Recent Developments in Chinese Corpus Research

Zhan Weidong[+,++], *Chang Baobao*[++], *Duan Huiming*[++], *Zhang Huarui*[++]

[+] Department of Chinese Language & Literature, Peking University
[++] Institute of Computational Linguistics, Peking University
{zwd; chbb; duenhm; hrzhang}@pku.edu.cn

Abstract

In this paper, the author firstly gives a brief overview of the history of developing Chinese corpora in mainland of China, especially focusing on some representative research projects in the last decade, such as the *General Contermporary Chinese Corpus* that is sponsored by the State Language Commission of China National Ministry of Education, and the *Chinese Corpus of Situated Discourse in Beijing Area* that is built up by China Academy of Social Science, and so on. And then the related works in this field made by Peking University on designing, annotating and using of corpus are elaborated. There are four parts are discussed in detail, including (1) a very large scale of wide time-span Chinese corpus using for linguistic research with an on-line KWIC concordance based on Web-Lucene search engine, (2) People Daily corpus which is processed with word segmentation and part-of-speech tagging, (3) a Chinese Treebank. Based on the Treebank, Chinese phrasal constructing rules can be extracted automatically, and the distribution of all kinds of phrases can be described through statistical approach. (4) a Chinese-English parallel corpus based on which a workbench prototype has been built to support Chinese-English lexicography. In the latter part of this paper, the author discusses briefly some issues, which have received more attention in this field recently, including the standardization of Chinese corpora encoding and the approaches to share large-scale Chinese corpora for researches and public use.

1. Introduction

From Ferdinand de Saussare to Noam Chomsky, the mainstream of modern linguistics continuously claims that a distinction can be drawn between langue, or language competence, and parole, or language performance. Only the former can be regarded as the scientific object for serious study. Although the distinction, to some extent, does make sense to linguistic research, the research results based only on the intuition or introspection of linguists are often insufficient to be used for related applications such as natural language information processing and second language learning, and so on. In such kind of fields, large-scale corpora, from which useful and practical knowledge about natural language can be extracted conveniently with help of computer software, are paid more and more attention. Nowadays corpora have reached a new stage with large scale, wide variety and greater accessibility.

In recent years, substantial progress has been made on Chinese corpora. This paper will give a broad overview of the present development of Chinese corpora. The rest of the paper is organized as follows. Section 2 will present a brief overview on the history of the development of Chinese corpora. Section 3 describes in detail several projects related to Chinese corpus implemented by Peking University. In section 4, the factors that affect the spread and use of Chinese corpora are discussed briefly. Section 5 will conclude this paper.

2. The brief history of the development of Chinese corpora

The history of the development of Chinese corpora can be roughly divided into three stages [8]:

In the first stage, from very early 20th century to 1980's, the age of pre-computer in China, people used to collect manually Chinese printed texts as corpora, on which frequency of Chinese characters can be counted by hand. The purpose of using corpora in this stage is to learn about the actual usage of Chinese characters in real world context according to statistical data. This work is apparently valuable to help the compilation of Chinese textbooks used for Children's learning to read and write Chinese characters in primary schools.

In the period between 1980's and the early 1990's, Chinese corpora enter into a new stage with using computer to store and process digital documents. In general, the size of a Chinese corpus in this period amounts to millions or even ten millions of Chinese characters. In order to count the frequency of Chinese words rather than characters in real world context, researchers segmented Chinese sentences into word sequences by hand with the aid of computer in this period. In contrast to corpora in the first stage, the basic unit of corpora in the second stage is word instead of Chinese character. The main applications on Chinese corpora in this stage include (1) compiling Chinese word frequency dictionary, (2) selecting most frequently-used words for using in Chinese textbooks, (3) drafting the specification for Chinese word segmentation, which was revised and finally issued as the national standard (numbered GB-13715) in October of 1990. The national standard, title as *The Segmentation Criterion for Modern Chinese Used for*

Information Processing, is the first guideline for automatically segmenting Chinese written language.

Since the middle of 1990's, computer and software on natural language processing are used more broadly in developing of Chinese corpora, including not only written text material but also colloquial Chinese, with deeply annotation. Some new trends of development of Chinese corpora in this stage can be outlined by the representative Chinese corpora as follows.

- Very large and deeply annotated corpus

Here we take the Contemporary Chinese corpus developed by the State Language Commission of China National Ministry of Education as an example [18]. The corpus, which was hoped to act as the Chinese national corpus, started to be built from 1991. Up to now, it has already contained 100 million Chinese characters. Half of texts in the corpus have been processed with word segmentation and part-of-speech tagging, among which there are 50 thousand sentences, or about 1 million Chinese characters, have been parsed and annotated for syntactic structures. In this corpus, the proportion of the documents wrote in the period from 1919 to 1992 is 70%. The other 30% is the documents wrote in the period from 1993 to 2002. There are 85% documents typed in by hand from previously printed texts, and 25% documents loaded into the corpus by downloading from the Internet in existing electronic format. Genres of the corpus include: (1) social sciences and humanity account for 50%, (2) natural sciences account for 30%, (3) various business and official documents account for 20%. Each sample text in the corpus contains about 2,000 Chinese characters in principle. Moreover, more than 20 metadata categories about the text, including the author, the title, the publisher, the date of publishing, and so on, are used in description of each sample.

- Multimedia corpus

Besides corpora constructed from common written texts, spoken Chinese in real world discourse, especially the materials from multimedia such as video and audio, have extracted much attention recently.

For instance, Chinese Academy of Social Sciences has been involved in developing such a corpus, titled as “Chinese Corpus of Situated Discourse in Beijing Area” since 1999[9],[10]. The goal of the corpus project is to collect information of human speech activities and discourses happened in various situations, including, discourses in the place of daily working, chats between family members in their own apartment (different types of family are took into consideration in the projects), utterances and behaviours presented in various types of social occasions (e.g. banquet, party) by the way of recording tapes and videos. Based on the so-called multimodel texts instead of only the traditional printed texts, it is more likely for discourse analysts to discover the complicated relationship among the social situations, individual behavior, and personal linguistic performance. Up to now, there are 600 hours audio and 50 hours video that have been collected.

