A tale of two C's: Comparing English varieties with Crown and CLOB (the 2009 Brown family corpora)

Jiajin Xu and Maocheng Liang
National Research Centre for Foreign Language Education,
Beijing Foreign Studies University

1 Background

Since 7 August 2010, the Corpus Linguistics team at Beijing Foreign Studies University had led the construction of two corpora of written contemporary English – Crown and CLOB. With the collaborative efforts of over 147 college English teachers and postgraduate students from over 107 universities across China, the corpus building project was nearly completed in December 2011, after which minor revisions had also been made till February 2012.

One of the underlying assumptions behind the development of the 2009 Brown family corpora is that corpora following a similar sampling frame at a certain interval can serve as good resources for describing language change. That is to say, Crown and CLOB can be viewed as good reference corpora for contrastive studies in terms of historical change (e.g. comparison among Brown, Frown, AmE06 and Crown) and regional variation (e.g. Brown vs LOB, Frown vs FLOB, BE06 vs AmE06, and Crown vs CLOB). Moreover, a bigger balanced corpus of contemporary written English can be created when the Brown corpora are merged.

The collection of English texts aimed at a balanced, in a modest sense, corpus of written contemporary English modelled after the sampling frame of the Brown Corpus. The name *Crown* is the fusion of the initial of *China* and the hind part of *Brown*, which means a new Brown family corpus collaboratively built by Chinese scholars. Likewise, *CLOB* gets its name with an initial *C*. We closely followed Kučera and Francis' sampling frame of the Brown Corpus.

The first standard release of Crown and CLOB was in June 2012. The corpora are available in raw texts (with and without metadata), PoS-tagged format (with CLAWS C7 tagset), and parsed format (with BFSU Stanford Parser). An online query interface of the corpora is also available for public access at http://124.193.83.252/cqp/.

2 Descriptive statistics of Crown and CLOB

2.1 The size of Crown and CLOB

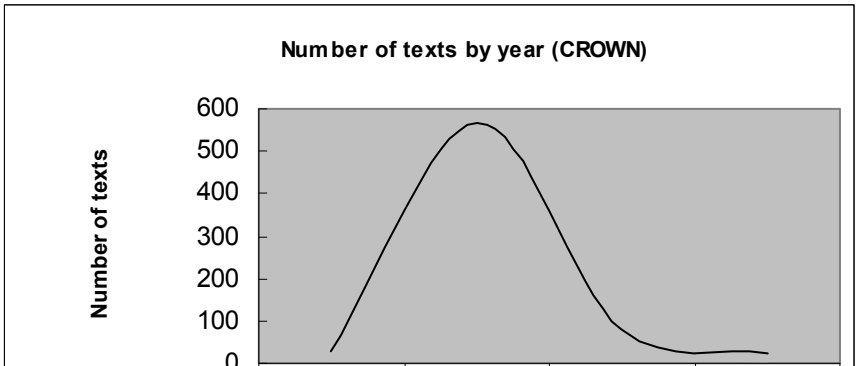
The total numbers of tokens as well as the tokens of four major genres of Crown and CLOB are provided in Table 1, together with those of Brown, LOB, Frown, and FLOB. (Note: four broad genres refer to newspaper texts (A–C, 88 texts), miscellaneous informative prose or general prose (D–H, 206 texts), learned and scientific English (J, 80 texts), and fiction (K–R, 126 texts).[1]

Table 1: The number of tokens of three generations of Brown family corpora[2]

	Genre	Sub-corpus tokens	Total tokens		Sub-corpus tokens	Total tokens
Brown 1961	Fiction	259,467	1,027,021	**LOB 1961**	258,722	1,018,785
	General prose	423,160			418,137	
	Learned	163,309			162,322	
	Press	181,085			179,604	
Frown 1992	Fiction	260,414	1,027,323	**FLOB 1991**	260,664	1,024,643
	General prose	421,933			419,990	
	Learned	163,228			163,286	
	Press	181,748			180,703	
Crown 2009	Fiction	259,250	1,026,226	**CLOB 2009**	259,484	1,023,466
	General prose	422,799			421,163	
	Learned	163,197			163,139	
	Press	180,980			179,680	

