The Lancaster Corpus of Mandarin Chinese: A Corpus for Monolingual and Contrastive Language Study

Anthony McEnery, Zhonghua Xiao

Department of Linguistics, Lancaster University, Lancaster, LA1 4YT, UK
{a.mcenery, z.xiao}@lancaster.ac.uk

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. LCMC is a one-million-word balanced corpus of written Mandarin Chinese. The corpus contains five hundred 2,000-word samples of written Chinese texts sampled from fifteen text categories published in Mainland China around 1991, totalling one million words. LCMC is XML-compliant and conforms to CES, with each document containing a corpus header giving general information about the corpus and a body of text. The corpus is segmented and POS tagged with a tagging precision rate of over 98%. The corpus is a useful resource for research into modern Chinese as well as the cross-linguistic contrast between English and Chinese.

1. Introduction

The Lancaster Corpus of Mandarin Chinese is a one-million-word balanced corpus of written Mandarin Chinese. The corpus was designed as a Chinese match for the FLOB (Hundt, Sand & Siemund, 1998) and Frown (Hunt, Sand & Skandera, 1999) corpora of British and American English and was created as part of a research project funded by the UK ESRC.[1] This paper first reviews the publicly available corpus resources for Mandarin Chinese. Following this we will outline our corpus design criteria and discuss the annotations undertaken on the corpus. Finally we will introduce some markup-aware tools to facilitate the exploration of the corpus.

2. What Corpora Are Available for Chinese?

As a result of the rapid development in Chinese corpus linguistics over the past decade, a number of publicly available corpora of Mandarin Chinese have been reported recently. One of the earliest of such corpora is the Academia Sinica Balanced Corpus of Modern Chinese. The corpus contains five million words of Mandarin Chinese as used in Taiwan.[2] The PH corpus contains around two million words of newswire texts published by the Xinhua News Agency during 1990 – 1991. The PFR corpus released by Peking University consists of one month's (January 1998) newspaper material published by the People's Daily.[3] The LDC has also released a number of corpora of news texts and telephone conversations in Chinese (e.g. TREC, Gigaword and Callhome Mandarin). The LIVAC (Linguistic Variation in Chinese Speech Communities) corpus, created by City University of Hong Kong, is near completion.[4] The corpus contains texts from representative Chinese newspapers and electronic media of Hong Kong, Taiwan, Beijing, Shanghai, Macau and Singapore and the collections of material from the diverse communities are synchronized. Other corpora are also planned – for example, a balanced corpus of Chinese, as reported in Zhou & Yu (1997), is under construction. A spoken Chinese corpus of situated discourse is (SCCSD BJ-500) being built under the auspices of the Chinese Academy of Social Science (see Gu 2002).

Corpus	POS	Bal.	Channel	Variety	Contr.
LCMC	Yes	Yes	Written	Mainland	E – C
Sinica	Yes	Yes	Mixed	Taiwan	No
PH	No	No	Written	Mainland	No
PFR	Yes	No	Written	Mainland	No
LIVAV	No	No	Written	Mixed	C – C
SCCSD	No	Yes	Spoken	Mainland	No
TREC	No	No	Written	Mainland	No
Gigaword	No	No	Written	Mainland	No
Callhome	No	?	Spoken	Mixed	No

Table 1: A comparison of available Chinese corpora

Table 1 compares the publicly available corpora of Mandarin Chinese. It is clear that while there are some Chinese corpus resources available, most of the written or mixed channel corpora are not balanced. The Sinica corpus is balanced. Yet 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.[5] As such, the Sinica corpus does not represent modern Mandarin Chinese as written in Mainland China. A further problem is that most of them are not suitable for cross-linguistic contrast. An exception is the LIVAC corpus, designed for comparative study. But it can only be used to explore regional variation in Chinese.

We built the LCMC corpus in response to the general lack of publicly available balanced corpora of Chinese as used in Mainland China.[6] As the corpus was designed principally with contrastive research in mind, it is a valuable resource for research into Chinese as well as a reliable basis for contrastive study of English and Chinese.

[1] We thank the UK Economic and Social Research Council for funding our project (Grant reference RES-000-220135).

[2] See http://www.sinica.edu.tw/SinicaCorpus for details.

[3] See http://www.ling.lancs.ac.uk/corplang for a brief introduction to the PH and PFR corpora.

[4] See http://www.livac.org/ for details.

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

[6] The corpus is distributed free of charge for use in non-profit making research. See http://www.ling.lancs.ac.uk/corplang/lcmc for instructions for accessing the corpus.

3. Design Criteria of LCMC

This section outlines the design criteria of the LCMC corpus including sampling frame and text collection, followed by a discussion of the markup and annotation schemes.

3.1 Sample frame and text collection

As the LCMC corpus was originally created for use on our research project *Contrasting tense and aspect in English and Chinese*, 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. After reviewing the available English corpora, we decided to create a match for FLOB, 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. Hence by building a match for FLOB we enabled a contrast of Chinese with the two major varieties of English.