The another instance worthily to be noted here is an ongoing corpus project that is devoted to collect the language materials from TV programs and broadcasting in China, including video, audio and also printed texts used for media. The corpus, titled as “Chinese media language corpus” is developed by Communication University of China [12], [17]. At present, there are 340 hours of materials from TV program, and 50 hours of materials from broadcasting that have been added to the corpus. Meanwhile, the program transcriptions that contain almost 50 millions Chinese characters are also added to the corpus. The transcriptions from TV program count for 80%, and the other 20% is from broadcasting program. It is no doubt that such kind of corpus is valuable as a reference for Chinese national language planning and researches of sociolinguistics.

- Cross-language corpus

Due to the rapid development of corpus-based machine translation and emerging of computer aided bilingual lexicography in China, cross-language corpus or parallel corpus, especially Chinese-English parallel corpus, is paid more and more attention in recent years. For instance, Peking University embarked to build a Chinese-English parallel corpus from 2002. Up to now, more than 1 million Chinese-English sentence pairs have been added to the corpus. The detailed description of the corpus can be seen in section 3.4.

- Special purpose corpus

In order to survey the actual conditions of learning Chinese by foreign students in China, Beijing Language & Culture University built the Chinese interlanguage corpus in 1995[7],[16]. The developers of the corpus collected 5,774 compositions and other exercises written in Chinese by 1,635 foreign students who came from 96 countries and regions to learn Chinese in 9 universities of Beijing and other cities. The original materials contain 3,528,988 Chinese characters, from which researchers extracted 1,731 texts (approximately 44,218 sentences, 1,041,274 characters) written by 740 students. Each sample in the corpus is annotated with 23 metadata categories to give the source description, including student’s name, sex, age, nationality, mother language, education degree, used textbooks, writing time, contributor, etc. Various types of errors, including wrongly used Chinese characters and words, unwell-formed sentences, are tagged and indexed for retrieval. The corrected counterpart of each error is also noted for the comparison between the error format and the correct format. Based on the data extracted from the interlanguage corpus, language teachers and textbook editors can know more about the factors that affect the effectiveness of Chinese learning and then adopt more active and effective measures to improve teaching performance and to compile better textbooks.

As the growing popularity of corpus methodology in Chinese academic communicty, there will be more Chinese corpora for special purpose following the forerunners, such as the LIVAC (Linguistic Variations in Chinese Speech Communities) synchronous corpus, has been developed by Hong Kong City University since 1995, is devoted to offer a window to investigate the lexical development of contemporary Chinese in different Chinese communities [13], [20], and the child language corpora were built in the mainland and Hong Kong during the last decade for the purpose of language acquisition research [2], [3], [19].

In addition, as the rapid development of Internet, the idea that the world-wide web can act as an immense, multilingual, freely available corpus is now spreading in academic communities. As an attempt to handle the tremendous amount of language examples in the Web, researchers in Tsinghua University are now developing a system for extracting example sentences from web pages, more precisely speaking, re-extracting from the search results returned by mutli web search engines, such as Google, Baidu, and so on [15]. The system is apparently different from the common web search engines, which return web pages or links to web pages as the responses to user's query. Tsinghua's system works as usual KWIC concordance system by listing only the sentences containing the key words as results. Besides the common functions as in usual concordance softwares, the system can delete automatically the repeated sentences from the set of results. And it also can detect automatically the collocations of the key words from the retrieved sentences by means of statistical approaches.

3. Corpus projects in Peking University

In Peking University, there are two organizations that are related to building Chinese corpora. One is the Center for Chinese Linguistics (CCL) of Department of Chinese Language & Literature, which is engaged in Chinese language research and teaching, the other is the Institute of Computational Linguistics (ICL), which is engaged in Chinese information processing.

It is well known that the language resources are very important to both theory-oriented linguistic researches and application-oriented Chinese information processing. A natural language processing system that falls short of linguistic data cannot work very well. Since 1986 when it was founded, ICL was continuously devoted to building Chinese language resources, including various Chinese lexicons correspond to different linguistic levels (i.e. morphemes, words, syntactic, semantic), which started to be built in the early stage, and Chinese corpora, which were being developed in recent years. ICL and CCL keep always a firm relationship of collaboration on developing Chinese language resources. In their works, the following three guidelines that are continuously kept in researchers' mind [25]:

(1) The research achievements in the Chinese linguistic theory should be taken fully advantage of in developing Chinese language resources. For instance, the set of Chinese part-of-speech tags, functional classification of phrasal structures, which play fundamental roles in annotation of Chinese corpus, are established on the basis of the researches on Chinese grammar system along the road of linguistic structuralism.

(2) Software based on the technology of natural language processing should be adopted broadly in building language resources for time and money saving. Accordingly, the developed language resources can correspondingly be used to improve the performance of natural language processing softwares.

(3) Both the language knowledge database and corpora can take their proper roles in different applications. And the two kinds of representation of language knowledge can be integrated as unified and comprehensive language resources.

The language resources developed in Peking University are listed in the following table (The tick mark stands for available).

Linguistic Resources		Contemorary Chinese	Traditional Chinese	Bilingual
language knowledge databasae	grammatical lexicon	√		
	semantic lexicon	√		√
	conceptual lexicon	√		√
	terminological database	√		√
	phrase structure rules	√		
corpus	sentence sgemented corpus	√	√	√
	word segmented & POS tagged corpus	√		
	treebank	√		

Table 1: Linguistic Resources of Peking University

The interactive relation between language knowledge base and corpus can be recognized more clearly by showing the workflow of processing of text into corpus (see Figure 1).

In general, it is necessary to take advantage of language knowledge base for corpus annotating automatically. The scale and quality of knowledge database can affect remarkably the accuracy of automatic annotating. And once annotated corpus is available, more reliable language knowledge data can

be automatically extracted to enlarge the scale of existing language knowledge-bases. It could be expected that the workflow will bring continuous progress for whole language resources.

