Aspect Marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for Contrastive Language Study[1]

Anthony McEnery, Zhonghua Xiao
Lancaster University, Lancaster, UK

Lili Mo
Ningbo University, Ningbo, China

Abstract

This paper presents the newly released Lancaster Corpus of Mandarin Chinese (LCMC), a Chinese match for the FLOB and Frown corpora of British and American English. We first discuss the major decisions we took when building the corpus. These relate to sampling, text collection, mark-up, and annotation. Following from this we use the corpus to study aspect marking in Chinese and British/American English. The study shows that although Chinese and English are typologically different, aspect markers in the two languages show a strikingly similar distribution pattern, especially across the two broad categories of narrative and expository texts. The study also reveals some important differences in the distribution of aspect markers in Chinese versus English and British versus American English across fifteen text categories, and provides an account of these differences.

Correspondence:
Anthony McEnery,
Department of Linguistics,
Lancaster University,
Lancaster LA1 4YT, UK.
E-mail:
a.mcenery@lancaster.ac.uk

1 Introduction

The Lancaster Corpus of Mandarin Chinese (LCMC) is a one-million-word balanced corpus of written Mandarin Chinese. The corpus was created as part of the research project *Contrasting tense and aspect in English and Chinese* funded by the UK Economic and Social Research Council.[2] We built the LCMC corpus in response to the general lack of publicly available balanced corpora of Chinese. Although there are some Chinese corpus resources, most of them, for example the PH corpus and the PFR People's Daily corpus,[3] are composed exclusively of news texts and are thus not balanced. Neither are the Chinese corpora released by the LDC balanced (e.g. TREC Mandarin, Chinese Gigaword, Mandarin Chinese News Text).[4] The latter contain only either newswire texts or

[1] We are grateful to the referees of *Literary and Linguistic Computing* for their insightful and constructive comments made on an earlier version of this paper. It was on their suggestion that the statistical tests for individual text categories were produced in Section 5.

official documents for written Chinese. The only balanced corpus of Mandarin Chinese is the Sinica Corpus, which was produced by Academia Sinica, Taiwan.[5] As a result of Taiwan being separated politically from mainland China for decades, the language used in Taiwan has diverged from that used on the mainland.[6] As such, the Sinica corpus does not represent modern Mandarin Chinese as written in mainland China. The balanced corpus of Chinese built in China, as reported in Zhou and Yu (1997), is not publicly available.

Given the available corpus resources for Chinese corpus linguistics and our desire to use a balanced corpus of modern Mandarin Chinese from mainland China to contrast English and Chinese, we decided to build the LCMC.[7] Our decision to build this corpus links directly to the organization of this paper; in this paper we need to both introduce LCMC and demonstrate why we believed it was useful to build such a resource for contrastive linguistics. We will introduce the corpus in Sections 2–4 of this paper, where we will outline the principal considerations involved in the construction of LCMC. To demonstrate the usefulness of the corpus, in Section 5 we will use LCMC to test the claim made recently by McEnery and Xiao (2002, pp. 224–5) that aspect markers in English and Chinese are significantly more frequent in narrative texts than in expository texts.[8] We will also compare the distribution patterns of aspect markers across the various text categories in LCMC and FLOB/Frown.

2 Sampling Frame and Text Collection

As the LCMC corpus was designed principally with contrastive research in mind, we first needed to make a decision regarding which English corpus we should use for contrastive purposes so that we could follow its sampling frame. Given the limited resources available to us, it was not feasible to create a Chinese match for the 100-million-word British National Corpus (BNC).[9] The limited availability of electronic Chinese texts from the early 1960s made the compilation of a match for the LOB (the Lancaster–Oslo–Bergen corpus, see Johansson *et al.*, 1978) or Brown (Francis and Kuãera, 1964) corpus infeasible. Having rejected building a match for LOB/Brown and the BNC, we decided to create a match for FLOB (Hundt *et al.*, 1998), a balanced corpus of British English, as FLOB sampled from a period in which electronic Chinese texts were produced in reasonable quantity (1991–1992). Also, FLOB, at one million words, was large enough to be useful, yet small enough for us to be able to build a Chinese match with relative ease. A further attraction of FLOB is that it has a matching American English corpus, Frown (Hunt *et al.*, 1999). Hence by building a match for FLOB we allowed a contrast of Chinese with the two major varieties of English.

FLOB, following the Brown/LOB model, contains five hundred 2,000-word samples of written British English texts sampled from fifteen text categories in 1991–1992, totalling one million words. The components of FLOB are given in Table 1.

[2] We thank the UK ESRC for funding our research project (Grant Ref. RES-000-220135), without which this corpus could not have been built.

[3] A brief description and the corpus can be accessed online at ftp://ftp.cogsci.ed.ac.uk/pub/chinese/. The PFR People's Daily Corpus is composed of newswire texts from *People's Daily* in 1998. A sample of the corpus is accessible online at http://icl.pku.edu.cn/Introduction/corpustagging.htm.

[4] For details of Chinese corpora available at the LDC, visit http://www.ldc.upenn.edu/Catalog/ and use 'Chinese' as the search word.

[5] See http://www.sinica.edu.tw/ftms-bin/kiwi1/mkiwi.sh?language=1 for a brief description of the corpus. The corpus query system can be accessed online at http://www.sinica.edu.tw/ftmsbin/kiwi1/mkiwi.sh?ukey=542949389 andlanguage=1andqtype=-1.

[6] In Taiwanese Mandarin, for example, 有 *you* can function as a perfective marker indicating the actualization of a situation, especially in conversations. Speakers of mainland Mandarin find this usage awkward and even ungrammatical (see Christensen, 1994).

Table 1 FLOB text category

Code	Text category	No. of samples	Proportion (%)
A	Press reportage	44	8.8
B	Press editorials	27	5.4
C	Press reviews	17	3.4
D	Religion	17	3.4
E	Skills, trades, and hobbies	38	7.6
F	Popular lore	44	8.8
G	Biographies and essays	77	15.4
H	Miscellaneous (reports, official documents)	30	6
J	Science (academic prose)	80	16
K	General fiction	29	5.8
L	Mystery and detective fiction	24	4.8
M	Science fiction	6	1.2
N	Western and adventure fiction	29	5.8
P	Romantic fiction	29	5.8
R	Humour	9	1.8
Total		500	100

In LCMC, the FLOB sampling frame is followed strictly except for two minor variations. The first variation relates to the sampling frame—we replaced *western and adventure fiction* (category N) with *martial arts fiction*. There are three reasons for this decision. First, there is virtually no western fiction written in Chinese for a mainland Chinese audience. Second, martial arts fiction is broadly a type of adventure fiction and as such can reasonably be viewed as category N material. It is also a very popular and important fiction type in China and hence should be represented. Finally, the language used in martial arts fiction is a distinctive language type and hence, given the wide distribution of martial arts fiction in China, once more one would wish to sample it. The language of the martial arts fiction texts is distinctive in that even though these texts were published recently, they are written in a form of vernacular Chinese, i.e. modern Chinese styled to appear like classical Chinese. Although the inclusion of this text type has made the tasks of part-of-speech (POS) tagging and the post-editing of the corpus more difficult, the inclusion of the texts has also made it possible for researchers to compare representations of vernacular Chinese and modern Chinese.

