Syntactic Positions of Prepositional Phrases in the History of Chinese: Using the Developing Sheffield Corpus of Chinese for Diachronic Linguistic Studies

Xiaoling Hu and Jamie McLaughlin
University of Sheffield, UK
Nigel Williamson
Sheffield Hallam University, UK

Abstract

This paper reports the completion of the first expansion phase of the Sheffield Corpus of Chinese (SCC). We describe the major improvements we made in expanding the corpus. They involve the coverage of time periods, choice of text types and categories, and selection of individual texts; the mark up scheme and the integral search and analysis tool. We use the developing SCC to examine Li and Thompson's (1974, 1975, 1976) controversial postverbal predominance hypothesis for prepositional phrases (PPs) in Archaic Chinese and their word order change hypothesis for PPs in general in the history of the Chinese language. Our study provides no evidence for the postverbal predominance hypothesis for PPs in Archaic Chinese and the word order change hypothesis for PPs in general from postverbal in Archaic Chinese to preverbal in Modern Chinese. Our findings show that postverbal and preverbal PPs have been in coexistence and there have always been more occurrences of preverbal PPs than postverbal PPs in all the time periods covered in the current SCC. Although use of some PPs declined in some time periods and use of others emerged in other time periods, there was never a predominant position for PPs in any time period in the history of Chinese. We show differences in the distribution of PPs in different time periods and provide an account of the syntactic positions of PPs in those time periods.

Correspondence:
Dr. Xiaoling Hu,
School of East Asian Studies,
University of Sheffield,
Floor 5, The Arts Tower,
Western Bank,
Sheffield S10 2TN, UK.
E-mail:
x.l.hu@sheffield.ac.uk

1 Introduction

The Sheffield Corpus of Chinese (SCC),[1] the outcome of a pilot project funded by the British Academy, was established in response to the lack of a publicly available corpus of marked-up Chinese texts for diachronic linguistic study. The first phase of the expansion was completed thanks to a Devolved Research Fund from the Arts and Humanities Research Council of the University of Sheffield. The objectives of this phase of the expansion were simply to add samples of text categories to each of seven time periods covered in SCC, and to test and refine the initial mark up system and the integrated search and analysis tool developed in the context of XML (eXtensible

Literary and Linguistic Computing, Vol. 22, No. 4, 2007.

doi:10.1093/llc/fqm017 Advance Access Published on 13 September 2007

Markup Language). At this stage, SCC contains 40 marked-up Chinese text samples and over 430,000 characters[2] with a mark up system and an integral search and analysis tool that can locate and display occurrences of any specified sequence of characters and of word classes, and can also count and display the frequencies by specified subcorpus, time period and genre. Having completed this expansion phase, we use SCC to examine Li and Thompson's (1974, 1976) controversial post-verbal prominence hypothesis of prepositional phrases (PPs) in Archaic Chinese and their word order change hypothesis of PPs in general in the history of the Chinese language. In Section 2, we describe the chronological framework and the progress in text sampling. Section 3 outlines the mark up system and the integral search and analysis tool. Section 4 tests and demonstrates SCC as a whole by examining the word order change hypothesis for PPs in the history of Chinese. We show that Li and Thomspon's (1974, 1976) post-verbal prominence hypothesis of PPs in AC and their word order change hypothesis for PPs in general in the history of Chinese are at best dubious. The syntactic positions of PPs in all the time periods in SCC are examined. In doing so, we show that even at this early stage of expansion SCC enables insightful investigation of linguistic development in Chinese.

2 Chronological Framework and Text Sampling

As described in Hu *et al.* (2005), we have based the major chronological framework of SCC [Archaic Chinese (AC, tenth century BC-AD 220), Medieval Chinese (MedC, 220–1368), and Modern Chinese (ModC, 1368–1911)[3]] on Peyraube (1996), for two reasons. One is that the division is based on syntactic criteria and the other is that it has taken into account the studies of many sinologists such as Wang (1958), Chou (1962), and Dobson (1959, 1962). Within this basic framework of three subcorpora Archaic, Medieval and Modern, SCC is further divided into seven[4] time periods based largely on dynasties as illustrated in Table 1.

The texts selected for the first expansion stage of SCC represent a wide range of kinds of writing found in different time periods and are organized in two major text types—literary and non-literary. Both types contain texts of different genres. The literary type contains works on poetry, drama, folklore and fiction (general, historical, romantic and mythological) and the non-literary type on philosophy, sciences, history, warfare, government, law, biographies and essays, religion and travelogues. In the current SCC, the contents distribution is: Archaic Chinese subcorpus 109,679 characters, Medieval Chinese subcorpus 147,492 characters and Modern Chinese subcorpus 175,522 characters. Sample text categories in SCC at the end of the first expansion phase are given in Table 2.

Table 2 shows the sixteen text categories currently represented in the corpus. A brief introduction to the content of text categories is given below.

The category for biographies and essays includes *Dunhuang Bianwen Ji* (A Collection of Dunhuang Popular Narratives) which represents a popular form of narrative literature flourishing in the Tang Dynasty with alternate prose and rhymed parts for recitation and singing, often on Buddhist themes (Sun, 1996).

In the fiction category, sample texts were selected from four subcategories: general, historical, romantic and mythological. *San Yan* (Three Words) by Feng Menglong (1574–1646), a compiler of anthologies of popular literature in the Ming Dynasty, contains collections of stories and is sampled in the general category. One chapter is selected from each of the three collections in Three Words: *Yu Shi Ming Yan* (Words to Instruct the World), *Jing Shi Tong Yan* (Words to Warn the World) and *Xing Shi Heng Yan* (Words to Awake the World). Historical fiction includes texts from *San Guo Yan Yi* (Romance of the Three Kingdoms) by Luo Guanzhong (1330–1400) which is the first historical novel written with chapters interwoven by the development of plots. We also included a sample chapter from *Shui Hu Zhuan* (Water Margin) by Shi Nai'an (1574–1645), one of the four most famous novels of

Table 1 Chronological framework of SCC

Sheffield Corpus of Chinese						
Archaic Chinese (Tenth BC to AD 220)		**Medieval Chinese (220–1368)**			**Modern Chinese (1368–1911)**	
Pre-Qin	**Han**	**Wei and Jin and Southern–Northern**	**Sui and Tang and Five Dynasties**	**Song and Yuan**	**Ming**	**Qing**
Tenth to 206 BC	206 BC to AD 220	220–581	581–979	960–1368	1368–1644	1644–1911

Table 2 Sample text categories in the current SCC

Code	**Text category**	**Total number of samples with breakdown into the seven time periods**	**Word count**	**Proportion (%)**
A	Biographies/Essays	3: 0, 1, 0, 1, 1, 0, 0	47,618	11.0
B	Drama/Play	2: 0, 0, 0, 0, 2, 0, 0	25,202	5.8
C	Fiction (general)	5: 0, 0, 0, 0, 0, 4, 1	82,296	19.0
D	Fiction (historical)	2: 0, 0, 0, 0, 0, 2, 0	30,114	7.0
E	Fiction (mythological)	3: 0, 0, 0, 0, 0, 1, 2	33,644	7.8
F	Fiction (romantic)	1: 0, 0, 0, 0, 0, 0, 1	12,506	2.9
G	Folklore	3: 0, 0, 0, 3, 0, 0, 0	10,988	2.5
H	Government	1: 0, 0, 0, 1, 0, 0 ,0	11,002	2.5
I	History	5: 1, 1, 3, 0, 0, 0, 0	68,236	15.8
J	Legal works	1: 1, 0, 0, 0, 0, 0, 0	10,384	2.4
K	Philosophy	6: 5, 0, 0, 0, 1, 0, 0	54,196	12.5
L	Poetry	3: 0, 0, 0, 2, 1, 0, 0	5,314	1.2
M	Religion	1: 0, 0, 1, 0, 0, 0, 0	6,515	1.5
N	Science/Technology	2: 0, 0, 0, 0, 1, 1, 0	18,664	4.3
O	Travelogue	1: 0, 0, 0, 0, 0, 1, 0	8,606	2.0
P	Warfare	1: 1, 0, 0, 0, 0, 0, 0	7,405	1.7
Total		40: 8, 2, 4, 7, 6, 9, 4	432,690	100.0