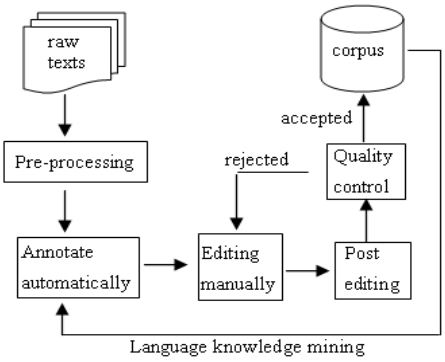

Figure 1: Workflow of Corpus Processing

If language resources are used more widely, it will be more possible to improve the quality of the resources. Besides delivering language resources by charging licence fee, we also provide language resources in part for public use freely via website at http://ccl.pku.edu.cn and http://icl.pku.edu.cn.

In the rest part of this section, we will give descriptions in detail about the 4 corpus projects in Peking University, including (1) a very large scale of wide time-span Chinese corpus, which is processed with sentence segmentation (denoted as PKU-CCL-CORPUS). (2) People Daily corpus, which is processed with word segmentation and part-of-speech tagging (denoted as PKU-ICL-PD-CORPUS). (3) Chinese Treebank (denoted as PKU-CTB). (4) Chinese-English parallel corpus (denoted as PKU-ICL-CE-CORPUS). Each corpus will be described from two perspectives: (1) the basic information about the corpus, such as corpus size, the manner of annotating, etc. (2) the applications and research works that rely on the corpus.

3.1 PKU-CCL-Corpus and its on-line query system

3.1.1 Basic information

The goal of building PKU-CCL-CORPUS is to collect Chinese written texts covering different eras as large as possible for using in Chinese language research and teaching, especially for finding proper examples from the corpus. On the one hand, such kind of corpus should be the larger the better, on the other hand, saving time and money should be taken into consideration seriously. The strategy we adopted includes two aspects. (1) The texts are not processed with word segmentation at all but annotated with some basic information, such as author, title, and so on. (2) The supporting query system could be developed on the basis of open source.

The project started from later 2003. Up to January 2006, The texts written in traditional Chinese in PKU-CCL-CORPUS have contained approximately 101 million Chinese characters (486 documents, 54 folders, 202,305,825 bytes), and the texts written in modern Chinese have contained 115 million Chinese characters (157 documents, 23 folders, 229,700,435 bytes). The total size of corpus data is 412 MB (approximately 200 million Chinese characters). The detailed information about composition of the corpus can be accessed via the website of CCL.

3.1.2 An on-line query system

An on-line query system supporting the PKU-CCL-CORPUS, named CCL-Lucene, has been developed based on the open source Lucene and its XML interface Web-Lucene.

Lucene is a great search engine purely based on Java technology [1]. A Lucene index consists of a sequence of documents. Each document consists of a sequence of fields. Fields have a name. The value of a field is a sequence of terms. As long as the texts in a corpus are organized logically in this structure, the corpus could be input into the Lucene indexer to create the so-called inverted index file easily. Lucene supports both incremental and batch indexing. The former feature allows easy adding of new texts to an existing index that is apparently critical for feeding a corpus with live data periodically to keep the pace of language changing. Based on the index, Lucene search engine can response user's query very quickly.

Web-Lucene [21] is an expansion of Lucene by adopting XML format as standard input stream, producing search result in XML format and providing result sorting with highlight support. The CJKTokenizer that supports to analyze Chinese, Japanese and Korean characters, as well as western languages, is also integrated into Web-Lucene.

CCL-Lucene is a customized Web-Lucene that can meets some special requirements of searching in PKU-CCL-Corpus. The following is a list of what we have done on CCL-Lucene.

- Defining the XML fields that will be used for corpus search

The original texts are segmented into a sequence of sentences and then converted into a XML-format document. The XML document consists of the fields, such as sentence, author, path, title, and so on, which are indexed by Lucene indexer. By searching in those fields (e.g. author, title), a corpus user can specify more clearly the scope of corpus searching.

- Designing the special query language

Besides supporting common query expressions, e.g. logical operators like AND, OR, some additional operators are added to CCL-Lucene query language. For example, for a user searching in corpus usually wants to query two words that maybe not occur close to

each other. In this case, the distance between the two words and the order of the two words need to be specified in user's query. The query expression "把 $5 打", which can be accepted by CCL-Lucene query analyzer, means to retrieve the sentences that contain two words "把" (ba3) and "打"(da3) in the conditions that the word "把" occurs before the word "打" and the distance between the two words should be less than 5 Chinese characters. Another example is to search some given patterns in Chinese, such as duplicated words, a kind of derived word format in Chinese. By the query expression "pattern: AABB", we can retrieve the sentences that contain the substring with the pattern AABB, where A and B stand for one Chinese character respectively.

- Optimizing the display of search result

The search results displayed by CCL-Lucene can highlight multi key words and put one of the key words at the center of each line in the result page. The search results also can be sorted by the right or left string around the key word. User can specify the exact number of Chinese characters around the key word to be displayed.

CCL-Lucene also can be used to provide searching service for the other corpora, such as those will be discussed below, with some modifications.

3.1.3 Usage report

With the support of Apache log4j package, it is very easy to track usage of an online corpus. PKU-CCL-Corpus and its online query system are published on CCL website in later December, 2004. According to the web log from 15:13:19, 2004-12-22 to 11:16:38, 2006-1-10, CCL-Lucene system runs on Resin Java server for 370 days. During this period, 18,020 IP addresses had accessed to the corpus. The total of requests is 574,009, among which there are 185,491 requests with unique form of query expression. That means 1,551 requests per day in average. The highest number of requests in one day is 9,794, which is recorded on 29 January 2005. In the table below, we list the top 10 words that are most frequently searched (Xiandai means searching in modern Chinese whereas Gudai means searching in traditional Chinese).

No.	User Query	Scope	Count
1	把	Xiandai	1280
2	很	Xiandai	1088
3	有	Xiandai	1028
4	给	Xiandai	931
5	被	Xiandai	858
6	下	Xiandai	794
7	的	Gudai	784
8	上	Xiandai	753
9	都	Xiandai	728
10	得	Gudai	716

Table 2: A list of most frequently searched words

The web log of corpus usage shows that it is a practical way to provide corpus service via Internet. By analyzing the web log, it is interesting to find what words and phrases are paid more attention by users.

