|vgn 了|utl 客轮|ng,|，很|dd 亲近|a 的|usd 擦|vgn 舷|ng 而|c 过|vgi 。|。兴奋|a 的|usd 旅游者|ng 们|ki 纷纷|dr 从|pg 客舱|ng 出来|vgi,|，挤|vgi 满|a 边舷|ng,|，向|pg 我们|rn 挥|vgn 手|ng 呼喊|vgi,|，我们|rn 也|dr 向|pg 他们|rn 挥|vgn 手|ng 致意|vgi 。|
*[When I first met Wang Mei, she was 13 years old and I was 20 years old. I was serving as a '**Sanqi**' gunner on a minesweeper in the navy at that time. While she was a middle school student, spending holidays at grandmother's home.*
In the early summer that year, we, students of naval academy went on an ocean-going voyage along the long coastline. We reached a well-known good harbor in the north, a summer resort, and sailed for some time side by side with a white vessel coming from the South, fully loaded with holiday spenders. While entering the harbor, our warship surpassed the passenger vessel with a very intimate touch on the side of the ship. Coming out from the cabins one after another and packing with the sides of the ship, excited tourists called to wave at us and we waved back.]

Recently, the National Key Laboratory for Intellectual Technology and System (LITS), attached to Tsinghua University, has compiled a Chunk Corpus (2 million Chinese characters), extracted from the *Hua Yu* corpus, and has annotated a small portion of this Chunk Corpus (including 200,000 words selected according to factors such as verb types, sentence length, etc.) The LITS has thus provided the syntactic annotation for a treebank. The samples of the sentence annotation are as follows.

– The sample of chunk annotation:

[从/p [他/rN 的/u 身上/s],/，[我们/rN [看/v 到/vB] 了/u] [一/m 位/qN] [跨越/v [中国/nS {CS 近代/t 和/c 现代/t } 的/u [共产主义/n 战士/n] [光彩照人/iV 的/u 楷模/n] 。/。]

– The sample of syntactic tree annotation:

[zj [dj [pp 从/p [sp 他/rN 的/u 身上/s]],/，[dj 我们/rN [vp [vp [vp 看/v 到/vB] 了/u] [np [np [mp 一/m 位/qN] [np [vp 跨越/v [tp 中国/nS [tp 近

代/t 和/c 现代/t]]] 的/u [np 共产主义/n 战士/n]]] [np 光彩照人/iV 的/u 楷模/n]]]]] 。/。
[We have seen a glittering model of a communist soldier crossing over Chinese modern and contemporary times from him.]

1.3.5 *Shanxi University*

The researchers at this university built a number of different corpora to fulfill different needs. The primary ones are:

- Chinese news corpus, 2.5 million words, built in 1988, including four parts:
  - *People's Daily*, 1.5 million Chinese characters;
  - *Beijing Science and Technology Daily*, 200 thousand characters;
  - *TV News* from China Central Television (CCTV), 500,000 characters;
- *Contemporary* (Journal), 300,000 characters.
- Standard-based Corpus: It is a comprehensive corpus in which *The Segmentation Criterion of Modern Chinese Words Used for Information Processing* (GB-13715) and the standard for PoS-tagging of the State Language Commission has been fully applied: 700,000 words of which 100,000 words are tagged on the level of fine granularity. This corpus can be used for the evaluation of word segmentation and PoS tagging.
- Special purpose corpora:
  - Annotation of Chinese place names: A corpus with 2.8 million characters has been tagged (containing 1793 different place names, the overall frequency of place names is 11,590) and a Chinese place name data-base was set up (containing 882,026 place names). A data-base of characters used for Chinese place names (3685 Chinese characters) and a data-base of Chinese words used for Chinese place names (3917 words) were compiled. The recognition precision is 86.7% and the recall is 93.8%.
  - Annotation of Chinese person names: A corpus of 3 million Chinese characters was tagged. A surname data-base was established. A frequency table of characters used for surnames (729 Chinese characters) and the frequency table of characters used for given names (3345 Chinese characters) were extracted from the corpus. The precision for recognizing Chinese person names is 87. 31% and the recall is 95. 23%.
  - Annotation of names translated from western languages: A corpus with 2.5 million Chinese characters has been tagged. The names translated from western languages are extracted from *The Translation Handbook of English Personal Names* and *Table of Common English Personal*

Names in the *New English-Chinese Dictionary*. The annotation precision is 63%, the recall is 98%.
  - Annotation of Chinese organization names: A corpus of 500,000 characters has been tagged. In the light of the rule-based approach, the syntactic structures of organization names have been parsed.
  - Annotation of new words and expressions from the Internet: A corpus of 1.5 million Chinese characters was tagged.
- Evaluation Corpus:
  - Data-base for the evaluation of overlapping ambiguous strings: A data-base with 78,000 characters has been compiled. In this data-base, every sentence is related to the overlapping ambiguous strings. Up to now, 5.1 million Chinese characters corpus has been segmented and the segmentation precision is 97%. At the same time, falsely-ambiguous strings which have an unambiguous tagging result account for 94% of the total amount of overlapping ambiguous strings.
  - Data-base for evaluation of combinative ambiguous strings: This data-base contains over 133 types of combinative ambiguous strings, giving an overall picture of the actual conditions of the combinative ambiguous strings in segmentation.