The second variation in the sampling frame adopted from FLOB was caused by problems we encountered while trying to keep to the FLOB sampling period. Because of the poor availability of Chinese electronic texts in some categories (notably F, D, E, and R) for 1991, we were forced to modify the FLOB sampling period slightly by including some samples ±2 years of 1991 when there were not enough samples readily available for 1991. As can be seen from Table 2, most of the texts were produced ±1 year of 1991. We assume that varying the sampling frame in this way will not influence the language represented in the corpus significantly.

LCMC has been constructed using written Mandarin Chinese texts published in mainland China to ensure some degree of textual homogeneity. It should be noted that the corpus is composed of written textual

[7] The corpus is distributed free of charge for use in non-profit-making research. The manual accompanying the corpus, as well as the details for ordering, can be accessed online at the corpus website http://www.ling.lancs.ac.uk/corplang/lcmc or the Chinese mirror site hosted by the Chinese Academy of Social Science http://www.cass.net.cn/chinese/s18_yys/dangdai/LCMC/LCMC.htm.

[8] The narrative versus expository distinction ‘might also be considered as distinguishing between active, event-oriented discourse and more static, descriptive or expository types of discourse’ (Biber, 1988, p. 109). Narrative discourse is basically event-oriented whereas expository discourse has an informational focus. See Biber (1988) for a discussion of the relation between discourse functions and linguistic features.

[9] See the BNC website http://www.natcorp.ox.ac.uk/.

Table 2 Sampling period of LCMC (all values are percentages)

Code	1989	1990	1991	1992	1993
A	—	22.7	72.7	2.3	2.3
B	7.4	14.8	51.9	3.70	22.2
C	—	5.9	88.2	5.9	—
D	5.9	17.6	41.2	11.8	23.5
E	—	23.7	44.7	10.5	21.1
F	6.8	25	29.5	13.6	25
G	1.3	10.4	64.9	16.9	6.5
H	—	—	100	—	—
J	1.2	7.5	72.5	17.5	1.3
K	—	—	79.3	13.8	6.9
L	—	8.3	62.5	16.7	12.5
M	—	—	100	—	—
N	3.4	13.8	48.3	31.1	3.4
P	10.3	6.9	55.2	20.7	6.9
R	—	—	44.4	22.2	33.3

data only, with items such as graphics and tables in the original texts replaced by <gap> elements in the corpus texts. Long citations from translated texts or texts produced outside the sampling period were also replaced by <gap> elements so that the effect of translationese could be excluded (McEnery and Xiao, forthcoming) and L1 quality guaranteed.

Although a small number of samples, if they were conformant with our sampling frame, were collected from the Internet, most samples were provided by the SSReader Digital Library in China. As each page of the electronic books in the library comes in PDG format, these pages were transformed into text files using an OCR module provided by the digital library. This scanning process resulted in a 1–3 per cent error rate, depending on the quality of the picture files. Each electronic text file was proofread and corrected independently by two native speakers of Mandarin Chinese so as to keep the electronic texts as faithful to the original as possible.

Although the digital library has a very large collection of books, it does not provide complete newspapers, but provides texts from newspapers or newswire stories instead. News texts in the library are grouped into a dozen collections of news arranged to reflect broad differences of text types (e.g. newswire versus newspaper articles) or medium (e.g. newspaper texts versus broadcast news scripts). These collections, however, represent news texts from more than eighty newspapers and television or broadcasting stations. The samples from these sources account for around two-thirds of the texts for the press categories (A–C) in LCMC. The other third was sampled from newswire texts from the Xinhua News Agency.[10] Considering that this is the most important and representative news provider in China, roughly analogous to the Associated Press in the USA/UK, we believe that the high proportion of material taken from the Xinhua News Agency is justified.

Unlike languages such as English, in which words are typically delimited by white space and thus word counts can be produced in

[10] The texts were sampled from the PH corpus compiled by Guo Jin. The corpus contains 2.4 million tokens of raw texts from the Xinhua News Agency, written between January 1990 and March 1991.

written texts relatively easily, Chinese contains running characters (see Section 4). Consequently, whereas it is easy to count the number of characters in a text it is much more difficult to count the number of words. The difficulty in word counting in Chinese posed a challenge for us, as we wanted to extract roughly 2,000-word chunks from larger texts. Rather than count the words by hand, which would have proved time consuming, we proceeded by estimating a character to word ratio. Based on a pilot study carried out by us we decided to adopt a ratio of 1:1.6, which meant that we needed a 3,200-character running text to gather a 2,000-word sample. When a text was less than the required length, texts of similar quality were combined into one sample. For longer samples, e.g. those from books, we adopted a random procedure so that beginning, middle, and ending samples of texts were included in all categories. It should be noted that in selecting chunks we operated a bias in favour of textually coherent chunks that fitted our sampling size, e.g. we favoured samples that did not split paragraphs over those that did. Although the character to word ratio we adopted worked on most text types, it also resulted in some samples of slightly more than 2,000 words, and some of slightly fewer than 2,000 words. This was typically the case where texts contained a large number of proper nouns or idioms, some of which are four-character or even seven-character words. Consequently, when these samples were processed and it was possible to count the number of words easily, we were forced to adjust the size of each sample that was finally included in the corpus. The adjustment was done by cutting longer samples to roughly 2,000 words while avoiding truncating the last sentence or reducing the whole sample to fewer than 2,000 words. Nonetheless, although some individual samples still contain fewer and a few more words than 2,000, the total number of words for each text type is roughly conformant to our sampling frame.

3 Encoding and Markup

Unlike writing systems typically encoded by single bytes, such as the Roman alphabet, the Chinese writing system typically requires two bytes. Currently there are three dominant encoding systems for Chinese characters: GB2312 for simplified Chinese, Big5 for traditional Chinese, and Unicode. Both GB2312 and Big5 are double-byte encoding systems. Although the original corpus texts were encoded in GB2312, we decided to convert the encoding to Unicode (UTF-8) for the following reasons: (1) to ensure the compatibility of a non-Chinese operating system and Chinese characters; (2) to take advantage of the latest Unicode-compliant concordancers such as *Xara* (Burnard and Todd, 2003) and *WordSmith Tools* version 4.0.

