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

Tony McEnery
Richard Xiao
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](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](http://www.sinica.edu.tw/SinicaCorpus)

- PH corpus
  - Ca. 2 million words of newswire text (1990-1991)
  - POS version available at [http://www.ling.lancs.ac.uk/corplang/](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](http://icl.pku.edu.cn/Introduction/corpustagging.htm)
  - Searchable at [http://www.ling.lancs.ac.uk/corplang/](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)

• Three Mandarin corpora released by LDC
  – TREC, Gigaword and Callhome
  – See the LDC catalogue
Chinese corpora: A comparison

<table>
<thead>
<tr>
<th>Corpus</th>
<th>POS</th>
<th>Bal.</th>
<th>Channel</th>
<th>Variety</th>
<th>Contr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCMC</td>
<td>Yes</td>
<td>Yes</td>
<td>Written</td>
<td>Mainland</td>
<td>E – C</td>
</tr>
<tr>
<td>Sinica</td>
<td>Yes</td>
<td>Yes</td>
<td>Mixed</td>
<td>Taiwan</td>
<td>No</td>
</tr>
<tr>
<td>PH</td>
<td>No</td>
<td>No</td>
<td>Written</td>
<td>Mainland</td>
<td>No</td>
</tr>
<tr>
<td>PFR</td>
<td>Yes</td>
<td>No</td>
<td>Written</td>
<td>Mainland</td>
<td>No</td>
</tr>
<tr>
<td>LIVAC</td>
<td>No</td>
<td>No</td>
<td>Written</td>
<td>Mixed</td>
<td>C – C</td>
</tr>
<tr>
<td>SCCSD</td>
<td>No</td>
<td>Yes</td>
<td>Spoken</td>
<td>Mainland</td>
<td>No</td>
</tr>
<tr>
<td>TREC</td>
<td>No</td>
<td>No</td>
<td>Written</td>
<td>Mainland</td>
<td>No</td>
</tr>
<tr>
<td>Gigaword</td>
<td>No</td>
<td>No</td>
<td>Written</td>
<td>Mainland</td>
<td>No</td>
</tr>
<tr>
<td>Callhome</td>
<td>No</td>
<td>?</td>
<td>Spoken</td>
<td>Mixed</td>
<td>No</td>
</tr>
</tbody>
</table>
# LCMC: Sampling frame

<table>
<thead>
<tr>
<th>Code</th>
<th>Text category</th>
<th>Samples</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Press reportage</td>
<td>44</td>
<td>8.8%</td>
</tr>
<tr>
<td>B</td>
<td>Press editorials</td>
<td>27</td>
<td>5.4%</td>
</tr>
<tr>
<td>C</td>
<td>Press reviews</td>
<td>17</td>
<td>3.4%</td>
</tr>
<tr>
<td>D</td>
<td>Religion</td>
<td>17</td>
<td>3.4%</td>
</tr>
<tr>
<td>E</td>
<td>Skills/trades/hobbies</td>
<td>38</td>
<td>7.6%</td>
</tr>
<tr>
<td>F</td>
<td>Popular lore</td>
<td>44</td>
<td>8.8%</td>
</tr>
<tr>
<td>G</td>
<td>Biographies/essays</td>
<td>77</td>
<td>15.4%</td>
</tr>
<tr>
<td>H</td>
<td>Miscellaneous</td>
<td>30</td>
<td>6%</td>
</tr>
<tr>
<td>J</td>
<td>Science</td>
<td>80</td>
<td>16%</td>
</tr>
<tr>
<td>K</td>
<td>General fiction</td>
<td>29</td>
<td>5.8%</td>
</tr>
<tr>
<td>L</td>
<td>Mystery/detective fiction</td>
<td>24</td>
<td>4.8%</td>
</tr>
<tr>
<td>M</td>
<td>Science fiction</td>
<td>6</td>
<td>1.2%</td>
</tr>
<tr>
<td>N</td>
<td>Adventure and martial arts fiction</td>
<td>29</td>
<td>5.8%</td>
</tr>
<tr>
<td>P</td>
<td>Romantic fiction</td>
<td>29</td>
<td>5.8%</td>
</tr>
<tr>
<td>R</td>
<td>Humour</td>
<td>9</td>
<td>1.8%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>500</td>
<td>100%</td>
</tr>
</tbody>
</table>
## LCMC: Markup (XML)