1.3.6 *Shanghai Normal University*

This university has compiled a raw corpus with 30 million Chinese characters. Based on the tagging norms of Beijing University, the team built an annotated corpus with 3 million Chinese characters. They also built an *Author's Digest* annotation corpus (1 million characters), selected from *Author's Digest* in 1997. The domains in this corpus include biographical literature, history, documentary literature, the personage's features, fiction, prose, commentaries, etc. Word segmentation, part-of-speech tagging, and the tagging of phrase structure relation and function were all carried out manually. This has been a very laborious task.

> Paradigm:
> [zw他/rp [db[zc期望/vz 着/ut]vp[db 打/vs [dz[sl一/mx 个/qi]mp[dz[zc 漂亮/ax 的/us]np[dz 大/ax 胜战/ng]np]np]vp]vp]jp 。/w
> *[He was anticipating winning a beautiful triumph in the battle.]*

Among them, zw (subject-predicate structure), db (verb-object structure), dz (modifier-head structure), sl (numerical-quantity structure) etc. are all tags of syntactical function.

1.3.7 *Treebank of Beijing University of Post & Telecommunication*

In this project, the grammatical rules of the Chinese Treebank of Language Data Consortium (LDC) in USA have been applied. The Chinese Treebank of LDC is composed of 325 articles of Xinhua News Agency from 1994 to 1998, containing 4185 trees and 100 thousand words. The Beijing project has built upon these results. The rules of grammar and parameters of analysis models are all acquired through the statistics and training of LDC Treebank.

Before extracting the rules, the following treatments were carried out in advance.

- Deletion of all empty words in the LDC Treebank;
- Removal of all the function labels of the non-terminal symbols in LDC Treebank;
- Removal of those nodes which have only a child node and this child node is tagged with a non-terminal symbol in LDC Treebank.

Based on these pretreatments and on the improved CYK algorithm, 3690 rules were automatically extracted. The format of rule is as follows:

parent_symbol|*current_symbol* → $RHS_1 \ldots RHS_n$	*log_probability*
Examples: *NP*|*NP* → *NN NN NN*	*−0.879602*

1.3.8 *The Contrastive Corpus of The Hong Kong City University*

The Language Information Research Center at this university has set up a Corpus of Linguistic Variety in Chinese Communities (LIVAC) to serve the purpose of surveying the similarities and differences of Chinese language varieties in different areas of Chinese Communities. This corpus has been planned since 1993 and data have been selected from five different areas: Hong Kong, Macao, Shanghai, Singapore and Taiwan. Every day one newspaper from these five different areas has been selected, extracted and put in storage. This corpus contains editorials, news and articles of the first edition, of the international edition, of local editions, features, comments, etc. Approximately 20,000 characters in all were collected every day. If the collected amount has already been more than 20,000 characters, some insignificant materials had to be deleted. During the two years from July of 1995 to June of 1997, the total number of characters in this corpus came to 15,234,551, in which about 8,869,900 words have been automatically segmented and manually proofread.

The statistics indicate that among the words used in all Chinese areas, double syllable words (types) take the first place, three-syllable words take the second place, four-syllable words take the third place and the monosyllabic words

take the last place. However, the frequency of monosyllabic words (tokens) is relatively higher, only next to the frequency of using double-syllable words and far beyond the sum of the frequency of other syllable words.

The statistics further indicates that the Macao and Hong Kong varieties have the highest similarity in word usage. Hong Kong and Taiwan, and Hong Kong and Singapore occupy the second place in the similarity in word usage. Shanghai and Hong Kong have the lowest similarity in word usage. In the light of the historical background and the social situation, this statistical result is acceptable. This is because Hong Kong is very close to Macao and both were once governed by the Western European countries for a long time. In addition, there is more common ground in the commercial situations and social structure between Hong Kong, Taiwan and Singapore than that between Hong Kong and Shanghai. Therefore, these features must have been reflected in word usage.

The statistics indicates yet that there are fewer different words (types) used in Singapore while there are relatively more different words (types) used in Shanghai. This can be explained by the fact that Singapore Chinese language is not the exclusive language in the local social life, whereas Shanghai has a special position in China and the economic activities are extremely active.

1.3.9 *Corpus of Taiwan*

The Taiwanese Academia Sinica has compiled a balanced corpus (Sinica Corpus) and a Tree Bank (Sinica Treebank). These two resources are both annotated corpora with a certain depth of processing.