3.2 PKU-ICL-PD-CORPUS

3.2.1 Basic information

At present, PKU-ICL-PD-Corpus consists of two years of newspaper articles with more than 50 million Chinese characters. In 1999, ICL, cooperating with Fujitsu, started to annotate a one-year newspaper corpus from People's Daily of China, which contains more than 26 million Chinese characters, approximately 14 million words [26]. The annotating work lasted for about three years. In early 2002, the one-year newspaper corpus (denoted as PD-1998) was built with word segmentation and part-of-speech tagging. ICL selected a sub-corpus of one whole month newspaper articles from the annotated texts, which was published in January 1998 on People's Daily, and put it on ICL website for free download (denoted as PD-Jan-1998). It contains more than 2 million Chinese characters. The set of part-of-speech tags used to be assigned to each word in PD-1998 contains 48 tags [24], evolved from a basic set of 26 tags, which is defined in the grammatical lexicon, titled as *Grammatical Knowledge Base of Contemporary Chinese* (GKBCC), which was developed by ICL over the past twenty years [27]. Soon after the accomplishment of building PD-1998, ICL started to annotate independently another one-year newspaper corpus, which was also published in People's Daily in 2000. It has the similar scale with PD-1998. At this time, a set of 106 tags was adopted [25], which is evolved from the set of 48 tags used to annotate PD-1998. Both of the two evolvements of tag set are from a coarse grain of tag set to a fine grain of tag set. That means the tag set used in the later corpus is compatible with the tag set used in the former one. There are two changes in the first evolvement. (1) Some word classes are subcategorized in terms of semantic features of words. For instance, noun is subdivided into nr (personal name), ns (name of place), nt (name of organization), nz (other proper nouns), nx (string consists of non-Chinese characters). (2) Some tags are added to represent functional variations of words. For instance, the tag "vn" is added to mark the verb that functions as a noun, and the tag "vd" is used to mark the verb that functions as an adverb. In the second evolvement of tag set, there are two kinds of changes too. (1) Some word classes are subcategorized in terms

of syntactic features. For instance, verb is subdivided into “vq (directional verb)”, “vx (dummy verb)”, “vu (auxiliary verb), etc. (2) Some tags are added to mark individual words, especially the Chinese particles. For instance, the tag “ui” stands for “地” (the adverbial marker in Chinese), the tag “us” stands for “所” (the nominalizing prefix in Chinese), the tag “ud” stands for “的”(the possessive and the structural particle in Chinese), etc.

3.2.2 Research works based on PKU-ICL-PD-CORPUS

Due to the very large scale and the syntactic annotation, PKU-ICL-PD-CORPUS could serve as an adequate linguistic data resource for both natural language processing and linguistic study. Here are some instances of research works based on the corpus.

(1) It is conceivable that the performance of a Chinese part-of-speech tagger based on statistical model can be improved remarkably, once it uses PKU-ICL-PD-Corpus as the training set, from which the n-gram data can be extracted easily. [28].

(2) For each word in the corpus is assigned a tag to represent its grammatical function, we can investigate the usage of words more detailedly. For instance, The word "在"(zai4) is ambiguous as it has 3 different part-of-speech tags. By investigating the actual occurrences of each senses of the ambiguous words in the corpus, we can learn that “在” functioned as preposition (p) counts 95.22%, “zai4” functioned as adverb (d) counts 2.42%, “在” functioned as verb (v) counts 2.36%. Similarly, the word "把" (ba3) is also ambiguous as it has 5 different tags. “把” functioned as preposition (p) counts 95.83%. “把” functioned as classifier (q) counts 2.80%. “把” functioned as verb (v) counts 1.14%. “把” functioned as numeral word (m) counts 0.14%. “把” used as a morpheme and functioned as noun counts 0.08%. Moreover, it is also easy to measure how many instances of all ambiguous words exist in the corpus. According to a survey implemented by [14] on PD-Jan-1998, in which contains 843,887 Chinese word tokens, and 44,600 Chinese word types. 397,777 word tokens are ambiguous (47.13%), and 10,079 word types are ambiguous (22.60%) correspondingly. It is obvious that ambiguous lexical items are very common in real world Chinese contexts.

In addition to calculate the frequency of words and tags in the whole corpus, we can investigate the occurrence frequency of each word in divided ranges of the corpus, e.g. the texts within a month, for the newspaper corpus can be divided naturally according to the time range. If a word is commonly used in a language, it will occur in different parts of a corpus. And if the word is used commonly enough, it will be well-distributed. Given this assumption, we can use a new statistical measure, called as Distributional Consistency (DC), together with word frequency to estimate distribution of a word. The DC of a word can be defined as follows:

$$DC = \frac{((\sqrt{F_1} + \sqrt{F_2} + ... + \sqrt{F_n})/n)^2}{(F_1 + F_2 + ... + F_n)/n}$$

Where n stands for the number of equally sized parts into which the corpus is divided, F_n stands for the occurrence frequency of a word in the n^{th} part of the corpus [29]. A corpus can be divided into parts according to different criteria, such as time range, domain, size, etc.

Once the concept of DC is established, we can avoid the problem of selecting words to build a core lexicon only depending on word frequency. For instance, the word “抗洪 (kang4 hong2, fight against the flood)” is used in high frequency in PD-1998. “抗洪/v” (the word is used as verb), which occurs 2322 times in the corpus, has frequency rank 766. “抗洪/vn” (the word is used as noun), which occurs 4622 times in the corpus, has frequency rank 347. It is no doubt that the word “抗洪” should be regarded as a frequently used word just according to the data above. However, once the DC rank of “抗洪” is taken into consideration, we have to take an opposite opinion. The DC of “抗洪/v” is 0.46516 at the rank 38,256, and the DC of “抗洪/vn” is 0.45074 at the rank 39,117. If we take a closer look at the actual use of the word in the corpus, we can find that the most occurrences of the word are in the texts within the three months from July to September of 1998 when the most areas along the Yangzi River of China were experiencing the disaster of flood. From the perspective of distribution, the word “抗洪” is not a real frequently-used word. Depending on the two statistical measures, i.e. frequency and DC, the property of words can be described in the way as follows. (a) The words that have high rank of both frequency and DC are stop words for common information retrieval systems. (b) The words that have high rank of frequency and low rank of DC should be selected by language teacher preferentially. (c) The words that have low rank of frequency and high rank of DC can be used to distinguish documents from each other. (d) The words that have low rank of both frequency and DC are really few used.