the Ming Dynasty. Water Margin is recorded in a colloquial style compounded with oral conventions and descriptive passages in prose narrative (Hanan, 1981). As such it is very useful for the historical study of the vernacular language during that time period. Romantic fiction includes sample chapters from *Hong Lou Meng* (sometimes translated as Dream of the Red Chamber) by Cao Xueqing (1715–63), one of the great masterpieces of Chinese fiction. Mythological fiction includes *Xi You Ji* (A Journey to the West) by Wu Cheng'en (1500–82), a novel that is regarded as representing the pinnacle of novels created in early Modern Chinese. Supernatural fiction such as *Liao Zhao Zhi Yi (Strange Tales of Liao Zhai)* by Pu Songlin (1640–1715) is an intermediate category and we decided to include it in the mythological fiction category.

In the government category, we included sample texts from *Zhen Guan Zheng Yao* (Administrative Principles of Zhenguan Reign) by Wu Jing (670–749). In the history category there are sample texts from *Shi Ji* (Records of the Grand Historian) by Sima Qian (about 145–90 BC), a Prefect of the Grand Scribes of the Han Dynasty and also known as the Father of Chinese history. This book not only records basic annals of dynasties or rulers, chronological tables and treatises among other things but also serves as a model for subsequent Chinese dynastic histories and is considered a representative text of late Archaic Chinese. *San Guo Zhi* (Chronicles of the Three Kingdoms) by Chen Shou (233–279), the official and authoritative historical text on the Three Kingdoms period of China, is representative of early Medieval Chinese texts. As Chronicles of the Three Kingdoms contains

three volumes giving smaller historical accounts of three rival states Wei, Shu and Wu, we selected one chapter from each of the three volumes.

In the category for legal works, the text sample is from *Shang Jun Shu* (Book of Lord Shang) by Shang Yang (390 to 338 BC). The Book of Lord Shang consists of a collection of the works written in the Legalist School represented by Shang Yang in the Warring States period (475 to 221 BC) and records the theory and the specific measures of the Shang Yang Reform led by Shang Yang in 361 BC.

In the philosophy category, sample chapters were selected from *Dao De Jing* (Classic of the Way and Its Power) by Lao Zi (around the sixth century BC) which is regarded as one of the core texts of the Chinese way of thinking known as Daoism. Sample texts were also taken from *Lun Yu* (The Analects) by Kong Zi (551 to 479 BC) and from *Meng Zi* (Mencius) by Meng Zi (372 to 289 BC). All these provide rich sources of prose texts of early Archaic Chinese.

The science and technology category includes *Meng Xi Bi Tan* (Notes Written at Mengxi) by Shen Kuo (1029–93) which is the first book written in China about science and technology and records scientific discoveries that the ancient Chinese had made in almost all sciences. There is also *Tian Gong Kai Wu* (Exploration of the Works of the Nature) by Song Yingxing (1587–1666) which is known as the first comprehensive book written in the world about agricultural and handicraft productions.

The travelogue category includes *Xu Xia Ke You Ji* (Travel Notes of Xu Xia Ke) by Xu Hongzu (1613–32) in the Ming Dynasty. In the warfare category there is *Sun Zi Bing Fa* (The Arts of War) by Sun Wu around the sixth century BC, the first book written in China about warfare.

As the summary above suggests, SCC mainly contains written textual materials from all the time periods in the history of the Chinese language. However two types of spoken-like data are included. One is a drama/play *Dou'e Yuan* (Dou's Case of Injustice) by Guan Hanqing (1271–1368), a famous playwright in the Yuan Dynasty. The other is the Medieval Chinese text *Zhu Zi Yu Lei* (Classified Quotations of Zhu Zi) by Zhu Xi (1130–1200), the most influential Chinese philosopher since the time of Confucius and Mencius. The text is characteristic of sermons and dialogues in the vernacular and represents Zhu Xi's actual speech as recorded by his disciples. Texts like these are generally regarded by scholars as a reflection of planned monologue study that represents, if not truly natural speech, some of the most 'spoken-like' registers (Halliday, 1991) available from earlier historical periods (Biber *et al.*, 1998). We believe that it is important to include such texts in SCC because they will provide researchers with useful comparative data for the analysis of written registers.

There are also a small number of translation texts in SCC such as the translation of a religious text *Jin Gang Jing* (The Diamond Sutra), a Buddhist scripture discovered in 1907 inside the Mogao Caves, from the Tang Dynasty (618–907). During the early Tang Dynasty the monk Xuan Zang went to Nalanda and other important sites to bring back scriptures. The Tang capital of Chang'an (today's Xi'an) became an important centre for Buddhist ideology. From there Buddhism spread to Korea and Japan. There is evidence in the text that Buddhist thought began to merge with Confucianism and Daoism, due in part to the use of existing Chinese philosophical terms in the translation of Buddhist scriptures. The Diamond Sutra was the first dated example of printed translation texts. Given that these translation texts were written by highly educated people and represented vernacular language used and spoken at the time, we believe that the use of these translations texts will not affect first language quality and that their inclusion in SCC is justified.

As we use natural chapters to sample historical texts (Hu *et al.*, 2005), there are differences in the lengths of text samples. For instance, a chapter from a biographical essay *Taiping Guang Ji* (Extensive Records of the Taiping Era) by Li Fang *et al.* from the Five Dynasties (907–979) contains 3370 characters whereas a chapter from a general fiction *Yu Shi Ming Yan* (Words to Instruct the World) by Feng Menglong contains 20,830 characters. SCC as established at the end of this expansion phase contains over 430,000 characters in forty text samples.

The texts for the first expansion stage of SCC were selected to cover most of the genres and

time periods. For this stage, easy availability of error-free texts in electronic form that are significant in the Chinese language as a whole was an important criterion. As the first-stage texts would be used to develop the annotation system we concentrated on classic texts that were significant in themselves in their time periods and had a lasting effect on subsequent writings. One example, the *Shu Jing* (Classic of History), a collection of documents and speeches alleged to have been written by rulers and officials of the early Zhou period and before, contains the best examples of early Chinese prose. The writings of Meng Zi (372 to 289 BC), along with others, contain extensive use of comparisons, anecdotes and allegories and developed a simpler and more concise prose style noted for its economy of words, which was effectively a template for literary form for the following 2000 years. Similarly the *Shi Ji* (Records of the Grand Historian) written by Sima Qian (between 145 to 90 BC) served as a model for historical texts for the following 2000 years. Another example, the *Dunhuang Bianwen Ji* (A Collection of Dunhuang Popular Narratives) represents a popular form of narrative literature flourishing in the Tang Dynasty (618–907) with alternate prose and rhymed parts for recitation and singing, often on Buddhist themes (Sun, 1996), and was crucially important for the development of fiction in Chinese literature as 'the predecessors of the later popular short stories' (Průšek, 1970, p. 240; also see Ma, 1976). The famous eighteenth-century romantic fiction *Hong Lou Meng* (Dream of the Red Chamber) established a lasting vernacular style and is widely regarded as a master work in Chinese literature. The crucial and complex question of balancing text samples in different genres and time periods will be a dominant aspect of the second expansion stage.