The Sinica Corpus is specially designed for the analysis of language. All its sentences are segmented on the word level and have part of speech tagging. The composition of the corpus is intended to represent the distribution of modern Chinese according to domains and language type and to use the corpus as a reference corpus of modern Chinese.

This corpus has been jointly compiled by the lexical corpus team at the Information Institute and the Linguistics Institute at Academia Sinica. The team embarked on this project in 1990. The compilation of the corpus was begun in 1991. In 1994, it was subsidized by the Chinese Information Cross-Institute Research Group Project of the Academia Sinica and the Taiwan National Science Association. After funding was secured, it started officially the PoS-tagging and completed the first version (2 million words) in July 1995. In November of 1996, the computer center completed the www version available to the public, and in 1997 the 3.0 version with approximately 5 million words was completed.

The numbers of Chinese characters, numbers of words and numbers of passages in this corpus are as follows.

Table 3. The distribution of the Sinica Corpus

Theme	Total Number of Words	Total Number of Chinese Characters	Number of Passages
Literature	777,050	1,169,801	1385
Life	858,750	1,398,791	2301
Society	1,610,997	2,711,720	3246
Science	629,838	1,054,738	994
Philosophy	439,955	673,080	695
Art	474,340	781,415	518
Other	101,394	160,306	89
Total	4,892,324	7,949,851	9228

For more information on the content of this corpus, you can visit the following website: http://rocling.iis.sinica.edu.tw.

1.4 Spoken Corpus

1.4.1 *The Linguistics Institute at the Chinese Academy of Social Science*

In this project, a modern natural spoken language corpus, including a spoken corpus of restaurant reservation dialogues, was compiled. This includes data from telephone conversations (altogether two hours) in which over 200 individuals have been involved. The project includes prosody segmentation and syntactic tagging. The corpus was set up in the format of a 'wav' file in which speech was labeled with SAMPA-C and prosody was labeled with C-ToBI2.0. The file was transcribed into a Chinese character text file. It includes a non-restricted natural conversation corpus with 14.2-hours of conversation in which 22 individuals have participated. This conversation corpus also has a prosody segmentation and a syntactic annotation. It was also set up in the format of a 'wav' file with speech labeled using SAMPA and prosody labeled using C-ToBI2.0. The corpus was transcribed into a Chinese text file.

The Linguistics Institute is also compiling a corpus of spoken modern Chinese dialects. It is based on1500 inductive topics and a variety of communicative situations. The topic part using the topic inductive method takes up 60% of the corpus. The spoken data collected in natural situations in which the speakers were not aware that they were being recorded account for 40% of the corpus.

What is more significant is that the Linguistics Institute also compiled the *Beijing District On the Spot Impromptu Conversation Spoken Corpus*. This is one of the key projects of the Chinese Academy of Social Science which is meant to serve the purpose of investigating the miscellaneous dynamic mechanisms of on-the-spot impromptu talk. A large amount of data was collected with the aim to reveal the conventions of on-the-spot impromptu talk. In this project, approximately 500-hours of audio data and approximately 50-hours of video data were collected. Using a hierarchically categorized sampling approach, instructions for the sampling and the sampling scale were developed. The manual labor was used for on-the-spot recording of the materials.

1.4.2 *Institute of Automation at the Chinese Academy of Science*

This institute has compiled a spoken corpus of travel agency conversation and a spoken corpus of hotel reservations. The aim is to model our understanding of this domain field in view of possibly developing statistics-based spoken language machine translation technology.

1.4.3 *Beijing Language University*

This university has compiled the *Contemporary Beijing Spoken Corpus* (of recorded texts) with the intention to archive the surviving dialects and sociolects of Beijing according to residence areas, sex, age, professions, nationality, and the degree of education, This large-scale investigation will lead to a corpus that can be used to study the differences between Beijing dialects and standard Mandarin. It will also enable the study of the evolution of the Beijing dialects and sociolects. While recording the speech data, the goal was to keep the intervention to natural language to a minimum, in order to ensure the reliability of the data. Basic conditions to ensure representativity, including topic control and quantitative sampling methods were observed.

1.5 The Construction of bilingual corpora

1.5.1 *English Chinese bilingual corpora*

– The bilingual corpus at the Institute of Computational Linguistics at Peking University: 50,000 English and Chinese matching sentences have been collected. The alignment tools and the bilingual corpus managing software have also been developed. A Chinese and English contrastive phrase corpus is being built on this base and the prospected scale will be over 100 thousand entries.