To make it more convenient for users of our corpus with an operating system earlier than Windows 2000 and no language support pack to use our data, we have produced a Romanized Pinyin version of the LCMC corpus in addition to the standard version containing Chinese characters. Although also encoded using UTF-8, the Pinyin version is more

Table 3 XML elements of text

Level	Code	Gloss	Attribute	Value
1	text	Text type	TYPE	As per Table 1 *Text category*
			ID	As per Table 1 *Code*
2	file	Corpus file	ID	Text ID plus individual file number starting from 01
3	p	Paragraph	—	—
4	s	Sentence	n	Starting from 0001 onwards
5	w	Word	POS	Part-of-speech tags as per the LCMC tagset
	c	Punctuation and symbol	POS	As per the LCMC tagset
	gap	Omission	—	—

compatible with older operating and concordance systems. This is also of assistance to users who can read Romanized Chinese but not Chinese characters.

Both versions of the corpus are composed of fifteen text categories. Each category is stored as a single file. The corpus is XML conformant. Each file has two parts: a corpus header and the text itself. The header contains general information about the corpus. The text part is annotated with five main features, as shown in Table 3.

These details are useful when using an XML-aware concordancer such as *Xara* version 1.0. With this tool, users can either search the whole corpus or define a subcorpus containing a certain text type or a specific file. The POS tags allow users to search for a certain class of words, and in combination with tokens, to extract a specific word that belongs to a certain class.

4 Corpus Processing

We undertook two forms of corpus annotation on the LCMC corpus: word segmentation and part-of-speech annotation. To deal with each of these in turn, word segmentation is an essential and non-trivial process in Chinese corpus linguistics (see Wu and Fung, 1994; Sun *et al.*, 1998; Swen and Yu, 1999). Segmentation, or tokenization, refers to the process of segmenting text strings into *word tokens*, i.e. defining *words* (as opposed to *characters*) in a running text. For alphabetic languages, as word tokens are generally delimited clearly by a preceding white space and a following white space or a new-line character, 'the one-to-one correspondence between an orthographic token and a morphosyntactic token can be considered the default case that applies in the absence of special conditions' (Leech, 1997, pp. 21–4).[11] However, for Chinese (and some other Asian languages such as Japanese and Thai),[12] word segmentation is not a trivial task, for, as noted already, a Chinese sentence is written as an unseparated string of characters.

Readers unfamiliar with Asian languages such as Chinese may think it strange that segmentation is such a vital process in Chinese corpus linguistics. Yet segmentation in Chinese corpus linguistics is vital for at

[11] The three exceptional conditions are multiwords, mergers, and compounds. For details refer to Leech (1997, pp. 21–4).

[12] See http://www.milab.is.tsukuba.ac.jp/wor-seg-ac199.

least two reasons. First, although a rough character to word correspondence in Chinese does at times exist, it is not possible to simply search for a character and assume it is a word, or always part of one word. Some characters in Chinese are meaningless. For example, 琵 *pi* is meaningful only when it goes with 琶 *pa* to form a word, i.e. 琵琶 *pipa* (a musical instrument). Second, the main purpose of segmentation is disambiguation. Consider the running text 他们不得不过一个灰色的圣诞节 *tamen budebu guo yi-ge huise de shengdanjie* 'They had to spend a grey Christmas'. The underlined part, when taken in isolation, can be segmented in different ways: (1) 不得 (must not) 不 (not) 过 (spend), (2) 不得 (must not) 不过 (but; only), or (3) 不得不 (have to) 过 (spend), although in this example only (3) is meaningful. Literate speakers of this language do not have any difficulty interpreting this sentence in its written form, precisely because they are actually segmenting the character string as they read it. Imagine that modern English did not use white space to delimit words in texts. When we search for the word *them* in the introduction of this paper, we would find both *them* and *the mainland*, which is not what we want. It is to avoid such meaningless corpus retrieval that segmentation is undertaken. As *words* are the basis of most corpus searching and retrieval tasks such meaningless retrieval is a real problem in Chinese corpus linguistics. It is for this reason that in Chinese corpus linguistics any string of characters in a corpus text must first be converted into legitimate words, typically prior to any further linguistic analysis being undertaken (see Feng, 2001; Xia *et al.*, 2000) because 'in computational terms, no serious Chinese language processing can be done without segmentation' (Huang *et al.*, 1997, p. 47).

The segmentation tool we used to process the LCMC corpus is the Chinese Lexical Analysis System developed by the Institute of Computing Technology, Chinese Academy of Sciences.[13] The core of the system lexicon incorporates a frequency dictionary of 80,000 words with part-of-speech information. The system is based on a multi-layer hidden Markov model and integrates modules for word segmentation, part-of-speech tagging and unknown word recognition (see Zhang *et al.*, 2002). The rough segmentation module of the system is based on the *n* shortest paths method (Zhang and Liu, 2002). The model, based on the two shortest paths, achieves a precision rate of 97.58 per cent, with a recall rate as high as 99.94 per cent (Zhang and Liu, 2002). In addition, the average number of segmentation candidates is reduced by sixty-four times compared with the full segmentation method. The unknown word recognition module of the system is based on role tagging. The module applies the Viterbi algorithm to determine the sequence of roles (e.g. internal constituents and context) with the greatest probability in a sentence, on the basis of which template matching is carried out. The integrated ICTCLAS system is reported to achieve a precision rate of 97.16 per cent for tagging, with a recall rate of over 90 per cent for unknown words and 98 per cent for Chinese person names (Zhang and Liu, 2002).

However, the POS system is in part under-specified, especially in the crucial area of aspect marking (see Section 5 for a discussion of aspect in

[13] We thank Kelvin H. Zhang for allowing us to use his system to annotate our corpus. Readers interested can visit http://mtgroup.ict.ac.cn/~zhp/ICTCLAS.htm to test the system or visit http://www.nlp.org.cn to download the Windows version of the system.

Chinese). For example, the system does not differentiate between the preposition *zai* and the aspect marker *zai*. Furthermore, as the system was trained using news texts, its performance on some text types (e.g. martial arts fiction) is poor. For example, although 道 *dao* is used much more frequently as a verb meaning 'say' in martial arts fiction than in other text types, it was tagged by the system as a classifier or noun (i.e. 'road').[14] As such, we decided to undertake post-editing of the processed corpus to classify all of the instances of the four aspect markers (*-le*, *-guo*, *-zhe*, and *zai*) according to the aspect annotation system of Xiao and McEnery (forthcoming). In addition, except for the three categories of news texts and the reports/official documents, on which the system performs exceptionally well, all of the processed texts were hand-checked and corrected. The post-editing improved the annotation precision to over 98 per cent.[15] As a final step, the post-edited corpus files were converted into XML format.