(3) As an implementation of language knowledge mining, the grammatical attribute-value pairs in GKBCC are re-described by taking advantage of statistical data acquired from PKU-ICL-PD-Corpus. For instance, each verb in GKBCC will be marked whether it can be modified by the Chinese most frequent negators “不”(bu4) and “没”(mei2) by two features, also named as “bu” and “mei”. Both features are assigned Boolean value, i.e. “yes” or “no”. Here we take the verb “到”(dao4, arrive) as an example. For the verb can be modified by both of the two negators, the values of both features “bu” and “mei” for the verb are “yes”. However, In PD-1998 corpus, 2045 occurrences

of the verb “到” are negated by “不”, and only 9 occurrences are negated by “没” [22]. The original Boolean value in GKBCC could be changed into numerical value to describe the probability of the verb “到” taking each negator as adverbial modifier.

In addition, there are several ongoing research works in Peking University based on PKU-ICL-PD-Corpus, including Chinese word sense disambiguation, automatic detection of Chinese basic phrases, e.g. basic noun phrase, prepositional phrase, temporal phrase, numerical classifier phrase, and so on.

3.3 PKU-CTB

3.3.1 Basic information

PKU-CTB, the ongoing project, started in later 2003 with the goal of creating a corpus of 1 million Chinese words with syntactic bracketing. The corpus will consists of four parts: (I) Chinese government white papers, (II) newspaper articles, (III) Chinese textbooks of preliminary/middle/high schools, (IV) test sentences used for machine translation evaluation. The sentences are collected from different sources, such as grammar books and linguistic research papers. The scale and composition of PKU-CTB at present are listed in appendix II.

There are many differences between PKU-CTB and the Pennsylvania Chinese Treebank (Penn-CTB) [5]. 33 part of speech tags are used in Penn-CTB. The most part of speech tags used in PKU-CTB follow the guideline of PKU-ICL-PD-Corpus with tiny modification. At present, there are two levels of tags with different granularities used in PKU-CTB. On coarse-granularity level, we have 26 tags, and on fine-granularity level, we have 97 tags. Penn-CTB has 17 phrases categories and 26 functional labels for marking the function of a phrase. PKU-CTB has 22 phrases categories (see appendix for detail) with only one additional marker “!” (exclamation mark). The marker is used to denote the head of a phrase that is preceded by the marker “!”. Given the phrase category label and the head marker assigned to a phrase, it is likely to mark the functions of its constituents automatically for most phrases. So we plan to annotate the function of each phrase in the later phase relying more on computer instead of human work. In addition to the difference of phrase labels between the two corpora, the difference of phrase bracketing is also distinct. In Penn-CTB, the tree structure of a sentence is drawn out under the paradigm of generative grammar whereas the guideline of drawing the tree structure for a sentence in PKU-CTB is the paradigm of traditional structuralism, especially the method of immediate constituent analysis. As a result, Penn-CTB assumes the null elements for analysis of sentence structure, which are not defined in PKU-CTB.

The workflow of building PKU-CTB can be described as follows. (1) An original text is segmented into lines of sentences automatically by the indication of punctuations. (2) The sentences are processed with word segmentation and part of speech tagging. (3) The annotated sentences are parsed by a Chinese Chart-based parser into tree structures. (4) The trees are edited and corrected by human annotator with the help of a visualized tool, TreeEditor. (5) The trees can pass over the quality control procedure will be accepted as the “gold standard” results finally.

By using of TreeEditor, treebank annotator can work directly on the graph of a tree instead of a sequence of brackets and words, which is apparently not a good interface for tree structure editing. The operations can directly be taken on a tree graph, including to change the location of a tree node, delete a tree node, merge two nodes, modify the label of a node, and so on. TreeEditor also provides advanced find & replace function that can be used for retrieving tree structures. It is very convenient for annotator to locate at the targeted sub trees and to change them into new structure in batch mode.

3.3.2 Extracting linguistic knowledge from PKU-CTB

It is well known that treebank corpus can be used as the training data for statistical parsers. We can also extract context free rules with frequency automatically from treebank for improvement of rule-based parsers by extending the coverage of rules and adopting score mechanism. The related research work is now under progress in Peking University.

Through the data acquired from treebank, we can find more linguistic facts. Here are some examples.

(1) Extracting phrase structure rules with occurrence frequency

4,853 rules are acquired from the current PKU-CTB. Among these rules, only 1180 rules, count 24.31%, occur more than 5 times in the corpus. Three-quarter rules are used in a very low frequency. This fact can be used to explain in part why human-written rules in traditional rule-based parsers are insufficient to work for parsing real world texts. On the one hand, the low frequently used rules are very easy to be ignored by rule writers. On the other hand, such rules have the high percentage of distribution in real world natural language texts. Moreover, it is worthy of noting that a rule in very low frequency, mostly one occurrence, may indicate a human error of annotating. It is a practical way to find a possible error in treebank by checking the rules with low frequency.

(2) Extracting phrase examples according to various conditions, e.g. the root label, the labels of branches, depth and width of a tree, which can be specified by treebank users. The following table illustrate the widths of prepositional phrases (pp) in T-I.

SPAN	2	3	4	5	6	7	8	9
freq.	231	189	135	102	88	76	54	37
SPAN	10	11	12	13	14	15	…	45
freq.	35	30	18	16	15	13	…	1

Table 3: SPAN frequency of preposition phrase

In the above table, the term "SPAN" means the length of phrase, i.e. the number of words a phrase covers. There are 1,093 **pp**s in T-I corpus. Among those **pp**s, the percentage of **pp**s whose span value is less than 8 is 80%. The biggest span is 45.