3 The Mark up System and the Integral Research and Analysis Tool

From the onset of the project, we have been aware of important issues associated with future international data exchange, especially in the creation of Asian language resources. We have therefore been careful to comply with the standards and practice set out in Xiao *et al.* (2004). As described in Hu *et al.* (2005), the mark up system for SCC is developed in the context of XML because XML is a well-supported open standard, ensuring compatibility and longevity for our data, and it also has an excellent support for non-Latin characters. We choose the UTF-8 character encoding standard for SCC, which accommodates Chinese characters and ensures compatibility between our corpus and all standards compliant XML processing tools.

A small number of samples were downloaded from the internet but most samples were collected freely from the Yifan Library, Guo Xue and China the Beautiful[5] which are willing to provide free access to their collections of texts written or dated before the beginning of the twentieth century. However, there were some problems that needed to be resolved before the texts available could be used. First, as nearly all the electronic texts we collected came in simplified Chinese characters, they were converted into traditional Chinese characters in order to be faithful to the original texts. Second, the texts contained errors in the first place and more errors occurred in the converting process. Together they gave rise to an error rate of about 2.5%. Most errors were due to the automatic blind conversion. Take the two characters 里[6] *li* and 裏 *li* as an example. Although both characters were frequently used in many historical Chinese texts, the conversion tool available only accepts 裏 as part of its traditional-form packaging. Every occurrence of 里 would be automatically converted into 裏 even when what fitted the context was the distance measure 里 'about half a kilometre' rather than 裏 'inside'. Errors like this were corrected manually. Each electronic text file in SCC was proof-read and corrected against the original texts (from various sources) independently by two native speakers of Chinese with a good knowledge of classical Chinese.

Another problem with historical texts is that many of the texts in text categories such as philosophy, government, history, religion and warfare have been published with notes and annotations written by different scholars or historians in mostly later time periods. For example, the text of the

Administrative Principles of Zhenguan Reign by Wu Jing (670–749) contained annotations written by Ge Zhi in the Yuan Dynasty (1279–1368). In order to keep the prose qualities of original texts, all the notes and annotations were removed from the sample texts selected to be included in SCC. The other problem we encountered in dealing with ancient Chinese texts is that a very small number of Chinese characters have become obsolete over time. They are so rarely used nowadays that they cannot be found in the CJK (Chinese/Japanese/Korean) Unified Ideographs. At present, they are represented as XML Internal Entities, which allows us to replace them with an editorial note, or even an image file we have created to depict the character.

The texts included in SCC range from tenth century BC to early twentieth century. This has made the tasks of word segmentation, part-of-speech (POS) tagging and the following post-editing of corpus text samples very difficult. McEnery *et al.* (2003), for example, point out that the essential and non-trivial process of word segmentation in Chinese corpus linguistics is a problem even in their context of homogenous contemporary Chinese texts. We are aware of segmentation tools such as the Chinese Lexical Analysis System (CLAS) developed at the Institute of Computing Technology, Chinese Academy of Sciences (Zhang *et al.*, 2002). That system is based on a core lexicon which incorporates a frequency dictionary of 80,000 words together with POS information and modules for word segmentation, POS tagging and unknown word recognition. However this system was developed using contemporary Chinese news texts and as McEnery *et al.* (2003) point out, it performs poorly on some genres, e.g. martial arts texts, and so even for their synchronic corpus the annotation of texts in all but few genres had to be manually corrected. We decided that for the SCC diachronic corpus, with its very wide range of genres and time periods, the CLAS system was inappropriate.

For SCC we decided to build a lexicon and an annotation system cumulatively from scratch using experience gained from the successive texts processed. The annotation process has an automatic stage (based on the corpus lexicon and grammatical rules) followed by manual editing. The automatic process for each new text uses the current lexicon containing all previously encountered lexical items and the set of interacting grammatical rules and specific examples. These elements are cumulatively increased and refined as texts are added to the corpus. Unresolved multiple tags and untagged new lexical items at the end of the automatic stage are resolved by manual post-editing. New lexical items (tokens) encountered, as texts are POS tagged, are added to the current lexicon. At the end of the first expansion stage, the lexicon has over 35,000 entries. The annotation system has nineteen basic word classes and 111 distinct tag labels (Hu *et al.*, 2005).

At this stage we have checked a few of the first texts to be tagged with the current lexicon and did not find inconsistencies in the tagging results. However a part of the next expansion stage will be systematic testing to ensure that the tagging system is consistent and reliable.

At the end of the first expansion phase, the integral search and analysis tool enables users to retrieve all occurrences of any specified sequence of characters or of any specified word classes or any combination of the two. Searches can be further restricted by subcorpus, time period and genre. When a search is executed, the web interface returns with a list of all occurrences within their immediate textual context and frequency tables displaying the proportions of occurrences in each subcorpus, time period and genre.

This functionality is presently implemented via a relational database which contains a record of every character, its position in the texts and the word classes to which it belongs. This database is generated automatically from our XML files, using SAXON.[7] All searches available on the current site are implemented via Standardised Query Language (SQL) and Java Server Pages (JSP). We chose JSP technology chiefly because of its entirely UTF-8 compliant architecture and ability to integrate with the various Java tools used in the creation of the corpus. Hence the aim is that researchers can use the in-house search and analysis tool that is an integral part of SCC. We believe that the improvement of the integral search and analysis tool at this stage was both important and necessary because it has made

it possible for researchers to carry out complex searching and analysis of SCC remotely, via a simple Web interface. We will of course consider the use of different retrieval and analysis tools as our project continues to expand, for instance, the XML Aware Indexing and Retrieval Architecture (XAIRA)[8] concordancing engine employed by, for example, LCMC (McEnery and Xiao, 2004; Xiao *et al.*, 2004).

4 Syntactic Positions of Prepositional Phrases in the History of the Chinese Language

In the early 1970s, Li and Thompson (1974, 1975) claimed that the Chinese language was changing from Subject-Verb-Object (SVO) to Subject-Object-Verb (SOV). A major piece of evidence they cited to support their claim was that PPs in AC were predominantly post-verbal and that pre-verbal PPs were a new development that emerged in MedC. As the first expansion stage of SCC neared completion we decided to use the developing corpus to test this assertion.

Li and Thompson's claim has met with adverse criticism (see, for example, He, 1984, 1985; Sun, 1991) and while the position of PPs is a significant factor in the study of language typology (Greenberg, 1963; Hawkins, 1983; Dryer, 1991; Liu, 2003) both Li and Thompson and their critics alike based their arguments on a very small number of texts. Their rather few examples include (1) below.

(1) $[_{VP}$ V PP] (AC) → $[_{VP}$ PP V] (ModC)

a. chu **yu** you gu (AC)
emerge from dark valley

b. **cong** you gu chulai (ModC)
from dark valley emerge

Both sentences contain an adjunct PP of location/source. It is post-verbal in AC but pre-verbal in ModC. However, not all the post-verbal PPs in AC can find pre-verbal counterparts in ModC. For instance, in the AC example (2), the PP of goal occurs in post-verbal position but its counterpart in ModC cannot occur in pre-verbal position as the ungrammaticality of (3b) shows.

(2) Li wang liu **yu** Zhi. (AC)
person-name emperor exile to place-name
'Emperor Li was exiled to the state of Zhi.'