5 Distribution of Aspect Markers in LCMC/FLOB/Frown

Having built LCMC, we decided to use the corpus to test a claim made by McEnery and Xiao (2002, pp. 224–5); those workers, based on a study of public health documents in Chinese and English, claimed that aspect markers occur significantly more frequently in narrative texts than in expository texts. However, McEnery and Xiao only studied one genre. Does this claim hold across a wider range of genres? Also, they only contrasted British English and Chinese. Is the claim true when American English and Chinese are contrasted, or American English and British English? We decided to explore these questions by examining the distribution of aspect markers in the fifteen text categories of the LCMC and FLOB/Frown corpora. In so doing, we were also able to compare the distribution patterns of aspect markers in Chinese and British/American English.

However, before proceeding to the analysis, a brief description of the aspect system of Chinese is needed, as Chinese has a very complicated aspect marker system. In Chinese the perfective aspect is marked by 了 *-le*, 过 *-guo*, verb reduplication and resultative verb complements (RVCs), whereas the imperfective aspect is marked by 在 *zai*, 着 *-zhe*, 起来 *-qilai*, and 下去 *-xiaqu* (see Xiao and McEnery, forthcoming).[16] In addition, covert aspect marking is also an important strategy used to express aspectual meanings in Chinese discourse (see McEnery and Xiao, 2002, p. 212). However, as the tagger we used only annotated 了 *-le*, 过 *-guo*, 在 *zai*, and 着 *-zhe*, we decided to explore these four aspect markers in LCMC in this study. The frequencies of these aspect markers in LCMC are as shown in Table 4.[17]

English is a less aspectual language with regard to grammatical aspect marking than Chinese. English only differentiates between the simplex viewpoints of the progressive, the perfect and the simple aspect, in

[14] In martial arts fiction, the monosyllable 道 *dao* is typically used as a verb meaning 'say'. Although in other text categories, 道 *dao* is also used as a verb to mean 'say', it typically occurs in disyllabic compound verbs such as 说道 *shuodao* 'say', 笑道 *xiaodao* 'say with a smile', 哭道 *kudao* 'say while crying' and 喊道 *handao* 'shout, yell'.

[15] We checked around 2,000 words from each text category and the precision rate quoted is the average result achieved in this evaluation.

[16] RVC is an acronym for 'resultative verb complement'. RVCs indicate the phase, resultant state or direction of the situation denoted by preceding verbs in a resultative compound, such as *open* in *push the door open*. In Chinese, there are three types of RVCs: *completive*, *result-state*, and *directional* RVCs. RVCs contribute both to situation aspect, by attaching a result to their preceding verbs, and to grammatical aspect, by marking the completiveness of a situation (see Xiao and McEnery, forthcoming).

[17] Readers can visit the corpus website given above to find out how to explore the corpus using *Xara*.

Table 4 Distribution of aspect markers in LCMC

Average	Text type	Words (10,000)	Frequency	Frequency per 10,000 words	Per cent
Above the average	K	5.8	1,674	289	12.00
	M	1.2	322	268	11.13
	P	5.8	1,384	238	9.88
	R	1.8	387	215	8.92
	L	4.8	1,024	214	8.88
	G	15.4	3,140	204	8.47
	N	5.8	1,107	191	7.93
	A	8.8	1,539	175	7.26
Average	Average of frequency per 10,000 words: 161 (6.68)				
Below the average	F	8.8	1,057	120	4.98
	C	3.4	365	108	4.48
	D	3.4	363	106	4.40
	B	5.4	561	104	4.32
	J	16.0	1,355	84	3.49
	E	7.6	412	54	2.24
	H	6.0	231	39	1.62

addition to the complex viewpoint of the perfect progressive (see Biber *et al.*, 1999, p. 461; Svalberg and Chuchu, 1998). In English, perfective meaning is most commonly expressed by the simple past (see Brinton, 1988, p. 52), although the perfect can also mark perfectivity (Dahl, 1999, p. 34). Imperfective meaning is typically signalled by the progressive, and less often by the perfect progressive. For the purpose of contrasting English aspect marking with Chinese we counted the distribution of the four aspects of English. The frequencies of aspect markers in FLOB and Frown are given in Tables 5 and 6.[18]

Tables 4–6 show that in both LCMC and FLOB/Frown, the text categories where the frequency of aspect markers is above average (categories L, M, N, P, R, and K) or near to the average (categories A and G) are the five fiction categories plus humour, biography, and press reportage. The text types where aspect markers occur least frequently include reports/official documents, academic prose, skills/trades/hobbies, press reviews, press editorials, religion, and popular lore. In both Chinese and the two major varieties of English considered here, there is a great difference in usage between the first and second groups of texts, which indicates that the two are basically different. Text types such as fiction, humour, and biography are narrative whereas reports/official documents, academic prose, and skills/trades/hobbies are expository. Press reportage is a transitory category that is more akin to narrative texts.

Log-likelihood (LL) tests indicate that in both Chinese and the two varieties of English, the differences between the distribution of aspect markers in narrative and expository texts are statistically significant (see Table 7).[19] In all of the three corpora, aspect markers occur in narrative texts twice as frequently as in expository texts (2.43 times in LCMC, 2.21

[18] Readers who wish to reduplicate this case study must note that (1) *had* as an auxiliary should not be counted as the simple past form of *have* and (2) the perfect does not include the perfect progressive, which is counted separately. We used *WordSmith* version 3 to extract the required frequency data from FLOB and Frown. Simple past forms include (1) all past forms of a lexical verb, verbs *do* and *be*; (2) all instances of the past form *had* (including the contracted form) not followed by a past participle within a four-word range to the right of the search word. Perfect constructions include all morphological forms of *have* (except *having*) followed by 0–2 words and then by a past participle, but not followed by a present participle within a four-word range to the right of the search pattern. The progressive forms (including the perfect progressive) can be extracted using the search pattern of all forms of verb *be* followed by 0–2 words and then the present participles of all verbs.

[19] For one degree of freedom, the calculated value must be greater than 3.84 (i.e. the significance level $P < 0.05$) for a difference to be statistically significant. The critical value for the significance level $P < 0.001$ is 10.83.

