DOI: 10.13140/RG.2.2.28488.39680

The UCLA Corpus of Written Chinese

Hongyin Tao Richard Xiao

UCLA, USA Lancaster University, UK

The UCLA Corpus of Written Chinese is designed as a Chinese counterpart for the FLOB and Frown corpora of British and American English for contrastive research, as well as a recent update of the Lancaster Corpus of Mandarin Chinese (LCMC) for diachronic studies of possible changes in written Chinese over the past decade. Since this period is of special significance because of the impact of the Internet on language, especially on Chinese, the corpus is an excellent complement to LCMC.

The samples in the corpus are all collected from written modern Chinese available from the internet, during the periods of 2000-2005 and 2005-2012, though some texts may have been converted from paper-based publications in earlier years. File types are matched as closely as possible to the Brown corpus model, with some variations (e.g. adventure fictions) to accommodate Chinese characteristics, while the proportions for different text categories may vary from the English counterparts and LCMC. Presently the genres covered, along with their sample sizes and number of files are shown in the table below. The differences in tokens between the two phrases of the project are also indicated in the table.

Code	Genre	Tokens (2000-2005)	Tokens (2005-2012)	Number of Files
A	Press: reportage	84302	3,880	41
B	Press: editorials	25155	28,213	64
C	Press: reviews	32223	2,261	36
D	Religion	5885	28,885	38
E	Skills, trades and hobbies	8925	39,415	58
F	Popular lore	24854	36,256	60
G	Essays and biographies	71169	78,884	68
H	Misc. (reports and official documents)	65705	0	34
J	Academic prose	27652	103,174	94
K	General fiction	40999	17,805	30
L	Mystery and detective stories	85317	0	27
M	Science fiction	60378	0	23
N	Adventure stories	55253	2,341	41
P	Romantic fiction	66849	0	32
R	Humour	32968	0	42
(No. of tokens		**687634**	**341114)**	
(No. of types		**32212**	**26637)**	**Total No of Files 688**

The corpus is Unicode and XML-compliant. Each corpus file is composed of a corpus header and a text body. The header gives general information of a corpus file. In the body part, paragraphs, sentences and tokens are marked up, with each sentence numbered and each token annotated for part of speech. The corpus was tagged by using both the Peking University Chinese POS tagset and the ICTCLAS tagger by 中国科学院.

The UCLA Chinese Corpus is a product of the joint effort of Professor Hongyin Tao (University of California Los Angeles) and the late Dr. Richard Xiao (UCREL of Lancaster University). Funding for this project was provided to Hongyin Tao by the UCLA Academic Senate during the academic years 2003-2012, while Richard Xiao was supported by the UK Economic and Social Research Council (Award Reference RES-000-23-0553). We are also obliged to Iris Li, Haiyong Liu, Hui Zhang (for the first phase of the project), and Danjie Su (for the second phase of the project) for their assistance in data collection, and to Huaqing Hong (Nanyang Technological University, Singapore) and Jiajin Xu (Beijing Foreign Studies University) for enabling web access to the corpus.

The UCLA Chinese Corpus can be cited as:

Tao, Hongyin and Richard Xiao (2007/2020). *The UCLA Corpus of Written Chinese*. Los Angeles, USA and UCREL, Lancaster, UK.