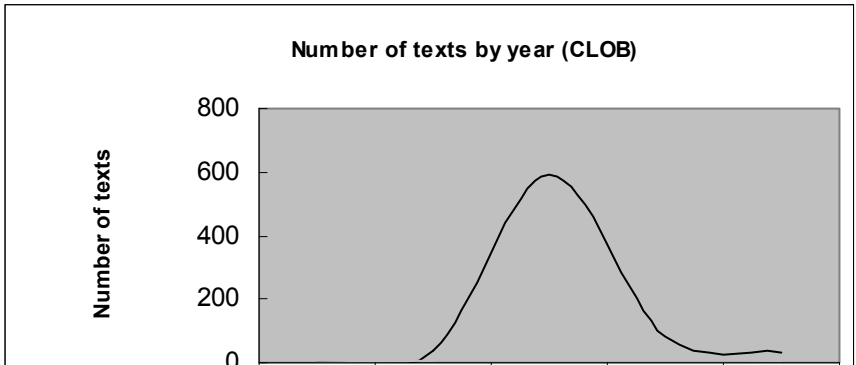
2.2 Years of publication of the texts in Crown and CLOB

The years of publication of the texts included in Crown and CLOB are summarised in Figures 1 and 2:



	2008	2009	2010	2011
Number of texts	30	566	79	25
Percentage (%)	4.29	80.86	11.29	3.57

Figure 1: Number of texts by year (Crown)



	2007	2008	2009	2010	2011
Number of texts	1	35	591	84	33
Percentage (%)	0.13	4.70	79.44	11.29	4.44

Figure 2: Number of texts by year (CLOB)

As seen in Figures 1 and 2, about 80 percent of Crown and CLOB texts were published in the year 2009, and 15 percent of the texts were published one year before or after 2009. Taken together, texts published between 2008 and 2010 account for over 95 percent of the total.

3 Data collection and processing

3.1 Sampling frame: The Brown Corpus model

The first standard release of Crown and CLOB contains one million words respectively, covering 15 categories (see Table 2) of texts published in 2009, or one year before or after 2009. In a few (less than 5%) cases, texts published in 2007 and 2011 were included.

Table 2: Text categories of Brown Corpus

	Text categories	No. of texts
A	Press: Reportage	44
B	Press: Editorial	27
C	Press: Reviews	17
D	Religion	17
E	Skill and hobbies	36
F	Popular lore	48
G	Belles-lettres	75
H	Miscellaneous: Government & house organs	30
J	Learned	80
K	Fiction: General	29
L	Fiction: Mystery	24
M	Fiction: Science	6
N	Fiction: Adventure	29
P	Fiction: Romance	29
R	Humour	9

3.2 The nationality of authors

Only the writings by the U.S. and the U.K. citizens or permanent residents were selected. The URLs, or the publishers, and the author profile pages were

recorded. In the case of multiple authors, the primary/first author should hold U.S. or U.K. citizenship or permanent residence.

3.3 The length of texts

The length of each text should be no shorter than 2,000 words. The word/token definition we used is [a-zA-Z0-9-]+. For some text types, one single article is not likely to be as long as 2,000 words. In that case, two or more articles of the same nature and on similar topics were gathered. If a text sample was going to be taken from a very long text, say, a novel, approximately 700 words from the beginning, the middle and the ending of the book respectively were selected.

3.4 Text cleaning and processing

The texts were cleaned and processed with regard to four aspects:

(1) To avoid, as much as we possibly could, mojibake or gibberish codes while processing the texts with corpus tools, some punctuation marks, mathematical symbols, non-English scripts (Greek, Nordic, German letters, etc.), and other special symbols were replaced with ASCII printable characters. For mathematical symbols, technical symbols, and Greek letters, the conversion table of HTML Symbol Entities Reference was consulted,[3] and replaced with ASCII printable characters accordingly. For instance, α, β and ϒ (in scientific discourse) were changed into 'α', 'β', and 'γ' respectively. All em dashes (—) and en dashes (–)[4] were replaced by two hyphens (--) with a white space before and after the two hyphens. ö, æ, ë, ê, and β (in German) in proper names were replaced by o, ae, e, e, and ss.

(2) The biographical information of all the authors was checked by the corpus linguistics team members at Beijing Foreign Studies University. Wikipedia and authors' personal and/or institutional web pages were searched through to make sure that the authors were born and got their education in the United States or the United Kingdom.

(3) Advertisements, tables, figures, sideline links etc. were deleted.