Table 5 Distribution of aspect markers in FLOB

		Words		Frequency per	
Average	**Text type**	**(10,000)**	**Frequency**	**10,000 words**	**Per cent**
Above (or near) the average	P	5.8	5,673	978	11.17
	L	4.8	4,624	963	11.00
	N	5.8	5,255	906	10.34
	K	5.8	5,169	891	10.17
	M	1.2	997	831	9.49
	R	1.8	1,313	729	8.32
	A	8.8	5,166	587	6.70
	G	15.4	8,257	536	6.12
Average	Average of frequency per 10,000 words: 584 (6.67)				
Below the average	D	3.4	1,317	388	4.43
	F	8.8	3,353	381	4.35
	E	7.6	2,724	358	4.09
	B	5.4	1,886	349	3.98
	H	6.0	1,740	290	3.31
	C	3.4	978	288	3.29
	J	16.0	4,524	283	3.23

Table 6 Distribution of aspect markers in Frown

		Words		Frequency per	
Average	**Text type**	**(10,000)**	**Frequency**	**10,000 words**	**Per cent**
Above (or near) the average	L	4.8	4,546	947	10.95
	M	1.2	1,119	933	10.78
	N	5.8	5,349	922	10.66
	P	5.8	5,238	903	10.44
	R	1.8	1,534	852	9.85
	K	5.8	4,815	830	9.59
	A	8.8	4,816	547	6.32
	G	15.4	7,799	506	5.58
Average	Average of frequency per 10,000 words: 577 (6.67)				
Below the average	F	8.8	3,397	386	4.46
	B	5.4	1,893	351	4.06
	E	7.6	2,617	344	3.98
	C	3.4	1,155	340	3.93
	D	3.4	1,053	310	3.58
	J	16.0	4,024	252	2.91
	H	6.0	1,368	228	2.64

Table 7 Distribution of aspect markers in narrative and expository texts

Corpus	**Discourse type**	**Categories**	**Words**	**Markers**	**LL score**	**Sig. level**
LCMC	Narrative	K–R, A, G	494,000	10,577	2796.53	<0.001
	Expository	B–F, H, J	506,000	4,344		
FLOB	Narrative	K–R, A, G	494,000	36,454	7771.37	<0.001
	Expository	B–F, H, J	506,000	16,522		
Frown	Narrative	K–R, A, G	494,000	35,216	7950.98	<0.001
	Expository	B–F, H, J	506,000	15,507		

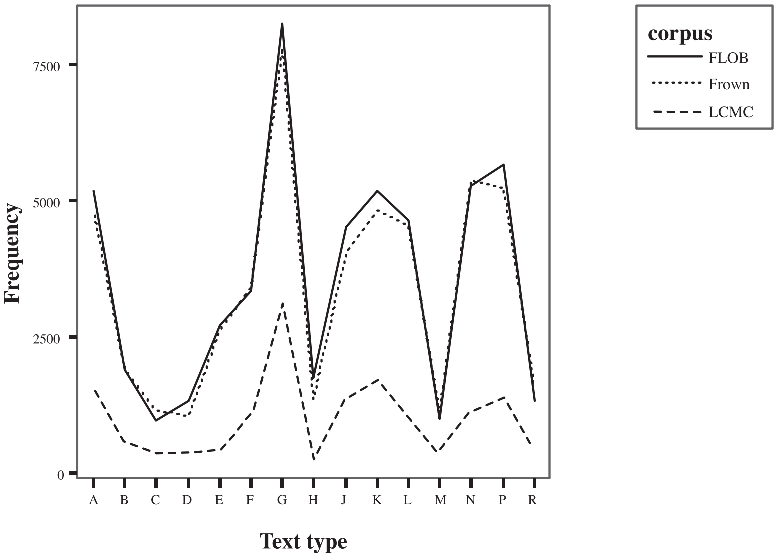

Fig. 1 Distribution of aspect markers (frequency).

times in FLOB, and 2.27 times in Frown), which means that the higher frequency of aspect markers in narrative texts over expository texts is a common feature of Chinese and the two major varieties of English.

These findings confirm those of McEnery and Xiao (2002) and allow us to generalize this claim from the domain studied by McEnery and Xiao, public health, to English/Chinese in general. As can be seen from Fig. 1, whereas the two languages differ typologically, they show a strikingly similar distribution pattern of aspect markers. It is also interesting to note that whereas British English and American English have developed variations in spelling (e.g. *behaviour* versus *behavior*), word choice (e.g. *petrol* versus *gasoline*), and grammar (e.g. American English has two participle forms for the verb *get*, namely *got* and *gotten*, whereas British English only uses the form *got*) (cf. Biber *et al.*, 1999, p. 19), their use of aspect is strikingly similar—the curves for the distribution of aspect markers for FLOB and Frown are almost identical (see Fig. 1).

Chinese and English, however, do show some differences in the distribution of aspect markers, as shown in Fig. 2. The figure shows the frequencies of aspect markers, as percentages, in the fifteen text categories in the three corpora. As can be seen, by comparison with the two major varieties of English, aspect markers in Chinese occur more frequently in categories G and K but less frequently in N, L, H, and E.[20] The relatively low frequency of aspect markers in category N (martial arts fiction) in relation to other fiction types, as noted in Section 2, is shown even more markedly in the contrast of the N category between LCMC and FLOB/Frown. British English and American English also differ in that the latter variety does not show such a marked fluctuation in aspect marking in narrative texts, notably in biography and the five types of fiction.

The general patterns as shown in Fig. 2, however, may mask some important differences in aspect marking in English and Chinese. They

[20] As the aspect and tense markers in English combine morphologically, English typically registers a considerably higher frequency of aspect/tense markers than Chinese. In terms of proportions, however, aspect markers are more common in Chinese for categories G and K but less frequent for categories N, L, H, and E (see Table 8).

may also mask some differences between the two major varieties of English, although the contrast between the varieties is not as marked as that between Chinese and English, as shown in Table 8. This table gives the log-likelihood scores and significance levels of individual text categories (one degree of freedom), where statistically significant values are highlighted. This table can be read in conjunction with Fig. 2 or Tables 4–6 to identify the text categories where aspect markers are significantly more (or less) common in Chinese and British/American English.[21]

The picture becomes clearer if we examine perfective and imperfective markers separately. Figure 3 shows the percentages of perfective markers occurring in each text category in the three corpora. As can be seen in the

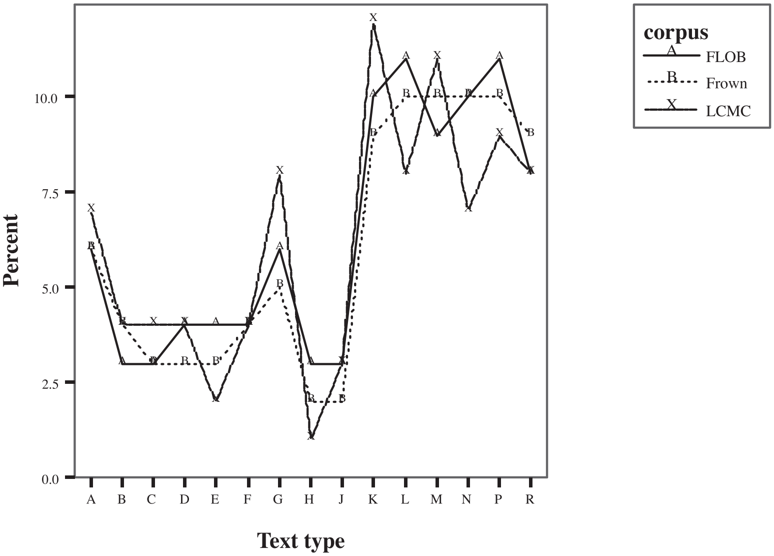

Fig. 2 Distribution of aspect markers (percentage).