(3) a. Li wang bei-liufang **dao** Zhi. (ModC)
person-name emperor PASSIVE-exile to place-name
'Emperor Li was exiled to the state of Zhi.'

b. *Li wang **dao** Zhi bei-liufang.
person-name emperor to place-name PASSIVE-exile
(PASSIVE = passive marker)

Examples 2–3 also show that even if there are some changes in syntactic positions of PPs from AC to ModC as illustrated in (1), it does not follow that all PPs in AC underwent similar changes in ModC. In his study of the distribution of PPs in two only (albeit large) texts from AC period, He (1984, 1985) observes that the obligatory occurrence of pre-verbal PPs amounts to 64% in one text and 80% in the other. He argues that in the two texts he studied, the *Zuo Zhuan* (Zuo's Commentary, 500 BC) and *Shi Ji* (Records of the Grand Historian, 100 BC), pre-verbal position is the obligatory position for the absolute majority of PPs in both texts. Sun (1991) observed that pre-verbal and post-verbal PPs are more or less equally distributed in AC period: 50% in one chapter of a text and 45% in one chapter of the other text and subsequently Sun (1996) noted that while most PPs had at least one occurrence in pre-verbal position, only half of them had occurrences in post-verbal position. Sun's investigations, like those of He and Li and Thompson, were based on a very limited number of texts. Sun's counts of the distribution of PPs were in one chapter from Zuo's Commentary and one chapter from *Meng Zi* (Mencius, 300 BC): the former contains 7,242 Chinese characters and the latter 3,115.

In order to provide a more thorough examination of the syntactic positions of PPs in all the time periods covered in SCC, we first counted the total number of the different PPs occurring in the corpus with the POS tag for preposition and we found twenty nine[9]. We then counted the distribution of

Table 3 Distributions of the 29 PPs in the current SCC (actual occurrences followed by frequencies per 10,000 words)

Sheffield Corpus of Chinese												
PPs	**AC (109,679)**				**MedC (147,490)**				**MC (175,521)**			
	Post-verbal		**Pre-verbal**		**Post-verbal**		**Pre-verbal**		**Post-verbal**		**Pre-verbal**	
以 yi	199	18.1	1019	92.9	152	10.3	842	57.1	32	1.8	170	9.7
於 yu2	568	51.8	87	7.9	483	32.7	151	10.2	179	10.2	43	2.4
于 yu1	374	34.1	20	1.8	94	6.4	2	0.1	18	1.0	0	0.0
為 wei	187	17.0	194	17.7	278	18.8	250	17.0	222	12.6	119	6.8
與 yu	16	1.5	200	18.2	45	3.1	229	15.5	111	6.3	362	20.6
自 zi	3	0.3	46	4.2	1	0.1	137	9.3	2	0.1	92	5.2
如 ru	40	3.6	39	3.6	144	9.8	45	3.1	76	4.3	85	4.8
由 you	5	0.5	41	3.7	0	0.0	21	1.4	0	0.0	6	0.3
在 zai	13	1.2	18	1.6	97	6.6	85	5.8	244	13.9	319	18.2
及 ji	12	1.1	15	1.4	19	1.3	5	0.3	8	0.5	0	0.0
至 zhi	11	1.0	22	2.0	57	3.9	50	3.4	61	3.5	21	1.2
像 xiang2	0	0.0	6	0.5	0	0.0	10	0.7	0	0.0	14	0.8
從 cong	0	0.0	14	1.3	2	0.1	125	8.5	1	0.1	118	6.7
比 bi	0	0.0	3	0.3	0	0.0	4	0.3	0	0.0	24	1.4
將 jiang	0	0.0	0	0.0	0	0.0	109	7.4	0	0.0	179	10.2
被 bei	0	0.0	0	0.0	0	0.0	64	4.3	0	0.0	86	4.9
向 xiang1	0	0.0	0	0.0	0	0.0	41	2.8	0	0.0	65	3.7
到 dao	0	0.0	0	0.0	32	2.2	5	0.3	166	9.5	97	5.5
把 ba	0	0.0	0	0.0	0	0.0	36	2.4	0	0.0	125	7.1
似 si	0	0.0	0	0.0	5	0.3	14	0.9	17	1.0	22	1.3
用 yong	0	0.0	6	0.5	0	0.0	16	1.1	0	0.0	51	2.9
對 dui	0	0.0	0	0.0	0	0.0	14	0.9	0	0.0	57	3.2
同 tong	0	0.0	0	0.0	0	0.0	14	0.9	0	0.0	54	3.1
往 wang	0	0.0	0	0.0	3	0.2	8	0.5	2	0.1	34	1.9
替 ti	0	0.0	0	0.0	1	0.1	6	0.4	0	0.0	32	1.8
跟 gen	0	0.0	0	0.0	0	0.0	5	0.3	0	0.0	9	0.5
拿 gei	0	0.0	0	0.0	4	0.3	0	0.0	7	0.4	3	0.2
據 na	0	0.0	0	0.0	0	0.0	2	0.1	0	0.0	26	1.5
據 ju	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	13	0.7
Total	1428	130.2	1730	157.7	1417	96.1	2290	155.3	1146	65.3	2226	126.8

all these PPs and the distribution of post-verbal and pre-verbal PPs in seven time periods in the three subcorpora of SCC. The distribution of these PPs (in the current SCC) broken down into post- and pre-verbal positions is given in Table 3 as occurrences followed by frequencies per 10,000 words. Log-likelihood (LL) tests on the occurrences of PPs in the two positions show that the differences between the distributions of PPs are statistically significant in all the time period covered in the current SCC (Table 4)[10] and, for example, that the changes in the proportions of post- and pre-verbal PPs between the three subcorpora are statistically significant (Table 5).

Fig. 1 and subsequent similar figures are produced using the data of Table 3 further broken down into the seven component time periods of the three subcorpora. Fig. 1 shows that post-verbal and pre-verbal PPs have coexisted in all the time periods and pre-verbal PPs have always been used more frequently than post-verbal PPs. A 2×7 Chi-square test on the post- and pre-verbal occurrences of PPs in the seven time periods gave the calculated value of 105.8 at the significance level of $P < 0.001$. Note that although there are twice as many occurrences of pre-verbal PPs as post-verbal PPs in ModC, it does not follow that post-verbal PPs will be entirely replaced by pre-verbal PPs because post-verbal PPs

Table 4 Distribution of post- and pre-verbal PPs in each of the seven time periods in the current SCC

Subcorpus	Time periods	Words	PP type	Frequency	LL score	Sig. level
AC	Tenth to 206 BC	72,133	Post-verbal	770	37.69	<0.001
			Pre-verbal	1030		
	206 BC to AD 220	37,546	Post-verbal	658	1.36	<0.05
			Pre-verbal	700		
MedC	220–581	42,448	Post-verbal	418	32.71	<0.001
			Pre-verbal	600		
	581–979	40,737	Post-verbal	558	76.39	<0.001
			Pre-verbal	889		
	960–1368	64,307	Post-verbal	441	105.86	<0.001
			Pre-verbal	801		
ModC	1368–1644	130,239	Post-verbal	841	226.03	<0.001
			Pre-verbal	1574		
	1644–1911	45,282	Post-verbal	305	128.73	<0.001
			Pre-verbal	652		

Table 5 Comparisons of the distribution of PPs in the two positions between the three subcorpora

PP type	AC versus MedC		AC versus ModC		MedC versus ModC	
	LL score	Sig. level	LL score	Sig. level	LL score	Sig. level
Post-verbal	65.48	$P < 0.001$	304.53	$P < 0.001$	95.24	$P < 0.001$
Pre-verbal	0.26	$P < 0.05$	46.06	$P < 0.001$	46.18	$P < 0.001$

and pre-verbal PPs have different syntactic functions.