(4) Partial or whole text duplicates were checked with Wcopyfind.exe.[5] Duplicates found were removed and new texts were added in place. In such cases, we deliberately looked for some texts which were published in 2011; thus duplicate texts would be less likely.

3.5 Format of texts and file names

Plain text files (in ASCII encoding) are named A01A.txt, A01B.txt, A02A.txt, A02B.txt, B01A.txt, B01B.txt, etc. (the fourth character A for American, and B for British). There are 700 texts in Crown corpus and 744 texts in CLOB corpus

instead of 500 texts of 2000 words each in the Brown corpus model, because short texts are saved separately in Crown and CLOB. For example, A03 text in the Brown model consists of four short texts in Crown corpus and they are named A03AA.txt (449 words), A03AB.txt (449 words), A03AC.txt (737 words) and A03AD.txt (433 words). Short texts are especially frequent in news reports. The strength of saving the short texts separately is that each text represents itself. It is quite easy to merge the related texts as a 2,000-word one, if this is necessary. However, once the short texts have been put together in one single 2,000-word file, without explicit section markers, it is almost impossible to save them as individual texts as they originally were.

3.6 No reprinted works

The work had to be first published in 2009 (± 1 year). Works reprinted in 2009 (in other words, they were originally published earlier than 2009) were not included. For instance, reprinted detective novels of *Sherlock Holmes* were not considered.

3.7 Metadata mark-up

Metadata were marked up as follows. Metadata Encoder 2[6] was used to add the following bibliographic information (see Table 3) to the texts:

Table 3: Metadata template for Crown and CLOB

Author	e.g. John Smith
Country	e.g. UK/US
Publication year	e.g. 2009
Publisher	e.g. New York Times
URL(s)	e.g. http://

4 Copyright and dissemination

The copyright of all the texts in Crown and CLOB belongs to the original copyright holders. We plan on a GNU[7] distribution of the texts, given that they are not used for commercial purposes in whatever manner. Everyone who makes a positive contribution to the data collection will get a copy of the Crown and CLOB corpora for their personal study. Other users can access Crown and CLOB at our CQPweb site: http://124.193.83.252/cqp/ (ID: test; password: test).

5 Case studies based on Crown and CLOB

The two new members of the Brown family, along with other Brown family members, can be a good dataset for studies on English language, for instance, for the contrastive study on lexis, phrases, grammar patterns, stylistic features favoured by the two varieties of English, synchronically and diachronically. The corpus research group at Beijing Foreign Studies University has investigated the lexical, grammatical, and stylistic aspects of the enlarged Brown family corpora.

Chen (2012) compares the 's-genitives' in the American and British press reportage. She concludes that: 1) American English and British English share similar 's-genitive' developmental patterns, i.e. the use of 's-genitive' increases from 1961 to 1991, and becomes stable after 1991; 2) Possessor animacy counts as one of the most powerful factors around the growing and declining popularity of 's-genitive' in contemporary modern English.

Ji (2012) conducted a quantitative and qualitative study on the variation of modals and semi-modals between American English and British English, and also along the timeline over nearly half a century from 1961–2009. Both tokens and modal meanings are examined. Extraction and comparison of modal occurrences are done with concordancers; modal meanings, such as obligation, permission, necessity, volition, possibility, however, are manually annotated. Data from the structural and semantic analyses show that possibility comes as the predominant meaning on top of all the other modal meanings, and it is used more in British English than American English, more in fiction than other genres, and sees a stable increase over half a century.

Zhang and Xu (2013) explore the diachronic changes of the distribution patterns and semantic shift of the English present progressive. The results show that, overall, frequencies of the English present progressive increase significantly since the 1960s, and the growth is more conspicuous at the beginning of the 21st century; as to genre and voice, the progressive instances exhibit some variability in addition to the general growing tendency; the frequencies of the non-present progressive uses (the futurate and attitudinal meanings) of the present progressive also increase dramatically. All these findings imply that the English present progressive in written English has a distinct tendency to be colloquialised, and in the meantime, the increasing uses of semantically shifted 'be V-ing' patterns seem to indicate that the usage of the present progressive is becoming more diversified and subjectified, suggesting that it is in a process of grammaticalisation, which is in concordance with the overall evolving tendency of English grammar.