Code	Text category	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/hobbies	38	7.6%
F	Popular lore	44	8.8%
G	Biographies/essays	77	15.4%
H	Miscellaneous	30	6%
J	Science	80	16%
K	General fiction	29	5.8%
L	Mystery/detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Western/adventure fiction	29	5.8%
P	Romantic fiction	29	5.8%
R	Humor	9	1.8%
Total		500	100%

Table 2: Text types covered in the FLOB corpus

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 2. 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. Firstly, there is virtually no western fiction written in Chinese for a Mainland Chinese audience. Secondly, martial arts fiction is broadly a type of adventure fiction and as such can reasonably 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. While 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 also makes 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 (around 87% of texts in the corpus occur ±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 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 and L1 quality guaranteed.

While 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 documents 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% 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.

While the digital library has a very large collection of books, it does not provide complete newspapers, providing 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 vs. newspaper articles) or medium (e.g. newspaper texts vs. 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 (cited from the PH corpus). Considering that this is the most important and representative news provider in China, roughly analogous to the Associated Press in the US/UK, we believe that the high proportion of material taken from the Xinhua News Agency is justified.

3.2 Corpus markup and annotation

LCMC is XML-conformant. Each text type in the corpus is stored in one file, which consists of a CES

header giving general information about the corpus, and the body of corpus text. The body is encoded with five main features, as shown in Table 3. These details are useful when using an XML-aware concordancer such as *Xara* (Burnard & Todd, 2003). 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 (see section 4).

Level	Code	Gloss	Attribute	Value
1	text	Text type	TYPE	As per Table 2 *Text Category*
			ID	As per Table 2 *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		
	gap	Omission	---	---

Table 3: XML elements in corpus text

While the original corpus texts were encoded in GB2312, we decided to convert the encoding to Unicode (UTF-8) for two 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* and *WordSmith Tools* version 4.0 (Scott 2003). In order 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. While also encoded using UTF-8, the Pinyin version is more compatible with older operating and concordance systems. This is also of assistance to users who can read Romanized Chinese but not Chinese characters.

We undertook two forms of corpus annotation on the LCMC corpus: word segmentation and part-of-speech annotation (the LCMC tagset consists of 50 POS tags. See the corpus website for details). 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. 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 (cf. Zhang, Liu, Zhang & Cheng, 2002). The rough segmentation module of the system is based on the *n*-shortest paths method (Zhang & Liu, 2002). The model, based on 2-shortest-paths, achieves a precision rate of 97.58%, with a recall rate as high as 99.94% (Zhang & Liu, 2002). In addition the average number of segmentation candidates is reduced by 64 times compared to 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% for tagging, with a recall rate of over 90% for unknown words and 98% for Chinese person names (Zhang & Liu, 2002).

However, the POS system is in part under-specified, especially in the crucial area of aspect marking. 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. 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 processed texts were hand-checked and corrected. The post-editing improved the annotation precision to over 98%.[7]

4. Corpus Exploration Tools

As the LCMC is marked up in XML, non-markup-aware concordancers will not allow users to easily exploit these corpora fully. Two Unicode-compliant markup-aware corpus tools that are available, *Xara* and *WordSmith* version 4.

Using *WordSmith* 4 to explore the two corpora is quite straightforward, though the LCMC Corpus needs to be converted from utf-8 to utf-16 first using a built-in utility of *WordSmith*. *Xara* is more powerful in that it allows users to build very complex queries, yet it is accordingly more difficult to use. The program is an XML-compliant extension of *SARA* (SGML-aware Retrieval Application) originally developed for the British National Corpus (cf. Aston and Burnard 1998). It can be used for both the local and remote access of a corpus.

In LCMC, the most important XML elements are *text* (text category), *file* (sample file), *s* (sentence) and *w* (word token).[8] The *text* element can be used to compare different genres while the *file* and *s* elements indicate the location of a concordance to provide a reference back in the corpus. Now suppose we want to extract all instances of the verb-final *-le* (tagged as *u*) immediately followed (the link type defined as *Next*) by a noun (tagged as *n*) in sentence number 0010 in all of the 500 sample files in the 15 text categories. This complicated query can be made using 'Query builder' of *Xara*. First, define the *scope node* (the left node in Query builder that indicates the context to search in) as '0010' using the *s* element (Fig. 1). In the *query node* (the right node in Query builder), select *AddKey* (POS) to define the first part of the query as *-le*

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

[8] Following the BNC style, punctuations and symbols in LCMC are tagged separately from word tokens using the *c* element.

and select the POS tag *u*, and the second part as *Any* and select the POS tag *n*. Then define the *link type* as *Next* (Fig. 2). The search result is shown in Fig. 3. The upper part of the concordance window gives the query text (Select *Query – Query text* from the main menu to display the query text) while the lower window displays the concordances. The status bar of the concordance window shows the name of the corpus, the current position of the pointer/mouse (i.e. concordance number 1), the total number of concordances (i.e. 25), the number of files in which the query is matched (10), the file name where the current concordance occurs (i.e. LCMC_A), and the file/sentence number for the current concordance (i.e. File A04 and sentence number sn0010). As we have searched in sentence number 0010 (in 500 sample files), this should be the sentence number for all of the concordances.