Table 3 and Fig. 1 show that PPs occur with substantial frequencies in both post-verbal and pre-verbal positions in all time periods. Post-verbal PPs and pre-verbal PPs are almost evenly distributed in AC: 42.8% versus 57.2% in early AC and 48.5% versus 51.5% in late AC (also see Table 4). Our findings strongly refute Li and Thompson's post-verbal predominance hypothesis for PPs in AC. We can also see that while post-verbal and pre-verbal PPs have been in coexistence in all time periods in the history of Chinese, there have always been more occurrences of pre-verbal PPs than post-verbal PPs. Our investigation provides strong evidence against Li and Thompson's claim that pre-verbal PPs did not become prevalent until the fifteenth or sixteenth centuries, which renders untenable their claim that the emergence of pre-verbal PPs was a new development. Moreover, it can also be seen that there has never been a historical time period in which one of the positions was strongly predominant as suggested by Sun (1991, 1996).

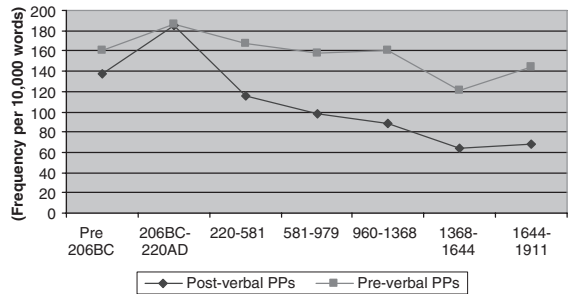

Fig. 1 Distribution of the 29 PPs in all time periods in the current SCC

However, when individual prepositions are examined, we see that the overall trend indicated in Fig. 1 is made up of many disparate trends which we now discuss in more detail.

4.1 Pre-verbal PPs

A closer examination of the PPs occurring in pre-verbal position in AC reveals that the majority of them occurred with 以 *yi*, the most frequently used

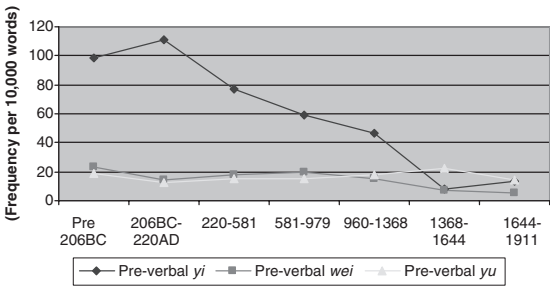

Fig. 2 Distribution of pre-verbal *yi*, *wei* and *yu* in the current SCC

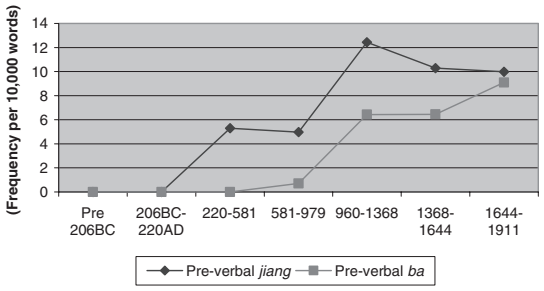

Fig. 3 Distribution of *jiang* and *ba* in the current SCC

preposition, and the other two commonly used pre-verbal PPs: 與 *yu* and 爲 *wei* (Table 3). Although all these three prepositions occurred in both post-verbal and pre-verbal positions in AC, a large proportion of them are found in pre-verbal position: 81.9% of *yi*, 91.2% of *yu* and 54.9% of *wei*. Together they make up 83.9% of all the occurrence of PPs in pre-verbal position in AC. *Yi* is mainly used in marking direct object of a double-object construction, the theme object of a spatial expression, a nominal functioning as an instrument and an attitudinal construction in *yi*...*wei* form (Zhu, 1957; Bennett, 1981; Peyraube, 1986; Sun, 1996). *Yu* is used to mark comitative and *wei* to mark object and passive. Distributions of pre-verbal *yi*, *yu* and *wei* are given in Fig. 2.

Fig. 1 shows that pre-verbal PPs increased from early AC to late AC (LL score 29.40, $P < 0.001$). When we examined further the distribution of pre-verbal PPs in these two periods in AC, we found that the use of pre-verbal *yi* increased from 88.0 (frequency per 10,000 words) in early AC to 102.3 in late AC, which is probably the main reason for the increase of pre-verbal PPs in these two time periods (Fig. 2).

The general occurrences of pre-verbal PPs remained stable in MedC. Fig. 2 indicates that the general decline of use of pre-verbal PPs was mostly related to *yi*'s decline (LL score 5.31, $P < 0.01$). Although the use of *wei* and *yu* remained fairly stable throughout the AC and MedC, the use of *wei* decreased slightly from the beginning of ModC (1368–1644), which we will return to shortly. Examination of the declining use of *yi* reveals that two competing instrument and object markers 將 *jiang* and 把 *ba* were emerging in early MedC. The distributions of *jiang* and *ba* are given in Fig. 3.

Fig. 3 shows that *jiang* started to emerge from the beginning of MedC (220–581), earlier than *ba*, which did not appear until middle MedC (581–979). This refutes the general assumption (Wang, 1958; Li and Thompson, 1974; Peyraube, 1989b, 1996; Sun, 1996) that *jiang* and *ba* began to be used to mark objects between 700–1000. Analysis of individual sentences marked by *jiang* at the beginning of MedC shows that the preposition was already associated with object marking. Of the thirty-three occurrences of *jiang* used as a pre-verbal preposition found in the period, twenty-eight (85%) showed its use as an object marker and five (15%) as an instrument marker. In late MedC (960–79), *jiang*'s use as a pre-verbal preposition increased slightly (LL score 2.12,[11] $P < 0.05$). From the sentences that we extracted that contained *jiang* in this period, we found that its usage also involved marking attitude and marking a theme in a spatial expression (Sun, 1996). However there were still many occurrences of pre-verbal *yi* near to the end of MedC. Of the 801 pre-verbal PPs, 235 (29.3%) were associated with *yi*.

Fig. 3 also shows that *ba* did not start to emerge until middle MedC. From the sentences that contained *ba* in middle MedC, we observed that when *ba* first appeared, it was largely used as an instrument marker. The use of *ba* as an object marker was developed in late MedC (960–1368). Of the thiry-two sentences that contained *ba* in this period, twenty-seven (87.5%) showed *ba* as an

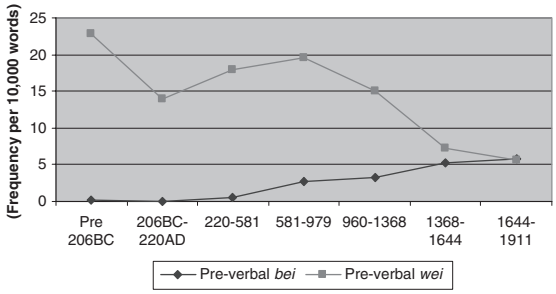

Fig. 4 Distribution of pre-verbal *bei* and *wei* in the current SCC

object marker. Following Zhu (1957), Sun (1996) argues that in marking object, the *ba*-form is a replacement of a similar construction marked by *yi* in early AC (Ye, 1988; Mei, 1990). In other words, *ba* is assumed to have come from *yi* by a simple lexical replacement. This lexical replacement hypothesis is not confirmed in our investigation. It is clear from Figs 2 and 3 that *yi* started to decline long before *ba* began to be used. By the time *ba* emerged, some use of *yi* as an object and instrument marker had already been shared by *jiang*. It is therefore implausible for the *yi*-form to serve as an analogical model for the *ba*-form as suggested by Peyraube (1988).