Lu (2012) makes an in-depth corpus-based exploration of discourse representation (Leech and Short 1981, 2007; Semino and Short 2004), or speech, writing and thought presentation (SW&TP) in English romantic fiction based on six Brown corpora (Brown, Frown, Crown, LOB, FLOB, and CLOB). Statistical results show that both American and British English prefer to employ thought representation, and that there is more 'telling' than 'showing' in American romantic fiction than in British romantic fiction. A general trend of increasing direct thought, free direct thought and free indirect thought is observed over time, which brings about more dramatic effect and vividness.

6 Concluding remark

The two Brown family corpora have been made publicly available among Chinese researchers for some time, which has encouraged quite a number of comparative studies of English lexico-grammar. Apart from the studies of English per se, more often than not, Crown and CLOB, alongside other Brown corpora, serve as reference corpora of present-day native English in contrastive interlanguage analysis, and corpus resources for Chinese-English contrastive studies. We are now finalising our 2009 Chinese Brown family corpus (temporarily called CC2009) developed at Beijing Foreign Studies University. We hope that the corpus building work is but the beginning of more Brown family corpus-related corpus construction and research in the far east.

Acknowledgements

Our greatest debt is to all the contributors who helped collect texts for the corpus project. A special thank then goes to Jie Ji for coordinating the BFSU Corpus Research Group to check the text format, encoding, relevance of text content to their genre categories, and sources of the texts with extreme care, and all other kinds of proofreading work. We would also like to thank Gong Chen, Jie Ji, Guobing Liu, Lei Liu, Xia Liu, Gong Peng, and Baicheng Zhang for their valuable comments on the earlier drafts of the article. The compilation of Crown and CLOB is supported by the National Social Sciences Foundation Project "A Bilingual Corpus Based Study on the English Translation of Chinese Complex Verbal Constructions" (Ref.: 12CYY060) and "Program for New Century Excellent Talents in University" (Ref.: NCET-12-0790). We are also grateful to the participants at the *Corpus Linguistics in China 2011* conference, as well as the audience at the *Corpus Technologies and Applied Linguistics* (2012) conference, for their critical comments.

Notes

1. http://icame.uib.no/archives/No_5_ICAME_News_index.pdf.
2. The token definition applied was all alphanumeric character strings plus hyphens, and the regular expression for the definition is [a-zA-Z0-9-]+.
3. http://www.w3schools.com/TAGS/ref_symbols.asp.
4. http://en.wikipedia.org/wiki/Dash.
5. http://plagiarism.bloomfieldmedia.com/z-wordpress/software/wcopyfind/.
6. http://ishare.iask.sina.com.cn/f/33571440.html.
7. Refer to more information about the GNU General Public License at http://www.gnu.org/licenses/glp.html.

References

Chen, Gong. 2012. A diachronic study on English '*s*-genitives': Evidences from Brown family corpora (1961–2009). Paper presented at *Corpus Technologies and Applied Linguistics* conference. 29 June 2012, Xi'an Jiao Tong Liverpool University, Suzhou, China.

Ji, Jie. 2012. Variational study on modal meaning – a corpus-based research of Brown family. Paper presented at *Corpus Technologies and Applied Linguistics* conference. 29 June 2012, Xi'an Jiao Tong Liverpool University, Suzhou, China.

Leech, Geoffrey and Mick Short. 1981. *Style in fiction*. London: Longman.

Leech, Geoffrey and Mick Short. 2007. *Style in fiction*. 2nd edition. London: Pearson Education.

Lu, Lu. 2012. The speech, writing and thought presentation in English romantic fiction: A diachronic and cross-cultural study. Paper presented at *Corpus Technologies and Applied Linguistics* conference. 29 June 2012, Xi'an Jiao Tong Liverpool University, Suzhou, China.

Semino, Elena and Mick Short. 2004. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

Xu, Jiajin. 2011. A tale of two C's: Crown and CLOB. Paper presented at the *Corpus Linguistics in China* Symposium. 19 November 2011, Beijing Foreign Studies University, Beijing, China.

Xu, Jiajin. 2012. Comparing English varieties with Crown and CLOB. Paper presented at *Corpus Technologies and Applied Linguistics* conference. 29 June 2012, Xi'an Jiao Tong Liverpool University, Suzhou, China.

Zhang, Baicheng and Jiajin Xu. 2013. A Brown family corpora based diachronic study on the English present progressive. *Foreign Language Education* 34: 42–47.