- The English and Chinese Bilingual Corpus at Harbin Polytechnic University: In 1998, there were 30,000 pairs of sentences with PoS-tagging. The corpus is being expanded to 400,000–500,000 pairs of sentences on three levels: sentence level, phrase level and word level.
- The English and Chinese bilingual corpus at North-East University: It was built on the basis of the bilingual chunk corpus. In 1999, a 100,000 bilingual chunk corpus was constructed and a chunk-based English-Chinese machine translation experiment was conducted. Now a 1 million bilingual chunk corpus is under construction using a combination of acquisition and human proofreading. This corpus is intended to be expanded to a 5 million bilingual chunks and it will be further expanded to a large-scale online electronic English-Chinese collocation dictionary with 10 million chunks for the sake of the exploration of the acquisition algorithms of collocation phrase from electronic dictionaries and the construction of a large-scale online electronic English-Chinese collocation dictionary.
- Parallel corpus in Foreign Language Teaching and Research Press:
  The English-Chinese literature parallel corpus contains 30 million words. For example, the Chinese-English Contrastive Corpus of *The History of Philosophy of China* — this book was written by the famous Chinese philosopher Feng Youlan, and the corpus consists of a Chinese version and an English version, the English-Chinese Contrastive Corpus of *Scientific and Technical History of China* — this book was written by the famous scientist Josef Needham, and the corpus consists of the English and the Chinese version.
- The English-Chinese bilingual corpus at Institute of Applied Linguistics: Feng Zhiwei (1996) and other researchers at Institute of Applied Linguistics attached to the Ministry of Education have compiled an English-Chinese bilingual corpus. This corpus contains the language material from the field of computer science and philosophy (including Plato's *Politeia*). Using this bilingual corpus, they have tested the limit entropy of Chinese characters and investigated English-Chinese bilingual alignment issues.
- The English-Chinese Bilingual Corpus of the Institute of Software at the Chinese Academy of Science: The research of bilingual alignment algorithms is being conducted in this institute. At present, an English-Chinese bilingual corpus (150,000 sentences) has been aligned, segmented and annotated.
- The English-Chinese Bilingual Corpus of the Institute of Automation at the Chinese Academy of Science: The institute has purchased the Hong Kong News English-Chinese Bilingual aligned Corpus and the Hong Kong Legal

English-Chinese Corpus; 25,000 pairs of sample sentences from an English-Chinese Dictionary have been extracted. Using these resources, an English-Chinese bilingual corpus has been compiled.

1.5.2 *The Japanese-Chinese translation Corpus*

The Beijing Center for Japanese Studies at Beijing Foreign Language Study University has compiled a Chinese-Japanese parallel corpus. This corpus is mainly composed of Chinese and Japanese literature works and also includes drama, prose, political commentaries, both in the original and a translated version. The center has also collected several different translated versions for some literary works. There are altogether 20 million words in this automatically segmented and PoS-tagged corpus, some of which are grammatically and semantically tagged.

1.5.3 *The German-Chinese Corpus*

The School of Language and Literature at Ocean University of Shandong Province designed and compiled a small scale German-Chinese Contrastive Corpus for the sake of the contrastive research. This corpus is composed of the Chinese and the German text of Wang Meng's novel *Butterfly* 《蝴蝶》. They primarily compared the particle "le" (了) in Chinese with the perfect tense of German verbs.

1.5.4 *The Chinese-Japanese-English Idiom Corpus in the Computer Department at Fudan University*

This corpus contains thousands of idiom categories and tens of thousands of texts in three languages.

1.6 Minority language corpora

In China, there are 56 different nationalities. The Mongol language, the Uighur language, and the Tibetan language are important minority languages. At the beginning of the eighties of the 20th century, the Institute of Mongolian at the Inner Mongolian University, in cooperation with Inner Mongolian Computer Center, cooperated to keyboard the *Mongolian Secret History* 《蒙古秘史》, a monumental Mongolian historical work, and to process it with retrieval and analysis software. This resource became, in 1984, the 'Middle Ages Mongolian Corpus' of the Institute of Mongolian at the Inner Mongol University, the core of the Mongolian Corpus. From 1984 to 1990, funded by the Chinese National Foundation of Social Science, the Institute of Mongolian at the Inner

Mongol University also compiled a 'Modern Mongolian Database' (1 million words) which was assessed by the Inner Mongolia Autonomous Region Commission of Science and Technology as 'the first modern Mongolian database in the world'.

The database contains resources such as:

a. works of fiction,
b. Mongolian teaching materials (from grade one in primary school to grade three in high school),
c. newspapers (one month of *Inner-Mongolian Daily*),
d. political comments (one-year, 12 issues of *Practice Magazine*).

They respectively account for 19.6%, 50.3%, 9.8% and 22.9% of the corpus.

The corpus contains only texts published and issued domestically after 1949. Since at the time of the completion of the Mongolian database, many domestic and overseas corpora had already reached or surpassed the scale of 5 million words, the corpus was enlarged to the scale of 5 million words. Subsidized by the Foundation of Human, Art and Social Science at the China Ministry of Education, a 'Modern Mongolian Database' with 5 million words was compiled in addition to some language material from the fields of mathematics, physics, chemistry, medical science, law, etc. However, this is a raw corpus.