We also can extract all phrases to count the average depth and width of each kind of phrase. The table in appendix III shows the average depth and width of most phrases in T-I corpus.

(3) Surveying the distribution of phrases

By extracting the context-free rules from treebank, we can survey the distribution of phrases to some extent. Given a context-free rule in the form of A → B C, where A, B, C stand for phrase categories. The distribution of phrase B can be described in part by A and C for A is the father node of B and C is the local context in which B exists.

The table in appendix IV shows the distribution of the first 10 **pp**s that have the most occurrences in T-I corpus (The mark "##" stands for empty or null, The tag "ude1" stands for the structural particle "的"). In the real world Chinese texts, prepositional phrase is used most frequently as the adverbial modifier of vp (see the first row in table). The punctuation comma is often used between prepositional phrase and its head phrase when the head is acted by dj or fj (see the second and the third row).

Such data extracted from treebank corpus as we demonstrated here would be helpful for Chinese language researchers and teachers to have a closer look at Chinese phrase structure at deep level.

3.4 PKU-ICL-CE-CORPUS

3.4.1 Basic information

Since 2002, ICL has been working with collecting Chinese-English texts and aligning them at the sentence level [6]. The following table lists the statistics reflecting the current scale of the ICL Chinese-English corpus.

Number of files:	7,228
Number of files: (Chinese-English direction)	2,074
Number of files: (English-Chinese direction)	5,154
Number of aligned Sentence pairs	1,023,077
Number of Chinese Sentences	1,059,825
Number of English Sentences	1,084,614
Number of Chinese Characters	29,431,702
Number of English Words	18,014,460

Table 4: Statistics of PKU-ICL-CE-Corpus

Among all the 7,728 files, 188 files are spoken language texts. They give 288,047 pairs of aligned sentences. In this spoken language portion, the average length of Chinese sentences is 10.323 characters and the average length of English sentences is 7.220 words. The other 7,040 files are written language texts and give 735,030 pairs of aligned sentences. Compared with the spoken language, the average length of both Chinese and English sentence is much longer. The length of Chinese sentences is 34.323 Chinese characters in average and for English sentence the average length is 20.10255 words. If counting the Chinese characters, 10.15% of the corpus is spoken language, which has 2,986,922 Chinese characters, and written texts account for 89.85% of the whole corpus with 26,444,780 Chinese characters.

To build the parallel corpus, once the source file and its translation file are collected, they are firstly automatically aligned by our sentence alignment software, and the results are then handed over to human verification with the help of software tools. At last the aligned texts are encoded with an XML-based markup language we defined for the parallel corpus.

The sentence alignment program developed by us is length-based and uses Dynamic Programming algorithm. It is language independent and could be easily used for constructing parallel corpus for language pairs other than Chinese and English. Actually, we are now also constructing a Chinese-Japanese corpus and a Chinese-Korean corpus, but both are still in very small scale.

3.4.2 Research works based on PKU-ICL-CE-Corpus

The Chinese-English corpus is no doubt very useful for the Chinese-English comparative lexical and grammar study. We extracted all the English sentences contained the word "by" and their Chinese translation from a small portion of PKU-ICL-CE-Corpus, to discover the knowledge that could be used by a MT system to correctly translate passive-form English sentences into Chinese. Among the 1,262 passive-form English sentences we got, 165 sentences (13%) can not be translated into passive-form Chinese sentences. The word order must be changed dramatically when translating these sentences. Therefore it's difficult for a MT system to cope with. Here are two examples shown as follows:

A. In this circumstance, **more memory *is required***.
这种情况下，就需要更多的内存。

B. In order to determine its breaking point, the simulated bridge deck has ***been tested*** for durability and strength by the expert group.

为了确定这个模拟桥面的断裂点，专家小组测试了它的耐力与强度。

For example A, the word "需要" is rarely used in passive form in Chinese. For example B, although the word "测试" has passive form "被测试", but it normally is not used in passive form when follow a purposive clause. Apparently, parallel corpus works in such comparative studies.

Chinese-English Translation Equivalent Pairs could also be automatically extracted from the parallel corpus. We used four different statistical measures to extract TEPs on a collection of 500 pairs of Chinese Enlish sentences and got the following results (the accuracy here means the ratio of the correct results in the top 100 results) [5]:

	Mutual Information	DICE coefficient	Log-likelihood	χ^2 score
accuracy	44%	6%	80%	81%

Table 5: Performance variations of different statistical measurements for TEP extraction

The following table shows some automatic TEP results ordered by χ^2-score.

Chinese	English	χ^2-score
成人_图书馆	adult_library	68620.5
影子_董事	shadow_director	68469.8
幕_墙	curtain_wall	68469.8
卤味_店	lo_mei	68282.1
橡胶_手套	rubber_glove	68041.9
橡胶_围裙	rubber_apron	67723.5
疾病_津贴	sickness_allowance	67433.1
计算机_软件	computer_software	67281.6
软_雪糕	ice_cream	67281.6
污水_隧道	sewage_tunnel	66626.8
工程_原理	engineering_principle	66626.8

Table 6: TEPs extracted automatically from Chinese-English parallel corpus

The TEP-extraction work is apparently useful for constructing bilingual lexicon for Machine Translation. It also can be used in corpus-based lexicography. At present, lexicographers are more and more interested in compiling dictionaries using corpus. We, cooperating with a major publisher in China, are developing a software platform for bilingual lexicography based on Chinese-English parallel corpus [4]. The platform under developing consists of two parts: (1) A Corpus Service Manager running on a high-speed server provides corpus and lexicon query service to the lexicographers. (2) Workbench for Lexicographers running on the users' personal computers provides friendly interface to the lexicographers. The two parts are linked together by Internet. Lexicographers are served by the Corpus Service Manager through the local workbench and make reasonable decisions for dictionary entry writing. The Corpus Service Manager can serve to multiple users at the same time, the services it provides include: querying the corpus, displaying query result as usual KWIC concordance, listing possible collocations, and so on.