By late MedC, *jiang* and *ba* appeared to have replaced *yi* in marking object, instrument and attitude though in the meantime their use in marking instrument and attitude started to decline as a result of emergence of other pre-verbal prepositions such as 跟 *gen* from late MedC and 用 *yong* and 拿 *na* from early ModC. *Gen, yong* and *na* are used to mark instrument and attitude in pre-verbal position in present-day Chinese.

The decline of pre-verbal PPs in late MedC also involves the use of 爲 *wei* as a pre-verbal preposition marking passive. *Wei* marks passive construction in two forms: 爲-V *wei-V* and 爲. . .所[12] -V *wei . . . suo-V*. The difference between *wei-V* and *wei . . . suo-V* is that the former suppresses the agent whereas the latter introduces the agent immediately after *wei*. Fig. 2 shows that the use of *wei* started to decline from late MedC (LL score 9.35, $P < 0.01$). In early MedC, the proportion of *wei*'s use as a passive marker was very high. Of a total of sixty-five occurrences of pre-verbal *wei* found in early MedC, forty-one (63.1%) were related to passive marking. Among these forty-one passive constructions, eight (19.5%) involved no agent. However, in late MedC the use of *wei* as a passive marker decreased. This seems to be related to the emergence of another passive marker 被 *bei* that emerged from early MedC. Like *wei*, *bei* marks passive construction in *bei. . . (suo)V* and *bei-V* forms in pre-verbal position and can also be used to suppress the agent in a passive construction. The distribution of *bei* and *wei* is given in Fig. 4.

Analysis of occurrences of *bei* and *wei* shows that although *bei* started to be used as a competing passive marker in pre-verbal position from early MedC, both *wei* and *bei* continued to be found in marking passive construction for a long time during MedC. It was not until early ModC that *wei*'s use started to decrease again (LL score 1.49, $P < 0.05$), coinciding with *bei* becoming dominant in passive marking. It is natural to ask why the use of *wei* declined when it was just as competitive and flexible as *bei*. A plausible answer is that it was due to the other pre-verbal usages of *wei* which would likely cause ambiguity (Sun, 1996). For instance, pre-verbal *wei* was also used to mark benefactive or purpose. At the end of MedC, the ratio of passive marking between *wei* and *bei* was 64% to 35.5%. In late ModC, *wei* was largely replaced by *bei* as 86.7% of passive constructions was marked by *bei*.

4.2 Post-verbal PPs

From Fig. 1 we can see that the use of post-verbal PPs started to decline from the beginning of MedC (LL score 26.32, $P < 0.001$). A closer examination shows that most of the post-verbal PPs occurred with three other most frequently used PPs in this time period: 于 *yu1*, 於 *yu2* and 爲 *wei*, which together made up 45.3% of all the occurrence of PPs in AC. *Yu1* is generally regarded as an older form of *yu2* (Wang, 1958). All these three prepositions occurred in pre-verbal and post-verbal positions but 79% of their occurrences are found in post-verbal position. The distributions of these three post-verbal PPs in AC are given in Fig. 5.

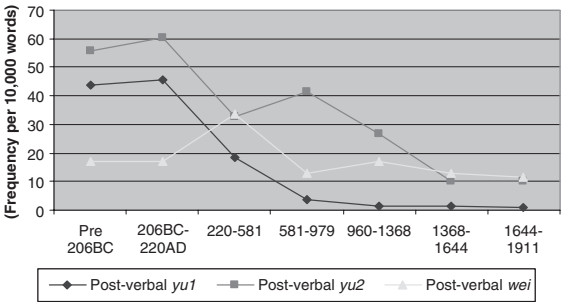

Fig. 5 Distribution of post-verbal *yu1*, *yu2* and *wei* in the current SCC

It is worth noting that these three prepositions all have multiple usages. Both *yu1* and *yu2* could be used in marking object, location, instrument, benefactive and passive construction, except for comparative construction where *yu2* tended to be used (Sun, 1996). Although both *yu1* and *yu2* have survived into present-day Chinese, their current use is limited to set phrases. *Wei* can be used in pre-verbal position to mark benefactive, purpose and passive construction. From Fig. 5, we can see that the use of post-verbal *wei* has been quite stable in most time periods except in early MedC where there was a noticeable increase (LL score 5.30, $P < 0.05$). Examination of sentences that contained post-verbal *wei* in this period shows that *wei* seemed to be involved in marking the second object of the main verb in a double-object construction where the main verb is subcategorised for two complements as shown in (4).

(4) 遂立子肯爲㠯大相。(*Dunhuan Bianwen: Wu Zixu Bianwen*)

sui li person-name *wei* guo daxiang
then establish person-name *wei* state prime-minister
'then made Zixu the prime minister of the state'

We suspect this phenomenon is related to the grammaticalization process of some transitive verbs in the history of the language. However this requires more detailed analysis which is inappropriate here.

Like the pre-verbal PPs we examined earlier, the occurrences of post-verbal PPs in AC also declined over time (LL score 36.76, $P < 0.001$). As *yu1*, *yu2* and *wei* are the most frequently used post-verbal PPs in AC, the general decline in the use of post-verbal PPs is inevitably related to their decline (overall) in their use. Fig. 5 shows that although the occurrences of both *yu1* and *yu2* started to decline from the beginning of MedC, *yu1* disappeared more quickly than *yu2* from middle MedC. As these two prepositions are mostly interchangeable in their uses except for comparative construction where *yu2* was preferred (Sun, 1996) as mentioned earlier, we refer to them as *YU* in the rest of our discussion.

One of the multiple uses of *YU* is to mark passive construction in post-verbal position like 'by' in a passive construction in English. The use of *YU* marking the passive started to decline from the beginning of MedC, which seems to coincide with the growth of *wei* as a passive marker in pre-verbal position during the same time period. Although *YU* was a commonly used passive marker in post-verbal position, it was unable to suppress an overt agent because Chinese generally does not allow preposition stranding.[13] On the other hand, *wei* marks passive construction in two forms: *wei-V* and *wei . . . suo-V*, which means that with pre-verbal *wei* there is an option of suppressing the agent in a passive construction. Of the thirty-six passive constructions marked by *wei* in early MedC (220-581), we found twenty-eight (77.8%) without agent. This flexibility made *wei* a very competitive passive marker for *YU* at the beginning of MedC. By middle MedC the proportion of *YU* in marking passive construction is reduced to only 0.7%. Our findings are in line with Peyraube (1996) that the lack of capability to suppress agent in a passive construction marked by *YU* led to its decline in marking passive construction.

However, passive PPs marked by *YU* are not the first post-verbal PPs that disappeared as claimed in Sun (1996) because other noticeable changes were observed in the meantime. One noticeable change is that *YU*'s use as a post-verbal source/location marker was gradually replaced by that of pre-verbal 在 *zai*[14] 'at' from early MedC. The distribution of *zai* is given in Fig. 6.