After the 90s of the 20th century, the Institute of Mongolian at the Inner-Mongolian University built a corpus for Khitan small characters (契丹小字, 2000), a corpus for Pags-pa scripts (八思巴文字, 2001), and a Mongolian Spoken Corpus.

In 2000, in combination with the Chinese-Mongolian machine translation system, the institute compiled a Chinese-Mongolian Contrastive Government Documentary Corpus (approximately 200,000 words). The institute is also compiling a Chinese-Mongolian parallel corpus (1.5 million-words), to be used with the exampled-based Chinese-Mongolian machine translation engine.

All the previously mentioned Mongolian corpora have their own goals and features. They differ in scale, hierarchy, purpose, and the degree of processing. Among them, the Modern Mongolian Corpus is at the center of corpus research at the Institute. It has now taken shape and will provide the linguistic basis for Mongolian information processing.

Due to the fact that during the initial compilation phase of the Modern Mongolian Corpus, the encoding guidelines used in China were word-form-based and therefore unable to display or store the different variants of the Mongolian alphabet with same pronunciation, the institute adopted the transcription approach and transcribed the corpus from the Mongolian alphabet into the

Latin alphabet. A set of transcription rules had to be developed. During this initial phase, some specialized word forms, diverse morphological attachments, person names, place names and compound words were tagged manually.

The institute has also conducted some tentative research in corpus processing. For instance, to resolve the problem of proofreading the corpus input, the institute designed an automatic verification program for the Latin-transcription of the Mongolian texts. Subsequently, it conducted a series of fundamental studies concerning automatic segmentation of Mongolian word stem, root, ending, etc. Currently, the institute is tagging the phrases in the Mongolian corpus, working on an approach that integrated the traditional Mongolian grammar with the current research of grammatical and semantic properties of Mongolian words. The PoS-annotation of this Mongolian corpus makes it a useful resource for the technical support of Mongolian information processing. The institute is now investigating semantic processing, and its results will be applied to the semantic annotation of the corpus.

All the Mongolian corpora have now adopted the Latin transcription. Recently the institute has been conducting an experiment of building a Mongolian parallel corpus using the XML format.

The Uighur is a minority nation in Xingjiang region of China. Uighur language processing is developing rapidly now. Xingjiang Normal University has compiled an Uighur corpus (4 million words) and a machine readable dictionary including 50,000 words with their grammatical features. The researchers at this university are carrying out PoS tagging, syntactic parsing and a semantic analysis of Uighur language.

Xingjiang University has also compiled an Uighur corpus (1.5 million words). A multiple level automatic processing of the corpus, including an automatic text classification, are being developed.

The Tibetan language is another important minority language in China. The Institute of Minority Language at the Chinese Academy of Social Science has compiled a Tibetan corpus (5 million words) and will also develop the segmentation and annotation guidelines for the Tibetan language. A spoken Tibetan corpus is being compiled at present.

1.7 Compilation of an English corpus

In the mid-eighties of the 20th century, the Shanghai Jiaotong University, under the direction of Yang Huizhong (2003) compiled a Corpus of English of Science and Technology, known under the acronym JDEST. From early on, English corpora in China were used for foreign language teaching. JDEST,

with its word-frequency statistics, was a great contribution to the Chinese College English Syllabus. This corpus won much attention and great respect from corpus linguists in Europe at that time. JDEST was the first of a generation of corpora used for language teaching all over the world. Later on, more English corpora in China have been compiled: the ICLE China Sub-corpus (Gui Shichun), the Corpus for English Learners in China (Gui Shichun, Yang Huizhong), the College Learner's English Spoken Corpus (Yang Huizhong), the Chinese English Major Learner's Spoken Corpus (Wen Qiufang), the CEC Chinese English Corpus (Li Wenzhong), the Middle School English Spoken Corpus (He Anping), etc.

These corpora are all used for foreign language teaching and learning in China. The reasons are as follows: On the one hand, many corpus researchers of English corpora in China are themselves English teachers. On the other hand, due to the intellectual property rights involved, most of the corpora cannot be used by other Chinese researchers. The researchers of English corpus linguistics in China have to develop their own corpora for which they have the copyright rather than to rely on the half-hidden and half-exposed attractions of western corpus resources, if they want to conduct genuine and seminal research.