4. The external factors that influence the development of Chinese corpora

There are two aspects of factors, which influence the development of Chinese corpora: internal factors and external factors. By "internal factor", we mean the scale and the quality of a corpus. The internal factors are apparently related to not only the technology of natural language processing, by which a corpus can be annotated more effectively, but also the linguistic theory, by which we can develop a better guideline to guarantee linguistic correctness and consistency of an annotated corpus. By "external factor", we mean the factors that influence availability, accessibility and usability of a corpus. Among such kind of external factors, there are two common issues related to Chinese corpora. One is the specification of corpora, and the other is the delivery channel of corpora.

The specification used for building a corpus usually includes two parts: (1) the corpus encoding standard to specify the valid format of annotating, and (2) the linguistic guideline to certify the well-accepted analysis for a linguistic unit. Apparently, the later issue is more complicated than the former. For there is lack of morphological cues in Chinese, it is more possible to be inconsistent in annotating a Chinese corpus by different research institutes than processing western languages. The problem is increasingly serious from annotating at the shallow level, e.g. word segmentation and tagging, to annotating at the deep level, e.g. syntactic bracketing and semantic relation tagging. We think it is no strange there have divergences between different researchers, and it is not only unnecessary but also impossible to make the request of commonly agreed guidelines for Chinese corpus annotating. In present stage, what we should pursuit for is to make different Chinese corpora more interoperable, in other words, the format of annotating of different corpora should be in common with each other. On the one hand, the growing popularity of XML in the Internet era can intensify the

migration to XML format for present Chinese Corpora. On the other hand, the influential encoding standards related to corpus annotating, such as Text Encoding Initiative (TEI) and Dublin Core (DC), should be mandated for use in building Chinese corpora. Further more, both the increasing demand for using corpus and more channels for delivering corpus could facilitate the normalization of corpus encoding.

Although a comprehensive delivery channel for spreading of Chinese corpora is still under construction, two efforts have been taken in the past few years. In 2003, Chinese LDC, an organization that is devoted to collect, create and share Chinese linguistic resources, was established with the support of the national high-tech and fundamental project from Ministry of Science and Technology of China [30]. Up to now, Chinese LDC has received and managed 41 items of Chinese language resources that can be accessed via its official website, compared with 313 items listed by LDC at University of Pennsylvania, which was founded in 1992. However, it is worthy of attention that the most linguistic resources managed by Chinese LDC now are collected mainly from the community of information technology of China rather than the discipline of linguistics. Chinese linguists and language teachers in China have yet not paid enough attention to activities of Chinese LDC. Meanwhile, it should be take into consideration for Chinese LDC to design proper terms in corpus license for different cases of using a corpus, for instance, using corpus for language study or using corpus to develop NLP system, online using or offline using, etc. By this kind of approach, it is possible for Chinese corpora to attract more users.

In the last two years, in order to accelerate and promote development of Chinese language resources, Ministry of Education of China, jointly with five universities respectively, established five centers for observing and analyzing the situation of contemporary Chinese language in social life. The five centers are the National Observation Center for Newspaper Language, setup in Beijing Language and Culture University, the National Obersvation Center for Cyber Language, setup in Huazhong Normal University, the National Obersvation Center for Multimedia Language, setup in Communication University of China, the National Observation Center for Overseas Chinese Newspaper Language, setup in Ji-Nan University, the National Observation Center for Textbooks Language, setup in Xiamen University. As one of the main task, the centers will build very extensive scale monitor corpus, which is updated periodically to record the changing nature of language. We hope the consciousness of corpus can be extended to broader areas, and public use of corpora data can be easier as the centers make great progress in developing modern Chinese corpora.

5. Conclusions

In this paper, we have presented the three phases of the development of Chinese corpora. There are four types of corpus that have been paid more attention by researchers in recent years. We gave a full introduction to the research works on Chinese corpora in Peking University. Corpora used to be favored by researchers in the field of Chinese natural language processing rather than Chinese linguists and language teachers. But as Chinese corpora are developed with broaden and deepen annotation, and correspondingly, using corpora with the help of software tools more conveniently, we can expect that Chinese corpora can be of better service for Chinese language study, Chinese teaching, and lexicography, etc., not be regarded as just a bank of language examples.

6. References

[1] Apache Lucene： http://lucene.apache.org/java/docs/index.html

[2] Cantonese - English Bilingual Child Language Corpus developed by Chinese University of Hong Kong: http://www.cuhk.edu.hk/lin/home/bilingual.htm

[3] Cantonese Child Language Corpus developed by Chinese University of Hong Kong: http://humanum.arts.cuhk.edu.hk/~cancorp/

[4] Chang, Baobao (2005), 基于语料库的双语词典编纂平台的构建，第六届全国双语词典学术研讨会。2005.11.26-28，广东外语外贸大学。

[5] Chang, Baobao, Pernilla DANIELSSON and Wolfgang TEUBERT (2001), Extraction of Translation Unit from Chinese-English Parallel Corpora, In: Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora, Birmingham University Press, Dec. 2002. Proceedings of 6th TELRI European Seminar on Multilingual Corpus Research, Nov. 2001.

[6] Chang, Baobao, Xiaojing Bai (2003), 北京大学汉英双语语料库标记规范，载 Journal of Chinese Language and Computing《汉语语言与计算学报》（新加坡），Vol.13, No.2，pp.195-214。

[7] Chu, Chengzhi, Xiaohe Chen(1993), 建立“汉语中介语语料库系统”的基本设想,《世界汉语教学》，1993 年第 3 期，pp.199-205.

[8] Feng, Zhiwei (2002), 中国语料库研究的历史与现状，载 Journal of Chinese Language and Computing《汉语语言与计算学报》（新加坡），Vol.12, No.1, pp.43-62。

[9] Gu, Yueguo, (2002), Sampling and representativeness in compiling SCCSD (written in Chinese), In Globalization and the 21st Century, pp.484-500. Beijing: Social Sciences and Archive Press.