Table 8 Contrasting the distribution of aspect markers

	LCMC versus FLOB		LCMC versus Frown		FLOB versus Frown	
Category	LL score	Sig. level	LL score	Sig. level	LL score	Sig. level
A	0.925	0.336	2.672	0.102	1.029	0.310
B	0.528	0.467	0.319	0.572	0.059	0.808
C	**7.461**	**0.006**	1.448	0.229	**5.160**	**0.023**
D	0.004	0.949	3.348	0.067	**8.127**	**0.004**
E	**20.245**	**<0.001**	**18.162**	**<0.001**	0.139	0.709
F	1.714	0.191	1.142	0.285	0.129	0.720
G	**15.925**	**<0.001**	20.211	**<0.001**	0.569	0.451
H	**21.608**	**<0.001**	**21.937**	**<0.001**	0.004	0.948
J	0.383	0.536	2.040	0.153	1.482	0.223
K	**6.467**	**0.011**	**11.530**	**0.001**	1.640	0.200
L	**9.269**	**0.002**	**8.844**	**0.003**	0.011	0.918
M	**5.553**	**0.018**	0.224	0.636	**8.036**	**0.005**
N	**13.032**	**<0.001**	**16.305**	**<0.001**	0.453	0.501
P	3.294	0.070	0.641	0.423	2.400	0.121
R	0.871	0.351	1.875	0.171	**12.264**	**<0.001**

[21] The calculations in Table 8 are based on standardized frequencies (per 10,000 words). To save space, we give here only the results, not the process of deriving the statistical tests. The following example shows how these values were obtained. To calculate the LL score of text category A in LCMC and FLOB, for example, we first found the standardized frequency of aspect markers in category A in LCMC (i.e. 175) and FLOB (i.e. 587). Then we subtracted these frequencies from the overall standardized frequency of aspect markers in LCMC (i.e. 2,409 minus 175) and FLOB (8,758 minus 587) to obtain the standardized frequency of aspect markers in other categories. The LL score 0.925 was obtained by cross tabulating the four frequencies.

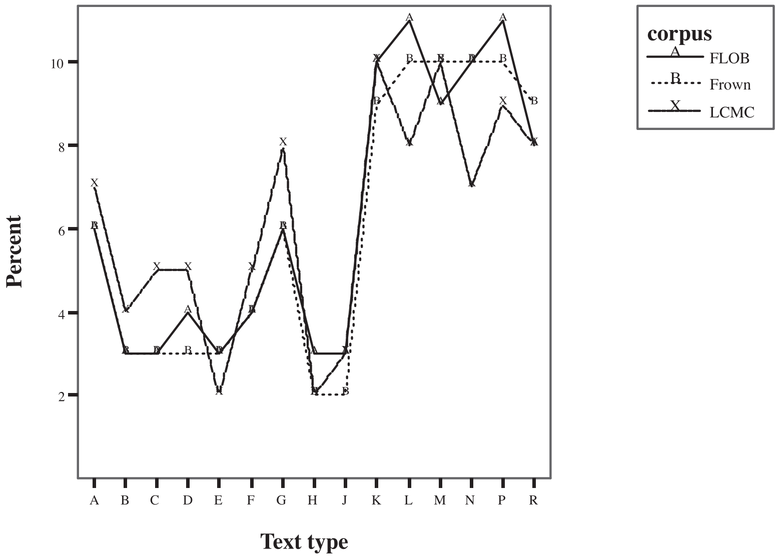

Fig. 3 Perfective aspect markers in LCMC/FLOB/Frown.

figure, in expository texts (barring category E) perfective aspect markers in LCMC generally occur more frequently than those in English, whereas in narrative texts (except for category G) perfective markers in English are generally more frequent than those in Chinese. The relatively high frequency of perfective markers in narrative texts and their lower frequency in expository texts in English can be accounted for by the fact that aspect markers in English express both temporal and aspectual meanings. In total, 82.5 per cent of the 48,902 perfective markers in FLOB, and 84.8 per cent of the 46,866 perfective markers in Frown, are simple past forms. Narrative texts are normally related to what happened in the past whereas expository texts are typically non-past. Hence the relatively higher frequency of perfective markers in narrative as opposed to expository texts is understandable.

As would be expected, the contrast between British English and American English is once again not as marked as that between Chinese and English (see Fig. 3). The two varieties of English show more similarity in expository texts than in narrative texts. In expository texts, the two varieties show a very similar distribution pattern except that British English registers a slightly greater percentage of perfective markers in categories D, H, and J. Similarly, in narrative texts (except categories M and R), British English generally shows a greater frequency of usage than American English. One possible explanation for this is that although the perfect aspect is typically more common in British English, the contrast in narrative texts is more marked than in expository texts, as shown in Fig. 4.

This finding is in line with Biber *et al.* (1999, p. 462), who observe that in British English news the perfect aspect is much more common than in American English news. Although the contrast between the three news categories in FLOB and Frown is not as marked as that observed by Biber

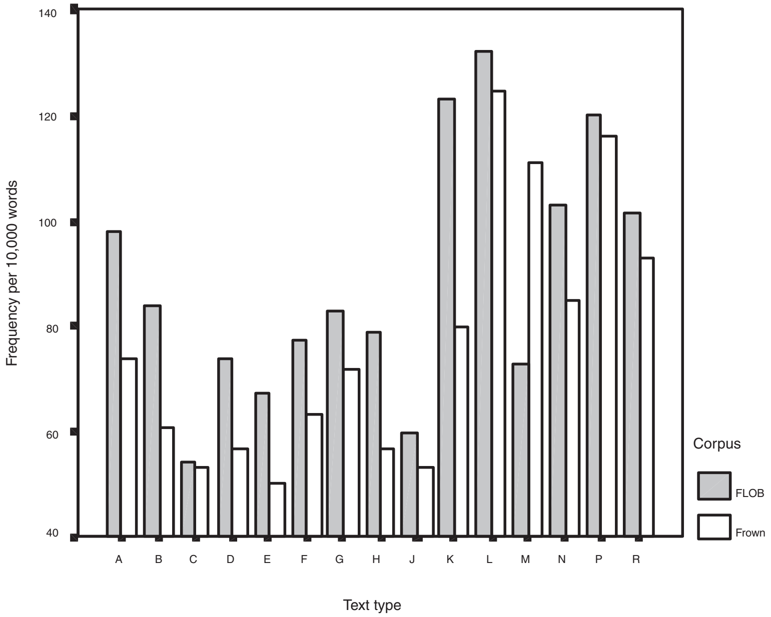

Fig. 4 Distribution of the perfect in FLOB/Frown.

et al. (1.29 times more frequent in FLOB news categories), the perfect is indeed more frequent in nearly all of the text types in FLOB (except for category M). The wide coverage of the perfect lends further credence to the claim of Biber *et al.* that 'BrE strongly favours the perfect in comparison with AmE' (Biber *et al.*, 1999). However, one should note that the ratio of the perfect in FLOB and Frown (1.15:1) is slightly lower than that reported by Biber *et al.* (1.33:1).