Another important feature of English corpus linguistics in China is that since its initial phase, it is inclined towards extensive cooperation and shared resources. This is mainly due to the following facts:

a. It is a matter of cost-effectiveness. A case in point is that nearly ten-year's endeavor of many linguists, statisticians, and software engineers was needed to develop the COBUILD corpus into the present Bank of English with its 450 million words (Sinclair 1991). Thus, large-scale corpora cannot only rely on a single university or organization.
b. The Chinese academy system does not give a prompt and sufficient financial support to this kind of interdisciplinary corpus research that brings together the arts and humanities with the social sciences, with natural science and technology, and with computer technology. The Chinese academy system treats the research of corpus as a common discipline of the arts and humanities which is outside of its funding remit. Such unfavorable conditions result in a serious underfunding of Chinese corpus research. Under the extremely arduous conditions, the researchers conduct their work by having to cooperate with institutions outside China.
c. The third feature of English corpus linguistics in China is its application-oriented tendency and intensive self-awareness. For years, most of Chinese

research in foreign language learning had been devoted to the introduction and interpretation of western linguistic theories while there has been not enough independent research fuelled by the linguistic traditions of China. This situation had made it difficult for Chinese foreign language teachers and researchers to develop their own theoretical frameworks, and thus it had limited the contribution they could make to the international research community. The applied research of corpus has now offered a new platform for English teachers and researchers from China on the international arena, to display their achievements, and to share and exchange their ideas by presenting the cornucopia of their achievements. Therefore, the intimate combination of corpus-driven applied research and the reality of English teaching and learning is a deliberate choice of scholars in the field of English language teaching in China that will propel corpus research in China and give Chinese linguistics an important voice in the international research community.

2. Corpus technology

2.1 Automatic segmentation

With regard to automatic segmentation for Chinese written text, the following technique of ambiguous segmentation are proposed: the relaxation approach (Fan C.K. & Tsai W.H. 1988), the approach based on ATN (Augmented Transition Network) (Huang Xiangxi 1989), the approach based on PSG (Phrase Structure Grammar) (Liang Nanyuan 1984; Kit Chunyu, Liu Yuan & Liang Nanyuan 1989; Yao Tianshun, Zhang Guiping et al. 1990; Yeh C.L. & Lee H.J. 1991; Han Shixin & Wang Kaizhu 1990; Liu Ting & Wang Kaizhu 1998), the approach based on Expert System (Xu Hui & He Kekang et al. 1991), the approach based on neural networks (Xu Bingzheng & Zhan Jian et al. 1993), the approach based on FSA (Finite State Automation) (Sproat R. & Shih C. et al. 1996), the approach based on HMM (Hidden-Markov Model) (Shen Dayang et al. 1997; Sun Maosong & Zuo Zhengping et al. 1998), the approach based on Brill's TBED (Transformation-Based Error-Driven, proposed by Eric Brill in 1993) (Palmer D.D. 1997, Feng Zhiwei 2001).

In addition, the technique of named entities recognition (person name recognition, place name recognition, organization name recognition) and new word recognition are also being studied.

2.2 Automatic annotation

In the case of automatic PoS tagging, the rule-based approach is primarily intended to resolve part of speech ambiguity. The researchers also use statistics-based approaches. Main approaches are: CLAWS algorithm, VOLSUNGA algorithm, HMM, TBED approach, etc.

2.3 Automatic phrase structure annotation

The results of phrase structure tagging can be expressed in the form of Phrase-Tree (P-tree) or in the form of Dependency-Tree (D-tree). Some systems adopt the transformational technique from P-Tree to D-Tree proposed by Feng Zhiwei (2001). Others adopt CYK algorithm (Cocke-Younger-Kasami algorithm) to parse the corpus sentences and to identify their phrase structure. Most commonly, phrase structure is expressed by bracketing.

2.4 Bilingual Alignment Technique

This technique primarily adopts the length-based approach, dictionary-based approach, and the mixed approach (the combination of length-based and dictionary-based methods). Both the Institute of Computational Linguistics at Peking University and Institute of Software at Chinese Academy of Science are exploring how to use aligned parallel corpora for the (semi-automatic) compilation of corpus-based bilingual dictionaries.

3. Several problems with the construction of corpora in China

3.1 Guidelines and standards for corpora

Researchers in the field of Chinese information processing have embarked on designing a national standard: *Standard for Word Segmentation in Chinese Information Processing*《信息处理用现代汉语分词规范》. The draft of this standard was tested over three years and revised several times. Eventually it was approved to become the national standard. The shelf mark of this standard is GB/T13715-92.

The main structure of the standard consists of five sections: topic and content, scope of application, referenced standards, terms, summary, and detailed instructions. Since there is no clear dividing line between morpheme, word, and phrase in Chinese, the standard, in addition to mostly adopting the definition

of the word in linguistics documents, particularly introduces the notion and definition of the 'segmentation unit' as "the fundamental unit with definite semantic and grammatical functions used in Chinese information processing". In addition, the standard points out that the segmentation unit refers to 'the word and phrase as prescribed in the standard'. This innovation brings to an end the dispute about the definition of the word and harmonizes the conflicts in the academic community.