[10] Gu, Yueguo,(2002), Towards an understanding of workplace discourse. In Christopher Candlin, ed. Theory and Practice in Professional Discourse. The City University of HK Press. pp.137 - 185.
[11] Guo, Huizhi, Hua Liu, Xuemin Xie, Pu Zhang, (2005) 《人民日报》标注语料的初步统计分析, 载孙茂松、陈群秀主编《自然语言理解与大规模内容计算》清华大学出版社 2005 年版。pp.187-192。
[12] Hou, Min (2005), 传媒语言语料库的建设与应用，语言学研究手段现代化问题学术研讨会, 2005.11.12-13，北京大学。
[13] LIVAC Corpus developed by Hong Kong City University: http://www.LIVAC.org
[14] Penn Chinese Treebank：http://www.cis.upenn.edu/~chinese/ctb.html
[15] Sun, Mao Song, Xiance Si, Wei Qiao (2005), 基于 Web 的汉语例句检索，语言学研究手段现代化问题学术研讨会，2005.11.12-13，北京大学。
[16] The Chinese interlanguage corpus developed by Beijing Language & Culture University: http://www.blcu.edu.cn/kych/H.htm
[17] The Chinese Meida Language Corpus developed by Communication University of China: http://ling.cuc.edu.cn/
[18] The Contemporary Chinese Corpus developed by the State Language Commission of China: http://www.china-language.gov.cn/, http://219.238.40.213:8080/
[19] The website of The Child Language Research Center of Huazhong Normal University: http://www.childes.cn/
[20] T'sou, Benjamin K., Tom B.Y. Lai, (2003), 汉语共时语料库与信息开发，载徐波、孙茂松、靳光谨主编《中文信息处理若干重要问题》，北京：科学出版社 2003 年版。pp.147 - 165。
[21] WebLucene：http://sourceforge.net/projects/weblucene/
[22] Yu, Shiwen, Huiming Duan, Xuefeng Zhu(2001), 汉语词的概率语法属性描述. 《语言文字应用》，2001 年，第 3 期，pp.21-26。
[23] Yu, Shiwen, Huiming Duan, Xuefeng Zhu, Bin Sun, Baobao Chang (2003), 北大语料库加工规范：切分· 词性标注· 注音. 载 Journal of Chinese Language and Computing《汉语语言与计算学报》（新加坡），2003 年 6 月，第 13 卷 2 期，pp.121-158。
[24] Yu, Shiwen, Huiming Duan, Xuefeng Zhu, Bin Sun (2002), 北京大学现代汉语语料库基本加工规范. 《中文信息学报》，2002 年，第 16 卷第 5 期，pp.49-64；第 6 期，pp.58-65。
[25] Yu, Shiwen, Huiming Duan, Xuefeng Zhu, Huarui Zhang (2004), 综合语言知识库的建设与利用, 《中文信息学报》2004 年第 5 期，pp.1-10。
[26] Yu, Shiwen, Huiming Duan, Xuefeng Zhu, Yasuhito Tanaka (2001), 大规模标注汉语语料库开发的基本经验. 载 Journal of Chinese Language and Computing《汉语语言与计算学报》（新加坡），2001 年，第 11 卷第 2 期，pp.101-110。
[27] Yu,Shiwen, Xuefeng Zhu, Hui Wang, Huarui Zhang, Yunyun Zhang, Dexi Zhu, Jianming Lu, Rui Guo (2003), 《现代汉语语法信息词典详解》（第二版），清华大学出版社 2003 年版。The Grammatical Knowledge-base of Contemporary Chinese – A Complete Specification (2nd Edition), Tsinghua University Press.
[28] Zhang, Huaping, Qun Liu, Xue-Qi CHENG, Hao Zhang, Hong-Kui Yu (2003). Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp.63-70, July 2003, Sapporo, Japan
[29] Zhang, Huarui, Chu-Ren Huang, and Shiwen Yu (2004), Distributional consistency: a general method for defining a core lexicon. In the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon, Portugal. May 26-28, 2004.
[30] Zhao, Jun, Bo Xu, Maosong Sun, Guangjin, Jin (2003), 中文语言资源联盟的建设和发展，载徐波、孙茂松、靳光谨主编《中文信息处理若干重要问题》，北京：科学出版社 2003 年版。pp.218 - 228。

Appendix I: Tags for phrases in PKU-CTB

Tag	Description	Tag	Description
hl	headline	npt	name of organization
zj	full sentence	npx	non-Chinese characters
yj	phrase surrounded by quotation marks	npz	proper noun
dj	finite clause	pp	prepositional phrase
fj	clause	qp	quantifier/classifier phrase
		sp	locative phrase
ap	adjective phrase	tp	temporal phrase
dp	adverbial phrase	vp	verb phrase
mp	numerical phrase		
np	noun phrase	yp	discourse element
npr	multi words proper name	ypc	parenthetical unit
nps	name of place	yph	form of address

Appendix II: Statistics of PKU-CTB

Corpus	Character types	Character tokens	Word types	Word tokens	Sentences	Average sentence length	phrases	rules
T-I	1553	51295	4917	35480	1268	27.981	26583	985
T-II	2415	151033	11763	93984	3553	26.452	69846	2778
T-III	1983	63499	5695	52202	4108	12.707	40887	1147
T-IV	2610	111494	9957	89794	10631	8.446	70223	1875
Total	3205	377321	22911	271460	19560	13.878	207539	4853

Appendix III: Average depth and width of phrases

phrase	zj	fj	dj	pp	vp	np	sp	ap	dp	tp	mp	qp
average depth	11	10	8	6	6	5	5	4	4	4	3	3
average width	28	29	13	5	7	4	4	2	2	3	2	2

Appendix IV: Distribution of preposition phrase

No.	root	left	phrase	right	Absolute freq.	Relative freq.	accumulative relative freq.
1	vp	##	pp	!vp	580	0.53065	0.53065
2	dj	##	pp	wco !dj	136	0.124428	0.655078
3	fj	##	pp	wco !fj	91	0.083257	0.738335
4	vp	##	pp	wco !vp	86	0.078683	0.817017
5	np	##	pp	ude1 !np	57	0.05215	0.869167
6	ap	##	pp	!ap	31	0.028362	0.89753
7	vp	!vp	pp	##	22	0.020128	0.917658
8	np	##	pp	ude1 !vp	17	0.015554	0.933211
9	dj	##	pp	!dj	9	0.008234	0.941446
10	dj	##	pp	wco !vp	7	0.006404	0.94785