Zai is a preposition that has been used in both pre-verbal and post-verbal positions in all the time

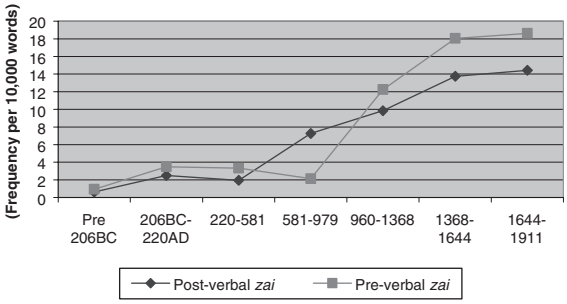

Fig. 6 Distribution of *zai* in all time periods in the current SCC

periods in the history of the Chinese language. However there is a difference in meaning depending on the position *zai* occupies in a sentence. Post-verbal *zai* marks direction whereas pre-verbal *zai* marks location (Li, 1975; Li and Thompson, 1975; Travis, 1984). Fig. 6 shows that the prepositional use of *zai* started from very early on in AC, albeit rarely. Fig. 6 also shows that the proportion of *zai*'s post-verbal use increased noticeably from seven (frequency value at per 10,000 words) in early MedC (581–979) to 179 at the beginning of ModC (1368–1644) (LL score 60.90, $P < 0.001$) and the proportion of *zai*'s pre-verbal use increased very quickly from twelve (frequency value per 10,000 words) in middle MedC to 235 at the beginning of ModC (LL score 70.27, $P < 0.001$).

Zai is not the only preposition that was replacing *YU* in marking locative constructions. There is also 到 *dao* 'to', which emerged in middle MedC to denote goal in post-verbal position. From late MedC *dao* was found to also mark location in pre-verbal position as illustrated in (5).

(5) 你爺兒兩個隨我到家中去來。
(*Dou'e Yuan*)

ni yer liangge sui wo *dao* jia zhong qu lai
you sir two follow me to home inside go come
'You two gentleman follow me and I'll take you to my home.'

By the end of ModC period, 74.6% of *dao*'s occurrence is post-verbal, whereas 25.4% pre-verbal. 向 *xiang1* 'toward' is another preposition that emerged at the beginning of MedC. Together with 自 *zi* 'since/from' and 从 *cong* 'from' that originated in early AC but developed rapidly from the beginning of MedC, *xiang1* has since been used in marking location/source in pre-verbal position as shown in (6) apart from goal for which only post-verbal *xiang1* tends to be used.

(6) 這的是衙門從古向南開,
(*Dou'e Yuan*)

zhede shi yamen cong gu *xiang1* nan kai
this is government-office from ancient towards south open
'It is true that ever since the ancient time *yamen* have been open to the south[15].'

From middle MedC, a few more new prepositions emerged in association with three of the many uses of post-verbal *YU*: the marking of comitative, dative/benefactive and instrument. There was 對 *dui* 'to', that emerged in middle MedC, and 同 *tong* 'with' and 跟 *gen* 'with' that appeared in late MedC. They are used to mark comitative in pre-verbal position. There was 替 *ti* 'for', that emerged near the end of MedC and 给 *gei* 'for', that appeared in late ModC. Both of them are used in marking dative/benefactive in pre-verbal position though *gei* sometimes also occurs in post-verbal position. 用 *yong* 'with' started to be used to mark instrument from middle MedC whereas 拿*na* 'with' was barely used until early ModC.

The use of *YU* in marking comparative and simile constructions was taken over by 似 *si* 'similar to' and 如 *ru* 'like' that emerged and developed in middle MedC and 同 *tong* 'with' and 比 *bi* 'compared to' that started to emerge from late MedC. *Ru* and *si* were found in marking similitude in both pre-verbal and post-verbal positions and in marking comparative construction in post-verbal construction. *Tong* and *bi* were found in marking similitude construction (Zheng, 1985) in pre-verbal position.

It seems that all these occurrences exerted a specific effect on the decline in the use of post-verbal *YU* and hence on the decline of general use of post-verbal PPs from early MedC. The evidence of our study supports Sun's (1996) suggestion that the decline in the use of post-verbal *YU* can be attributed to its multiple functions that are likely to cause ambiguity in communication.

5 Conclusion

On the basis of the current SCC, Li and Thompson's (1974, 1976) post-verbal predominance hypothesis for PPs in AC is refuted. Our study provides no evidence for their word order change hypothesis for PPs from post-verbal in AC to pre-verbal in ModC. Contrary to Li and Thompson's (1976) claim, there is clear evidence that pre-verbal and post-verbal PPs have coexisted in all the time periods in the history of the Chinese language, and that pre-verbal PPs have always been used more frequently than post-verbal PPs. Our study also provides strong evidence against Li and Thompson's claim that pre-verbal PPs did not become prevalent until the fifteenth to sixteenth centuries, which renders untenable their claim that the emergence of these pre-verbal PPs was a new development. However, it is clear that the proportions of individual PPs in the two positions change over time: uses of some post-verbal PPs disappear whereas uses of other pre-verbal PPs emerge. Proportions of uses of both positions also change over time: marking of location/source that used to be denoted by post-verbal PPs is now taken over by PPs in pre-verbal position though marking of other uses such as goal has always been denoted by PPs in post-verbal position. The decline in the use of some of the most frequently used PPs in AC, whether in post-verbal or pre-verbal position, is consistent with the suggestion that the ambiguity that is caused by the multiple functions that these PPs had could be the main cause for their decline because the emerging PPs that replace them have more specific functions. Although there are over twice as many occurrences of PPs in pre-verbal position as in post-verbal position in ModC, it does not necessarily follow that post-verbal PPs will be replaced by pre-verbal PPs. Different syntactic positions entail differences in meaning, which determines that the different functionality of post-verbal PPs cannot be replaced by PPs in pre-verbal position and vice versa.

This article reports the completion of the first expansion phase of the SCC. We described some major developments of SCC in terms of text sampling, the mark up system and the integral search and analysis tool. A diachronic linguistic study carried out on the word order change of PPs in the history of the Chinese language using the developing SCC shows the usefulness of the corpus under construction.

Acknowledgements

We are pleased to acknowledge the constructive and insightful comments on the first version of this paper by the referees of *Literary and Linguistic Computing*.

References

Bennett, P. (1981). The evolution of passive and disposal sentences. *Journal of Chinese Linguistics*, **9**: 61–90.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investing Language Structure and Use*. Cambridge: Cambridge University Press.

Chou, F. (1962). Zhongguo gudai yufa: gouci bian. Academia Sinica, Institute of History of Philology, Monograph No. 39.

Dobson, W. A. C. H. (1959). *Late Archaic Chinese*. Toronto: University of Toronto Press.

Dobson, W. A. C. H. (1962). *Early Archaic Chinese*. Toronto: University of Toronto Press.

Dryer, M. S. (1991). SVO languages and the OV:VO typology. *Journal of Linguistics*, **27**: 443–82.

Greenberg, J. H. (1963). Some universals of language with special reference to the order of meaningful elements. In Greenberg, J. H. (ed.), *Universals of Language*. Cambridge, MA: MIT Press, pp. 73–113.

Halliday, M. (1991). Corpus studies and probabilistic grammar. In Aijmer, K. and Altenberg, B. (eds), *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman, pp.30–43.

Hanan, P. (1981). *The Chinese Vernacular Story*. Cambridge, MA.: Harvard University Press.

Hawkins, J. A. (1983). *Word Order Universals*. New York: Academic Press.

He, L. (1984). Zuozhuan, Shiji jiebing duanyu weizhi de bijiao (A comparison between the positions of the prepositional phrases in Zuozhuan and Shiji). *Yuyan Yanjiu*, **1**: 57–65.

He, L. (1985). Shiji yufa tedian yanjiu (A study of the grammar in Shiji). In Cheng, X. (ed.), *Lianghan hanyu yufa yanjiu (Studies in the grammar of Han Chinese)*. Jinan (P.R.C.): Shangdong jiaoyu chubanshe, pp. 1–261.

Hu, X., Williamson, N., and McLaughlin, J. (2005). Sheffield Corpus of Chinese for diachronic linguistic study. *Literary and Linguistic Computing*, **20**(3): 281–93.