Another achievement is the *List of Commonly-Used Modern Chinese Words for Information Processing* 《信息处理用现代汉语常用词表》. Due to the extreme complexity to distinguish the morpheme, word and phrase in the modern Chinese language, almost every rule for this distinction has its exceptions. Therefore, the standard proposes a principle for segmentation: 'close combination and stable application' of the segmentation unit, this principle is used as a guideline for assessing *whether or not* a string can be regarded as a segmentation unit. Nevertheless, this principle is not specific enough to be understood by the operators, who inevitably segment the words according to their own understanding of the standard. As a result, there are no uniform norms for different systems. Therefore, some researchers have suggested that apart from the standard there should also be a word list on the basis of the standard to further clarify how to implement the standard. Adopting the strategy, 'standard plus word list' has proved to be a farsighted approach. In 1994, using the statistical results of word frequency of modern Chinese, the key designers of the standard, Liu Yuan and his colleagues, issued the *List of Commonly-used Modern Chinese Words for Information Processing*, containing 43,570 lexical entries. Unfortunately the list does not tackle the bottleneck appropriately in word segmentation and therefore is not authoritative enough in this domain.

In Taiwan, a different *Word Segmentation Norm for Chinese Information Processing* 《资讯处理用中文分词规范》was designed. This norm introduces three fundamental principles for word segmentation. (1) The word segmentation unit should accord with the requirement of linguistic theories. (2) Word segmentation should be practicable regarding information processing. (3) The segmented words should ensure the consistency of text processing. In addition, there were also some supplementary principles (the principles of merging and the principles of segmenting), to determine whether to merge or segment. This norm, based on the degree of difficulty and simplicity, offers word segmentation guidelines on three different levels: faithfulness level, expressiveness level and refinement level. The level of faithfulness is the criterion for normal information exchange. The level of expressiveness is the criterion for natural language processing such as machine translation and information retrieval.

The level of refinement is the highest virtual level for word segmentation. This approach is very useful for coping with word segmentation.

In addition, the State Language Commission compiled a *Standard Word List for Chinese Information Processing* 《信息处理用现代汉语规范词表》 for the sake of designing a list of modern Chinese words of general use from the perspective of the government. This word list was meant to serve as a standard. The planned size of the list is up to 60,000 to 80,000 lemmas. This project is still in the process of completion. The modern Chinese vocabulary is a complex system including scientific terminology, dialect words and phrases, classical Chinese words and phrases, named entities (person names, place names, names of organizations etc.), diverse idioms (idioms, proverbs, set phrases etc.), apart from the generally-used words. The vocabulary is constantly evolving. With the development of the society, there will be more new words and phrases with which information processing will have to cope. Consequently, this project will have to include setting up scientific term lists in different domains, dialect word lists, a list of classical Chinese words, lists of named entities, idiom lists, lists of neologisms etc., in addition to the table of words in general use. This is a huge task. It will, in due course, have a profound impact on the further development of Chinese information processing.

In addition, China drafted a *Norm of a PoS Tag Set for Chinese Information Processing* 《信息处理用现代汉语词类标记集规范》. Such a norm was originally proposed by the Computational Linguistics Section at the Institute of Applied Linguistics at Ministry of Education. It lists 18 upper-level categories of PoS. There are three main principles for devising this norm. (1) The grammatical function of a word is the primary principle for the classification. The sense of word sometimes plays a subservient role. (2) It is permitted that words have multiple PoS tags. From the perspective of statistical research, some modern Chinese words possess multiple grammatical functions, which have different probabilities of distribution. (3) The upper-level categories of PoS tag sets should cover the entire modern Chinese vocabulary. It is planned that this norm will become a national standard.

During the last two decades, some Chinese researchers were keeping a close watch on the evolution of International Standard Generalized Markup Language. The Association for Computers and the Humanities (ACH), in cooperation with the Association of Computational Linguistics (ACL) and the Association of Literature and Language Computation (ALLC), initiated the Text Encoding Initiative (TEI) in 1989. This initiative was meant to devise a set of uniform encoding norms for electronic texts in order to standardize natural language corpora and to facilitate the exchange of real language data.

The Corpus Encoding Standard (CES), initiated collaboratively by European MULTEXT, EAGLES, and VASSAR/CNRS, is now widely applied in the compilation of new corpora.

In 1986, ISO officially issued the international standard, Standard Generalized Markup Language (SGML). Its shelf number is ISO 8879–1986. In 1995, China also adopted SGML as the national standard with the shelf signature GB 14814. In 1998, Feng Zhiwei published an essay in which he explained the SGML language for a Chinese audience: *Standard Generalized Markup Language and Its Application in Natural Language Processing in Contemporary Linguistics*.

Extensible Markup Language (XML), a sub-set of SGML is now widely used as the meta-language for corpus annotation. Using Document Type Definition (DTD) and Schema to standardize XML documents, XML has excellent flexibility of extension and contraction, and it can separate the mark-up from the content and the norm from the result. The construction of Chinese corpora will apply SGML and XML in the meta-linguistic description.

3.2 Availability of sharing the corpus resources

The practices for sharing corpus resources are as follows:

a. Selling the corpus as a product;
b. Implementing a membership system;
c. Granting user rights;
d. Providing academic non-profit organizations with a user right.

