The Lancaster Corpus of Mandarin Chinese

- 1 million word balanced corpus of written Mandarin Chinese
- 500 2000-word samples from 15 text categories
- Chinese match for FLOB (British English) and Frown (American English) corpora
- Publicly available, downloadable as a .zip file
  - http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
- XML compliant
- Segmented and POS tagged – tagging precision rate >98%
- Produced by Anthony McEnery and Richard Xiao
- Contrastive language study

Creation: Sample Frame and Text Collection

- Follow FLOB's sampling frame of 500 2000-word samples
- 15 text categories from 1991-1992, total 1,000,000 words (with a few variations)
  - E.g. used martial arts fiction to replace western and adventure fiction
  - Samples ±2 years of 1991 for popular lore, religion, skills/ trades/hobbies, and humor
- Written Mandarin Chinese texts published in Mainland China
- SSReader Digital Library in China
  - Large collection of books
  - Newspaper and newswire stories

Markup and Annotation

- XML-conformant
- Each text type stored in one file
  - Consist of CES header
- Original corpus text encoded in GB2312 → converted to Unicode UTF-8
- Word segmentation and part-of-speech tagging
  - 50 POS tags
  - Chinese Lexical Analysis System
  - Post-editing to increase accuracy

Usage

- Used to contrast aspect marking between
  - American English, Mandarin Chinese, and British English
- Test the claim: aspect markers occur significantly more frequently in narrative texts than in expository texts
  - According to LCMC, FLOB, and Frown, this is true.
  - i.e. higher frequency of aspect markers in narrative texts over expository texts is a common feature of Chinese, and the two varieties of English
  - Some differences were noted across various categories within the two types of texts

References

- McEnery, A. and Xiao, Z. (2003). Aspect Marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for Contrastive Language Study. *Literary and Linguistic Computing*, 18(4):361–378. DOI: 10/1093/llc/18.4.361
- McEnery, A. and Xiao, Z. (2004). The Lancaster Corpus of Mandarin Chinese: A Corpus for Monolingual and Contrastive Language Study. *Religion*, 17:3–4. (http://www.lancs.ac.uk/~xiaoz/papers/231.pdf)

ISLRN

Title	LCMC
Full Title	The Lancaster Corpus of Mandarin Chinese
Resource Type	Text
Source/URL	http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
Format/MIME Type	Text/XML
Size/Duration	1,000,000 words
Access Medium	Online
Description	The Lancaster Corpus of Mandarin Chinese (LCMC) addresses an increasing need within the research community for a publicly available balanced corpus of Mandarin Chinses. LCMC has been constructed as part of a research project undertaken by the Linguistics Department of Lancaster University The corpus is designed as a Chinese match of the Freiburg-LOB Corpus of British English (FLOB), and, as such, will provide a valuable resource for contrastive studies between English and Chinese as well as a sound basis for monolingual investigations of Chinese.
Version	?
Media Type	Text
Language	Mandarin Chinese
Resource Creator	Anthony McEnery, and Richard Xiao
Distributor	European Language Resources Association, and Oxford Text Archive
Rights Holder	Richard Xiao