In marked contrast, as can be seen in Fig. 5, imperfective aspect markers show a totally different distribution pattern from perfective markers. In expository texts, imperfective markers in both varieties of English typically occur more frequently than those in Chinese whereas in narrative texts, imperfective markers in Chinese are generally more frequent than those in English.

This phenomenon can be explained as follows. First, the Chinese progressive marked by *zai* can only signal progressiveness literally. In contrast, 'the progressive in English has a number of other specific uses that do not seem to fit under the general definition of progressiveness' (Comrie, 1976, p. 37). Although the different uses of the progressive in English and Chinese account for the slightly higher frequency of the English imperfective markers in expository texts, this cannot explain the relatively low frequency of these markers in narrative texts. Nevertheless, we can find an answer in the Chinese imperfective marker *-zhe*, which accounts for 88 per cent of the 3,836 instances of imperfective markers in LCMC. This marker has three basic functions: to signal the durative nature of a situation, to serve with a verb as an adverbial modifier to provide background information, and to occur in locative inversion to indicate existential status (Xiao and McEnery, forthcoming). Of the three

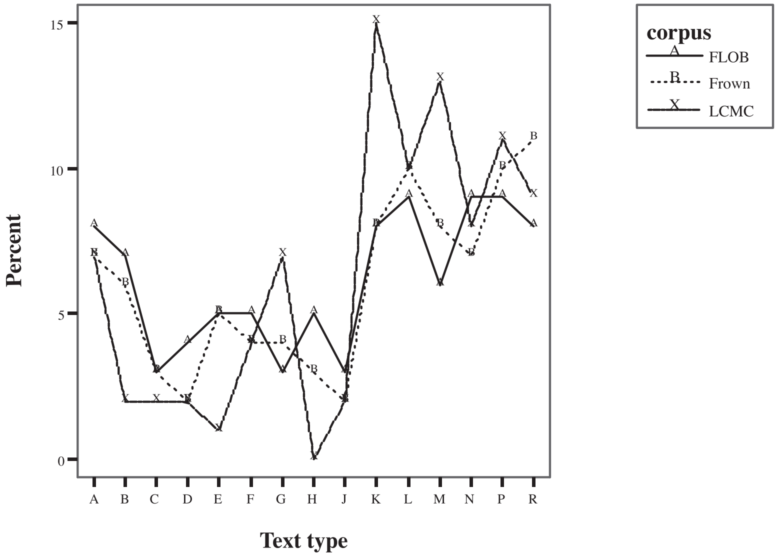

Fig. 5 Imperfective aspect markers in LCMC/FLOB/Frown.

functions of *-zhe*, only the first is used in expository texts. Hence, in spite of the high overall frequency of *-zhe* in LCMC, only about 20 per cent of all examples of *-zhe* occur in expository texts. In contrast, all of the three functions of *-zhe* apply to narrative texts. Furthermore, in addition to inducing a background effect, *-zhe* can also be used in an apparently ‘foregrounded’ situation to move narration forward (see Du, 1999; Xiao and McEnery, forthcoming). As such, it is hardly surprising that Chinese imperfective markers occur more frequently in narrative texts than English imperfective markers.

Figure 5 also shows some important differences in the distribution of imperfective markers in British English and American English. In expository texts, imperfective markers in British English are typically more common than in American English whereas in narrative texts (except for category N and less markedly for the transitory category A), imperfective markers in American English generally occur more frequently than in British English (see Fig. 6).

The narrative texts in FLOB/Frown are distributed mainly in the five fiction types plus humour. Yet imperfective markers in American English are more frequent in four of these six categories. Although imperfective markers in British English are slightly more frequent than in American English for category K, the difference is not significant (58 versus 57 instances per 10,000 words). According to Biber *et al.* (1999, p. 462), the progressive aspect in American English conversation is much more common than in British English conversation. The imperfective markers we counted in this case study are the progressive and the perfect progressive. As perfect progressive verb phrases are extremely rare in all categories (less than 0.5 per cent of all verb phrases according to Biber *et al.* (1999)), the influence of the perfect progressive on the overall frequency may, in effect, be discarded. Fiction and humour typically dwell

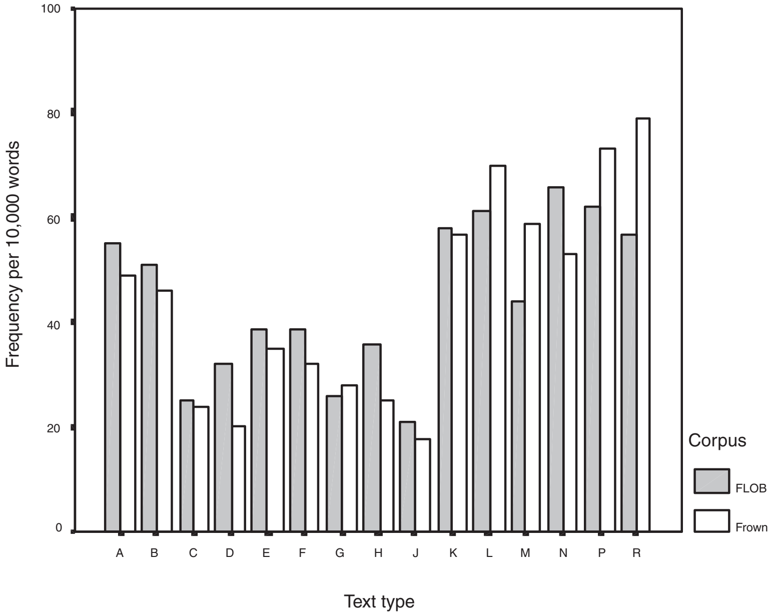

Fig. 6 Distribution of imperfective markers in FLOB/Frown.

on dialogue and thus bear a close resemblance to conversation (see Biber, 1988).[22] As such, it is hardly surprising that imperfective markers in American English are more common than those in British English.