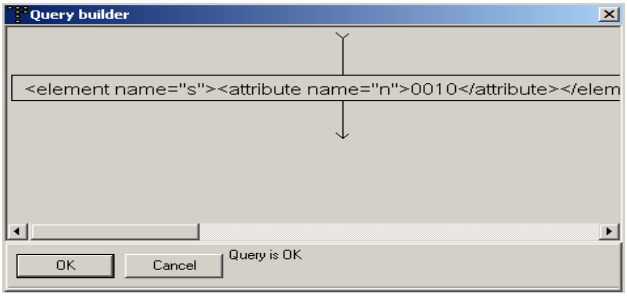

Figure 1: Defining the scope node

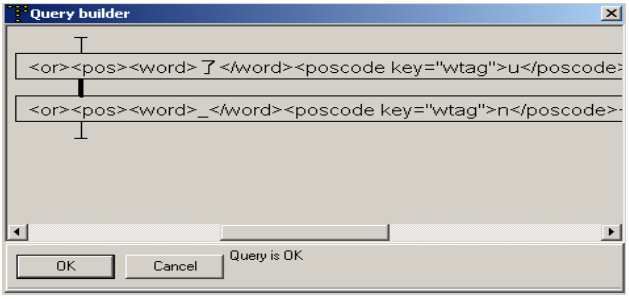

Figure 2: Defining the query node

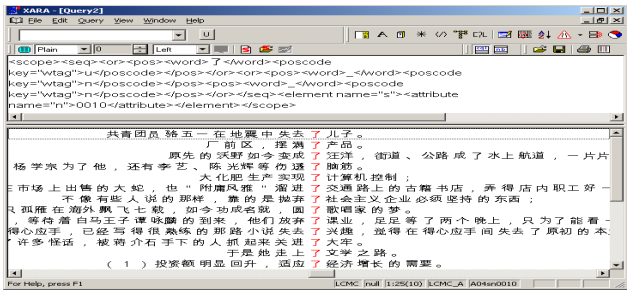

Figure 3: The concordance window

By comparison to many other corpus tools, one advantage of *Xara* is that it displays complete sentences while also centering the search query. Users are also given options to display concordances in the *page* (giving more context) or *line* mode (i.e. KWIC, as shown in Fig. 3), in XML or plain text. Additionally, users can define their own style sheet to display selected XML elements. *Xara* can also compute significant collocates automatically using a statistic selected from those available by the user.

While LCMC can be explored most efficiently with *Xara*, we have also developed a web-based concordancer (*WebConc*) for use with LCMC, which is more user-friendly than *Xara*. *WebConc* allows users to search in the standard character version or the Romanized Pinyin version of the LCMC corpus using a token, POS tag or their combination. Users can also select text categories for inclusion in their search. The search result can be displayed in the sentence or KWIC mode (both displaying complete sentences), in XML or plain text. The concordancer also gives a summary of the query, including the query text, the date the corpus is accessed, raw and normalized (per million words) frequencies in each text category, and the total frequency in the text categories users have selected. The *WebConc* can be accessed at the LCMC website.

5. Conclusion

This paper presented the newly released Lancaster Corpus of Mandarin Chinese, a Chinese match for the FLOB and Frown corpora of British and American English. We first reviewed the publicly available corpora of Mandarin Chinese. Following this we outlined the design criteria of LCMC and discussed the annotations undertaken on the corpus. Finally a number of markup-aware tools were introduced to facilitate corpus exploration. 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

Aston, G. & Burnard, L. (1998). *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Burnard, L. & Todd, T. (2003). Xara: an XML aware tool for corpus searching. In *Proceedings of Corpus Linguistics 2003* (pp. 142-4). Lancaster.

Christensen, M. (1994). *Variation in Spoken and Written Mandarin Narrative Discourse*. Ph.D. thesis, Ohio State University, Columbus.

Gu, Y. (2002). Towards an understanding of workplace discourse. In C. Candlin (ed) *Research and Practice in Professional Discourse* (pp. 137-86). Hong Kong: City University of Hong Kong Press.

Hundt, M., Sand, A. & Siemund, R. (1998). *Manual of information to accompany the Freiburg - LOB Corpus of British English ('FLOB')*.

Hunt, M., Sand, A. & Skandera, P. (1999). *Manual of information to accompany the Freiburg - Brown Corpus of American English ('Frown')*.

Scott, M. (2003). *WordSmith Tools Manual*.

Xiao, Z. & McEnery, A. (Forthcoming). *Aspect in Chinese*. John Benjamins, Amsterdam.

Zhang, H. & Liu, Q. (2002). Model of Chinese words rough segmentation based on N-shortest-paths method. *Journal of Chinese Information Processing*, 16(5), 1-7.

Zhang, H., Liu, Q., Zhang, H. & Cheng, X. (2002). Automatic recognition of Chinese unknown words based on role tagging. In *Proceedings of the 1st SIGHAN Workshop, COLING 2002* (pp. 71-7). Taipei.

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