Li, C. (1975). Synchrony vs. diachrony in language structure. *Language* **51**: 873–76.

Li, C. and Thompson, S. (1974a). Co-verbs in Mandarin Chinese: verbs or prepositions? *Journal of Chinese Linguistics*, **2**: 257–78.

Li, C. and Thompson, S. (1974b). An explanation of word order change SVO<SOV. *Foundations of Language*, **12**(2): 201–14.

Li, C. and Thompson, S. (1975). The semantic function of word order: a case study in Mandarin. In C. Li (ed), *Word Order and Word Order Change*. Austin: University of Texas Press, pp. 163–95.

Li, C. and Thompson, S. (1976). Development of the causative in Mandarin Chinese: interaction of diachronic processes in syntax. In Shibatani, Masayoshi (ed.), *Syntax and Semantics*. New York: Academic Press, pp. 477–91.

Liu, D. (2003). *Word Order Typology and Theory of Prepositions*. Beijing: Commercial Press.

Ma, Y. (1976). The beginning of professional storytelling in China: a critique of current theories and evidence. *Paris: Bibliothèque de l'Institu des Hautes Êtudes Chinoises*. Vol. XXIV. pp. 227–45.

McEnery, A. and Xiao, Z. (2004a). Character encoding in corpus construction. In Wynne, M. (ed.), *Guide to Good Practice*. Oxford: Oxford University Press.

McEnery, A. and Xiao, Z. (2004b). The Lancaster Corpus of Mandarin Chinese: a Corpus for monolingual and contrastive language study. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC) 2004, pp. 1175–78. Lisbon, May 24–30, 2004.

McEnery, A., Xiao, Z., and Mo, L. (2003). Aspect marking in English and Chinese: using the Lancaster Corpus of Mandarin Chinese for contrastive language study. *Literary and Linguistic Computing*, **18**: 361–78.

Mei, T. (1990). Tang song chuzhishi de laiyuan (The origin of the disposal construction in the Tang and Song dynasties). *Zhongguo Yuwen*, **216**: 191–206.

Peyraube, A. (1986). Shuangbin jiegou cong handai zhi tangdaide lishi fazhan (The historical developments of double object constructions from Han Dynasty to Tang Dynasty). *Zhongguo Yuwen*, **3**: 204–16.

Peyraube, A. (1988). *Syntaxe diachronique du Chinois. Evolution des constructions dative du 14e av.J.-C siècle au 18e siècle*. Paris: Collège de France.

Peyraube, A. (1989a). History of the passive construction in Chinese until the 10th century. *Journal of Chinese Linguistics*, **17**(2): 335–71.

Peyraube, A. (1989b). Zaoqi BA ziju de jige wenti (Several questions on the early BA construction). *Yuwen Yanjiu*, **1**: 1–19.

Peyraube, A. (1996). Recent issues in Chinese historical syntax. In Huang, C.-T.J. and Audrey, L.Y.-H. (eds.), *New Horizons in Chinese Linguistics. Studies in Natural Language and Linguistics Theory 35*. London, Dordrecht and Boston: Kluwer.

Průšek, J. (1970). *Chinese History and Literature*. London, Dordrecht and Boston: Kluwer.

Sun, C. (1991). The adposition YI and word order in Classical Chinese. *Journal of Chinese Linguistics*, **19**(2): 202–18.

Sun, C. (1996). *Word-order Change and Grammaticalization in the History of Chinese*. Stanford, CA: Stanford University Press.

Travis, L. (1984). *Parameters and Effects of Word Order Variation*. PhD dissertation, MIT.

Wang, L. (1958). *Hanyu shigao (A draft history of Chinese grammar)*. Beijing: Kexu Chubanshe.

Wu, Z. and Tseng, G. (1993). Chinese text segmentation for text retrieval: achievements and problems. *Journal of the American Society for Information Science*, **44**: 532–42.

Xiao, Z. and McEnery, A. (2004). A corpus-based two-level model of situation aspect. *Journal of Linguistics*, **40**: 325–63.

Xiao, Z., McEnery, A., Baker, P., and Hardie, A. (2004). Developing Asian language corpora: standards and practice. *Proceedings of the 4th Workshop on Asian Language Resources*, pp. 1–8. March 25, 2004, Sanya, China.

Ye, Y. (1988). Sui Tang chuzhishi neizai yuanyuan fenxi. *Zhongguo Yuyan Xuebao*, **16:** 55–71.

Zhang, H., Liu, Q., Zhang, H., and Cheng, X. (2002). Automatic recognition of Chinese unknown words based on role tagging. In Tsou, B., Kwong, O. and Lai, T. (eds), *Proceedings of the 1st SIGHAN Workshop, COLING 2002*. Taipei, Academia Sinica, pp. 71–7.

Zheng, Y. (1985). Bijiaoju zhong 'ru', 'si', 'bi' de wenti (Some problems of the words 'ru' 'si' 'bi' in the

comparative construction). *Youshi Xuezhi (Taiwan)*, **18**(4): 143–59.

Zhu, M. (1957). Zaoqi chuzhishi (On the early disposal form). *Yuyanxue Luncong*, **1**: 17–33.

Notes

1 See www.hironline.ac.uk/scc/

2 In this article, we distinguish between 'characters' (individual Chinese characters) and 'words' (tokens, or individually tagged text fragments) (also see Wu and Tseng, 1993; McEnery *et al.*, 2003; Hu *et al.*, 2005).

3 Peyraube's Modern Chinese period has been expanded to 1911 because it was the year when the last dynasty in that period ended.

4 At the end of the pilot project there were eleven projected subdivisions but for the first expansion phase we decided to reduce them to seven to avoid extreme differences in the size of their contents. Western-Han and Eastern-Han were combined to form a subdivision Han (206 BC to AD 220), Sui, Tang and Five Dynasties to form Sui and Tang and Five Dynasties (581–979), and Song and Yuan to form Song and Yuan (960–1368).

5 We thank www.yifan.net, www.guoxue.com and http://www.chinapage.com/china.html for providing us with free access to their collections of classical Chinese texts written or dated before the beginning of the twentieth century.

6 In written Chinese used in areas and regions including mainland China where simplified Chinese is the official form of writing, 里is the simplified version of 裏.

7 See http://www.saxonica.com

8 See www.oucs.ox.ac.uk/rts/xaira/

9 Prepositions with a total occurrence ≤ 10 are excluded in this investigation.

10 The critical LL score for $P < 0.001$ is 10.83.

11 For one degree of freedom, a calculated value of 3.84 or higher is significant at the level $P < 0.05$.

12 所suo is regarded as a clitic prefixed to a transitive verb followed by an overt object (Wang, 1958; Peyraube, 1989a; Sun, 1996).

13 Preposition stranding refers to the phenomenon in which a preposition is not followed by an overt noun phrase. Some languages such as English allow preposition stranding. For example, in the sentence 'This is the North Face I told you about' the preposition 'about' is not followed by a noun, but the sentence is grammatical.

14 Apart from being a preposition, 在*zai* (and 正在 *zheng zai)* can also be used as a progressive aspect marker before action verbs to indicate an ongoing action as shown in (1). This use of *zai* was excluded from the frequency data in Fig. 6.

(1) 只見諸王貴人正在堂上飲宴。 (Er Ke Pai An Jing Qi: 2)
just see all royal distinguished person just PROGRESSIVE hall on eat feast
'(He) just saw that all the royal members were having dinner.'

15 We should point out that the phrase 'open to the south' does not mean simply that *yamen* entrances are on the south side. In fact they are but this is an idiomatic expression complaining that *yamen* or government officials operate for themselves rather than for the people as a whole.