The Lancaster Corpus of Mandarin Chinese (LCMC): Collocations in English and Chinese

Tony McEnery

Richard Xiao

LANCASTER UNIVERSITY

Celebrating 40 Years of Excellence

The LCMC Corpus: Aims

- Built for the ESRC project *Contrasting tense and aspect in English and Chinese* (Grant Ref. RES-000-220135)
  - See http://www.regard.ac.uk
- A Chinese match for FLOB/Frown for BrE/AmE
- A publicly available balanced corpus of Mandarin Chinese
- Distributed free of charge for use in non-profit-making research

LCMC: Profile

- One million words
- Standard character and Romanized Pinyin versions
- 1990-1993 (ca. 87% of samples from 1991-1992)
- 15 text categories
- 500 text samples
- Major text provider: SSReader Digital Library, China
- Unicode (UTF-8)
- XML-conformant mark-up
- Marked for paragraphs and sentences
- POS-tagged (precision rate 98%+)

Major Chinese corpus resources (1)

- Sinica Corpus of Mandarin Chinese
  - 5 million words of Mandarin as used in Taiwan
  - http://www.sinica.edu.tw/SinicaCorpus
- PH corpus
  - Ca. 2 million words of newswire text (1990-1991)
  - Available at ftp://ftp.cogsci.ed.ac.uk/pub/chinese
  - POS version available at http://www.ling.lancs.ac.uk/corplang/
- PFR People's Daily Corpus
  - Newspaper text from People's Daily 1998
  - Sample (01/98) available at http://icl.pku.edu.cn/Introduction/corpustagging.htm
  - Searchable at http://www.ling.lancs.ac.uk/corplang/

Major Chinese corpus resources (2)

- Linguistic Variation in Chinese Speech Communities
  - Text from newspapers and electronic media in six Chinese speech communities
  - http://www.livac.org/
- Spoken Chinese Corpus of Situated Discourse (SCCSD)
  - See 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.
- Three Mandarin corpora released by LDC
  - TREC, Gigaword and Callhome
  - See the LDC catalogue

Chinese corpora: A comparison

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
LIVAC	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

LCMC: Sampling frame

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	Adventure and martial arts fiction	29	5.8%
P	Romantic fiction	29	5.8%
R	Humour	9	1.8%
Total		500	100%

LCMC: Markup (XML)

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 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	---	---

LCMC: Annotations

- Word segmentation
- POS tagging
  - Applying the Peking University tagset
    - 26 Level 1 POS tags
    - 50 Level 2 POS tags
  - POS tagger (ICT Chinese Lexical Analyzing System)
    - Developed by the Institute of Computing Technologies, the Chinese Academy of Sciences
  - Automatic tagging with a precision rate of 97.16%
  - Post-editing improved the precision to over 98%

LCMC Level 1 POS tags

- a. adjective
- b. non-predicative adj.
- c. conjunction
- d. adverb
- e. interjection
- f. directional locality
- g. morpheme
- h. prefix
- i. Idiom
- j. abbreviation
- k. suffix
- l. fixed expression
- m. numeral
- n. noun
- o. onomatopoeia
- p. preposition
- q. classifier
- r. pronoun
- s. space word
- t. time word
- u. auxiliary
- v. verb
- w. punctuation/symbol
- x. unclassified item
- y. modal particle
- z. descriptive

LCMC: corpus exploration tools

- Unicode-compliant, XML-aware corpus tools
  - *WebConc* designed for use with LCMC
    - http://www.ling.lancs.ac.uk/corplang/cgi-bin/conc.pl
  - *Xaira* (XML-aware *Sara*)
    - *Sara:* SGML-aware Retrieval Application
      - Originally developed for use with the British National Corpus (BNC)
    - Known as *Xara* before beta version 1.06
    - Documentation available at http://www.oucs.ox.ac.uk/rts/xaira/
    - A tutorial available at the LCMC website
  - The *WordSmith Tools* version 4
    - Beta version available
      - http://www.lexically.net/wordsmith/version4/index.htm

Collocation and Semantic Prosody: the cross linguistic perspective

- Most studies of both phenomena to date conducted on English language corpora
- Cross linguistic studies rare
- Exception – Berber-Sardinha and Tognini-Bonelli
- But what of two genetically distinct languages?

Our Goal

- Explore translation equivalents in Chinese of words/expressions on which research on semantic prosody had been undertaken in English
- Are semantic prosodies peculiar to English? Are the presence or absence of collocations in a language determined, to a degree, by the language's dependence on word order restrictions?