<table>
<thead>
<tr>
<th>Level</th>
<th>Code</th>
<th>Gloss</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>text</td>
<td>Text type</td>
<td>TYPE</td>
<td>As per Table 2 Text Category</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ID</td>
<td>As per Table 2 Code</td>
</tr>
<tr>
<td>2</td>
<td>file</td>
<td>Corpus file</td>
<td>ID</td>
<td>Text ID plus file number starting from 01</td>
</tr>
<tr>
<td>3</td>
<td>p</td>
<td>Paragraph</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4</td>
<td>s</td>
<td>Sentence</td>
<td>n</td>
<td>Starting from 0001 onwards</td>
</tr>
<tr>
<td>5</td>
<td>w</td>
<td>Word</td>
<td>POS</td>
<td>Part-of-speech tags as per the LCMC tagset</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Punctuation and symbol</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gap</td>
<td>Omission</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
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
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?
<table>
<thead>
<tr>
<th>Author</th>
<th>Negative prosody</th>
<th>Positive prosody</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinclair (1987, 1990, 1991)</td>
<td>break out, happen, set in</td>
<td></td>
</tr>
<tr>
<td>Louw (1993, 2000)</td>
<td>bent on, build up of, end up <em>verbing</em>, get oneself <em>verbed</em>, a recipe for</td>
<td>build up a</td>
</tr>
<tr>
<td>Stubbs (1995, 1996, 2001b, 2001c)</td>
<td>accost, cause, fan the flame, signs of, underage, teenager(s)</td>
<td>provide, career</td>
</tr>
<tr>
<td>Moon (1998)</td>
<td>fan the flame</td>
<td></td>
</tr>
<tr>
<td>Partington (1998)</td>
<td>commit, peddle/peddler, dealings</td>
<td></td>
</tr>
<tr>
<td>Hunston (2002)</td>
<td>sit through</td>
<td></td>
</tr>
</tbody>
</table>
Our Study

- the consequence group
- the cause group
- commit
- price(s)
(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 fan-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, fan 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)
- *daozihi* (导致, 79)
- *cushi* (促使, 44)
- *zhishi* (致使, 23)
- *yinfa* (引发, 11)
- *cucheng* (促成, 11)
- *niangcheng* (酿成, 4)
<table>
<thead>
<tr>
<th>Synonyms</th>
<th>Negative</th>
<th>Positive</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>zhishi</td>
<td>199 (99%)</td>
<td>0</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>niangcheng</td>
<td>92 (98%)</td>
<td>2 (2%)</td>
<td>0</td>
</tr>
<tr>
<td>zaocheng</td>
<td>190 (91%)</td>
<td>3 (2%)</td>
<td>15 (7%)</td>
</tr>
<tr>
<td>daozihi</td>
<td>60 (76%)</td>
<td>2 (3%)</td>
<td>17 (21%)</td>
</tr>
<tr>
<td>yinfu</td>
<td>138 (69%)</td>
<td>40 (2%)</td>
<td>22 (11%)</td>
</tr>
<tr>
<td>dailai</td>
<td>64 (49%)</td>
<td>36 (27%)</td>
<td>31 (24%)</td>
</tr>
<tr>
<td>yinqi</td>
<td>83 (43%)</td>
<td>28 (15%)</td>
<td>81 (42%)</td>
</tr>
<tr>
<td>chansheng</td>
<td>111 (31%)</td>
<td>88 (24%)</td>
<td>162 (45%)</td>
</tr>
<tr>
<td>xingcheng</td>
<td>34 (10%)</td>
<td>85 (26%)</td>
<td>215 (64%)</td>
</tr>
<tr>
<td>cushi</td>
<td>2 (5%)</td>
<td>26 (59%)</td>
<td>16 (36%)</td>
</tr>
<tr>
<td>cucheng</td>
<td>2 (1%)</td>
<td>171 (98%)</td>
<td>2 (1%)</td>
</tr>
</tbody>
</table>
(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 zaocheng de yanzhong houguo fuze (LCMC: K)
  ‘You must be responsible for the serious loss you have caused.’

- (b) woshi de chuangu meiyou guan, bobo de chuanglian zai yefeng li piaopiaofofo, zaocheng 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.