6 Conclusion

This paper presented the newly released Lancaster Corpus of Mandarin Chinese, a Chinese match for the FLOB/Frown corpora. We first discussed the principal considerations of the corpus construction; namely, the corpus sampling, mark-up, and annotation. The case study presented in this paper has demonstrated that the corpus is a valuable resource for research into Chinese and, in combination with FLOB and/or Frown, for the contrastive study of Chinese and English. It is our hope that the release of LCMC will stimulate corpus-based research both into modern Chinese itself and into modern Chinese in contrast with English.

References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brinton, L. (1988). *The Development of English Aspectual System*. Cambridge: Cambridge University Press.

Burnard, L. and Todd, T. (2003). Xara: an XML aware tool for corpus searching. In Archer, D., Rayson, P., Wilson, A., and McEnery, A. (eds), *Proceedings of Corpus Linguistics 2003*, pp. 142–4. UCREL, Lancaster.

[22] Over 70 per cent of instances of quotations in FLOB and Frown are found in the five types of fiction and humour (76.78 per cent in FLOB and 70.36 per cent in Frown). In terms of the frequency of quoted words, 48.29 per cent in FLOB and 45.08 per cent in Frown are found in the six categories. We also found a positive correlation between the frequency of imperfective markers and the number of quoted words. In FLOB there are 1,161 markers in 232,376 words in quotations and 2,913 markers in 767,624 words not in quotations, with an LL score of 59.99 and a significance level less than 0.001. In Frown, there are 1,346 markers in 246,749 words in quotations and 2,511 markers in 753,251 words not in quotations, with an LL score of 199.83 and a significance level less than 0.001. Category N

Christensen, M. (1994). Variation in spoken and written Mandarin narrative discourse. Ph.D. thesis, Ohio State University, Columbus.

Comrie, B. (1976). *Aspect*. Cambridge: Cambridge University Press.

Dahl, Ö. (1999). Aspect: basic principles. In K. Brown and J. Miller (eds), *Concise Encyclopaedia of Grammatical Categories*. Oxford: Elsevier, pp. 30–7.

Du, W. (1999). Locative inversion and temporal aspect in Chinese. *Chicago Linguistic Society 35th Regional Meeting*, Chicago, IL, April 1999. Available at URL http://wwwvms.utexas.edu/~juliet/papers/cls35f.pdf. Accessed on May 20, 2003.

Feng, Z. (2001). Hybrid Approaches for Automatic Segmentation and Annotation of Chinese Text Corpus. *International Journal of Corpus Linguistics*, **6** (Special issue): 35–42. Available at http://nlplab.kaist.ac.kr/people/Feng/hybrid.doc. Accessed on May 21, 2003.

Francis, W. and Kuãera, H. (1964). *Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers*. Revised edn 1971; revised and augmented (with Henry Kuãera) 1979. Providence, RI: Department of Linguistics, Brown University.

Huang, C., Chen, K., Chen, F., and Chang, L. (1997). Segmentation standard for Chinese natural language processing. *Computational Linguistics and Chinese Language Processing*, **2**(2): 47–62.

Hundt, M., Sand, A., and Siemund, R. (1998). Manual of information to accompany the Freiburg–LOB Corpus of British English ('FLOB'). Online. Available at http://www.hit.uib.no/icame/flob/index.htm. Accessed on July 5, 2003.

Hunt, M., Sand, A., and Skandera, P. (1999). Manual of information to accompany the Freiburg–Brown Corpus of American English ('Frown'). Online. Available at http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM. Accessed on July 5, 2003.

Johansson, S., Leech, G., and Goodluck, H. (1978). Manual of information to accompany the Lancaster–Oslo/Bergen corpus of British English, for use with digital computers. Oslo: Department of English, University of Oslo.

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, and A. McEnery (eds), *Corpus Annotation*. Harlow: Longman.

McEnery, A. and Xiao, Z. (2002). Domains, text types, aspect marking and English–Chinese translation. *Journal of Languages in Contrast*, **2**(2): 211–29.

McEnery, A. and Xiao, Z. (forthcoming). Parallel and comparable corpora: what are they up to? In G. James (eds), *Corpus Linguistics and Translation Studies*. Clevedon: Multilingual Matters.

Sun, M., Shen, D., and Tsou, B. (1998). Chinese word segmentation without using lexicon and hand-crafted training data. In S. Kahane and A. Polguere (eds) *Proceedings of COLING-ACL '98* Vol. 2, pp. 1265–71. Montreal, Montreal University. Available at http://acl.ldc.upenn.edu/P/P98/P98-2206.pdf. Accessed on June 18, 2003.

Svalberg, A. and Chuchu, H. (1998). Are English and Malay worlds apart? *International Journal of Applied Linguistics*, **8**(1): 27–60.

Swen, B. and Yu, S. (1999). A graded approach for efficient resolution of Chinese word segmentation ambiguity. *Proceedings of NLPPRS 99*. Beijing. In K. Choi (ed), *Proceedings of NLPPRS 99*. Beijing, Tsinhua University.

(adventure and western fiction) is peculiar in that unlike other fiction types, British English registers a significantly greater proportion of imperfective markers over American English in this category. It might be speculated that adventure and western fiction may focus on action more than dialogue. However, the data in FLOB and Frown prove that this is not the case. In FLOB the average frequency of quotations in the six categories (five fiction types plus humour) is 198 instances per 10,000 words (12.78 per cent) whereas the normalized frequency of quotations in category N is 260 (16.68 per cent). In Frown the average frequency of the six categories is 202 instances per 10,000 words (11.73 per cent) whereas the normalized frequency of quotations in category N is 252 (14.62 per cent). The apparently anomalous nature of category N begs further investigation.

Wu, D. and Fung, P. (1994). Improving Chinese tokenization with linguistic filter on statistical lexical acquisition. *ANLP-94*. Stuttgart. In S. Schmid and S. Laderer (eds), *ANLP-94*. October 1994. Stuttgart, University of Stuttgart. Available at http://acl.ldc.upenn.edu/A/A94/A94-1030.pdf. Accessed on August 23, 2002.

Xia, F., Palmer, M., Okurowski, M., Kovarik, J., Huang, S., Kroch, T., and Marcus, M. (2000). Developing guidelines and ensuring consistency for Chinese text annotation. In M. Gravilidou, G. Caravannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer (eds), *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, LREC. Accessed online on June 26, 2003 at http://www.cis.upenn.edu/~chinese/.

Xiao, Z. and McEnery, A. (forthcoming). *Aspect in Chinese*. Amsterdam: John Benjamins.

Zhang, H. and Liu, Q. (2002). Model of Chinese words rough segmentation based on N-shortest-paths method. *Journal of Chinese Information Processing*, **16**(5): 1–7. Online. Available at http://www.nlp.org.cn. Accessed on June 12 2003.

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

Zhou, Q. and Yu, S. (1997). Annotating the contemporary Chinese corpus. *International Journal of Corpus Linguistics*, **2**(2): 239–58.