• Author	• Negative prosody	• Positive prosody
• Sinclair (1987, 1990, 1991)	• break out • happen • set in	
• Louw (1993, 2000)	• bent on • build up of • end up *verb*ing • get oneself *verb*ed • a recipe for	• build up a
• Stubbs (1995, 1996, 2001b, 2001c)	• accost • cause • fan the flame • signs of • underage • teenager(s)	• provide • career
• Moon (1998)	• fan the flame	
• Partington (1998)	• commit • peddle/peddler • dealings	
• Hunston (2002)	• sit through	

Our Study

- the *consequence* group
- the *cause* group
- *commit*
- *price(s)*

COMMIT

(1)

- (a.) Indeed, William Zinsser describes how aspiring writers set out to commit an act of literature, an impossible task. (Frown: R)
- (b.) <…> because I don’t want to commit an overt, non-rational act and I don’t want to lose control of self <…> (Frown: D)
- (c.) But Rickman endows his character with such an intense inner life that you suspect that, at any moment, he might be about to commit some monstrous act of violence. (FLOB: A)
- (d.) Hitler understandably regarded people who could commit such acts against Britain as his natural allies. (Frown: J)

(2)

- a. The sole function of the other one, as far as we could tell, was to apologize to us on behalf of the hotel for having committed this monumentally embarrassing and totally unforgivable blunder. (Frown: R)
- b. At least Rovers battled until the bitter end and Castleford did their best to help, committing a series of handling errors while watching prop Keith England sin-binned after he hit-out at home sub Wayne Jackson at a play-the-ball. (FLOB: A)

- *fan* (犯)

(3)

- (a.) *ta xiukui de di-xia tou, nene de shuo, "youpai… youpai jiu shi <u>fan</u>-le cuowu de ren."* (LCMC: K)
- 'With his head lowered in shame, he said in a faltering voice, "Rightists…Rightists are those who have made a mistake."'
- (b.) *ta you tuo ren daixin gei tewu zuzhi, <u>fan</u> you panguo toudie zui (LCMC: A)*

'He also committed the crime of treason and defection to the enemy by asking someone to take a message to the secret service.'

The Case of Near Synonyms – the CAUSE group

- *chansheng* (产生, 361 instances in the LCMC corpus)
- *xingcheng* (形成, 334)
- *zaocheng* (造成, 208)
- *yinqi* (引起, 192)
- *dailai* (131)
- *daozhi* (导致, 79)
- *cushi* (促使, 44)
- *zhishi* (致使, 23)
- *yinfa* (引发, 11)
- *cucheng* (促成, 11)
- *niangcheng* (酿成, 4)

•Synonyms	•Negative	•Positive	•Neutral
•zhishi	•199 (99%)	•0	•1 (1%)
•niangcheng	•92 (98%)	•2 (2%)	•0
•zaocheng	•190 (91%)	•3 (2%)	•15 (7%)
•daozhi	•60 (76%)	•2 (3%)	•17 (21%)
•yinfa	•138 (69%)	•40 (2%)	•22 (11%)
•dailai	•64 (49%)	•36 (27%)	•31 (24%)
•yinqi	•83 (43%)	•28 (15%)	•81 (42%)
•chansheng	•111 (31%)	•88 (24%)	•162 (45%)
•xingcheng	•34 (10%)	•85 (26%)	•215 (64%)
•cushi	•2 (5%)	•26 (59%)	•16 (36%)
•cucheng	•2 (1%)	•171 (98%)	•2 (1%)

(4)

- (a.) *renmen keyi zhaochu xuduo zhishi ta duoluo de yuanyin* (LCMC: C)

  ‘We can find many causes for his degeneration.’

- (b.) *ruguo yushang zhongda qingkuang, zheyang caoshuai xingshi nanmian niangcheng dahuo* (LCMC: J)

‘In critical situations, taking hasty action like this would inevitably lead to a great disaster.’

(5)

- (a) *ni bixu dui ni <u>zaocheng</u> de yanzhong houguo fuze* (LCMC: K)

  ‘You must be responsible for the serious loss you have caused.’

- (b) *woshi de chuanghu meiyou guan, bobo de chuanglian zai yefeng li piaopiaofofo, <u>zaocheng</u> yi-zhong ji ju langman qingdiao de, feidong de yinxiang, zheng xiang nuzhuren xinuwuchang, zaodong bu ning de xingge* (LCMC: P)

  ‘The window of the bedroom was open. The thin curtain was fluttering gently in the night wind, giving an impression of romantic appeal and flying, just like the restlessly changing moods of its hostess.’

Conclusion

- The construction of suitable comparable corpora building upon existing monolingual corpora a fruitful way of enabling contrastive language studies
- Collocation and semantic prosodies exist in Chinese
- Both also show marked similarities – but some differences – to collocation/prosodies in assumed equivalents in English.