In 2003, the Chinese Linguistic Data Consortium (Chinese LDC) was founded. The Chinese LDC was founded voluntarily by the researchers in the domain of the construction of Chinese language resources (text corpora, speech corpora, dictionaries, etc.). It is an academic, non-profit association. This organization is a powerful social instrument established to improve the construction of Chinese language resources, and to assist Chinese language information processing with language resources. The goal of the association is to unite the researchers in the area of Chinese language resources and to establish a reference corpus of the Chinese language, which can be viewed as balanced and representative. The consortium will also assist the basic research and applied exploration of Chinese information processing, enhancing the performance of the available technology. The establishment of the Chinese LDC will boost Chinese corpus linguistics.

3.3 Intellectual property rights of corpora

The more extensively corpora are used, the more serious becomes the issue of intellectual property rights. All corpora derived from published materials encounter this problem. It is strongly suggested that the relevant government departments develop applicable guidelines and empower the Chinese Information Processing Association to deal with this problem.

3.4 Statistical noise in corpus processing

Due to the increased use of electronic texts, we have more and more access to the acquisition of corpus resources. In China, there is already a large reference corpus with 500 million Chinese characters. The Data Collection Initiative (DCI) of the Association of American Computational Linguistics has pointed out that if real language data is stored in text format, corpus size can easily be extended to trillions of words (tokens). There appears to be a serious problem, however. While statistically analyzing successive pairs of strings in corpus, Song Rou (1996) has found out that 'with the size of the corpus increasing, the noise in newly added successive pairs will gradually occupy most of the linguistic data'. The noise is mainly caused by errors in named entity recognition in word segmentation. While it is essential to identify all novel language phenomena, for instance neologisms, in the added corpus materials, they have to be distinguished from this kind of noise. Examples for such new words and phrases are: "喷塑" (spray plastic), "蒜农" (garlic farmer), "危改" (the reconstruction of the houses in danger), "市话" (city call), etc. All these new words and phrases can be found hiding in the noise caused by erroneous segmentations and misprints. Therefore tools are needed to separate relevant information from obstructive noise. This will become a brand new project for the processing of very large corpora.

References

Brill, E. (1993). Transformation-based Error-driven Parsing. *International Workshop on Parsing Technologies.*

Huizhong, Yang. (2003). Opening Address on the 2003 International Conference on Corpus Linguistic, Shanghai, Shanghai Jiaotong University.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sproat, R., Shih, C. et al. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3), 377–404.

Zhiwei, Feng. (2001). Hybrid Approaches for Automatic Segmentation and Annotation of a Chinese Text Corpus. *International Journal of Corpus Linguistics*, Special Issue, 2001.
北京语言学院语言教学研究所, (1985), 汉语词汇的统计与分析，外语教学与研究出版社。
陈鹤琴, (1928), 语体文应用字汇，商务印书馆。
冯志伟, (1996), 汉字的极限熵，载《计算机时代的汉语和汉字研究》，清华大学出版社。
冯志伟, (1998), 标准通用置标语言SGML及其在自然语言处理中的应用，《当代语言学》, 1998年，第4期。
冯志伟, (1999), 语料库语言学与机器翻译，信息网络时代与日本研究，山东大学出版社。
冯志伟, (2001), 计算语言学基础，商务印书馆。
顾曰国, (2003), 北京现场即席话语语料库的取样与代表性问题, 全球化与21世纪（首届“中法学术论坛”论文集，社会科学文献出版社。
韩世欣, (1990), 基于短语结构文法的自动分词系统，中文信息学报，第4卷，第4期。
揭春雨，刘源，梁南元, (1989), 汉语自动分词方法，中文信息学报，第3卷，第1期。
梁南元, (1984), 书面汉语的自动切分与一个自动分词系统，北京航空学院学报，1984年，第1期。
刘开瑛, (2000), 中文文本自动分词和标注，商务印书馆。
刘挺，王开铸, (1998), 关于歧义字段切分的思考与实验，中文信息学报，第12卷，第2期。
刘挺，吴岩，王开铸, (1998), 串频统计和词形匹配相结合的汉语自动分词系统，中文信息学报，第12卷，第1期。
宋柔, (1996), 二元接续关系及其在分词和校对中的应用，ICCC'-96, National University of Singapore, Singapore.
孙茂松，左正平, (1998), 汉语真实文本中的交集型歧义，汉语计量与计算研究。
孙茂松, (1999), 高频最大交集型歧义切分字段在汉语自动切分中的作用，中文信息学报，第13卷，第1期。
姚天顺, (1990), 基于规则的汉语自动分词系统，中文信息学报，第4卷，第1期。
俞士汶、朱学锋、段慧明, (2000), 大规模现代汉语标注语料库的加工规范，中文信息学报，第14卷，第6期。

Author's address

Prof. Feng Zhiwei
Institute of Applied Linguistics
Chaonei Nanxiaojie 51
100010 Beijing
China