Prosody in text: Patterns of word-length choices in written Chinese corpora

San Duanmu (端木三)
University of Michigan
March 4, 2010
City University of Hong Kong
Abstract

• There is a well-known word length problem in Chinese, where certain length combinations are preferred over others.
• Several analyses have been offered.
• However, no quantitative data have been produced to show how widely spread the word length problem is, or to support of the predictions of the phonological analysis.
• I show that quantitative data from written corpora support previous observations and the phonological analysis.
Outline

1. The word length problem
2. Analysis
3. Words with no length choices
4. Predictions
5. Corpus examination
6. Variation among corpora
7. Conclusions
1. The word length problem

- The dual vocabulary in Chinese: Many words can be monosyllabic or disyllabic, with more or less similar meanings

- Examples:

<table>
<thead>
<tr>
<th>English</th>
<th>Chinese 1</th>
<th>Chinese 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>coal</td>
<td>煤炭</td>
<td>煤</td>
</tr>
<tr>
<td>study</td>
<td>學習</td>
<td>學</td>
</tr>
<tr>
<td>worker</td>
<td>工人</td>
<td>工</td>
</tr>
<tr>
<td>store</td>
<td>商店</td>
<td>店</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The word length problem

• In a two-word expression, there are four theoretically possible length combinations:
  
  \[ 2 + 2 \]
  
  \[ 2 + 1 \]
  
  \[ 1 + 2 \]
  
  \[ 1 + 1 \]

• However, some length patterns are better than others.
The word length problem

- In $[\textbf{N N}]$, 1+2 is not favored
  
  2+2 技術 工人 ‘skill worker’
  
  2+1 技術 工
  
  *1+2 技 工人
  
  1+1 技 工

- In $[\textbf{V O}]$, 2+1 is not favored
  
  2+2 學習 繪畫 ‘study painting’
  
  *2+1 學習 畫
  
  1+2 學 繪畫
  
  1+1 學 畫
More examples

• In [N N], 1+2 is not favored
  2+2 煤炭 商店 ‘coal store’
  2+1 煤炭 店
  *1+2 煤 商店
  1+1 煤 店

• In [V O], 2+1 is not favored
  2+2 种植 大蒜 ‘plant garlic’
  *2+1 种植 蒜
  1+2 种 大蒜
  1+1 种 蒜
Summary

• Length pattern preferences are quite robust

• Preferred length patterns are sensitive to syntax:

  \[\text{NN}: \ast 1+2\quad \text{coal} \quad \text{store}\]

  \[\text{VO}: \ast 2+1\quad \text{plant} \quad \text{garlic}\]

• The word length problem has been observed for a long time (e.g. 吕叔湘 1963, 吴为善 1986, 陆丙甫 1989, Lu & Duanmu 1990, 2002, 冯胜利 1998)
2. Analysis

• Several analyses have been offered to account for the word length problem.

• For illustration, let us consider the analysis in metrical phonology (see Duanmu 2008 for a review of other analyses)

• Metrical requirements (S indicates a syllable):
  
  **Binary foot**: (S+S) or (SS)

  **Phrasal stress** as in English: [N N] and [V O]
Analysis

• Metrical requirements:
  
  **Binary foot**: (S+S) or (SS)
  
  **Phrasal stress**: [N N] and [V O].

• Foot structures for [N N] (offending foot underlined)
  
  2 + 2 (SS)(SS)
  2 + 1 (SS)S
  *1 + 2 *(S)(SS)
  1 + 1 (S+S)

• Foot structures for [V O] (offending foot underlined)
  
  2 + 2 (SS)(SS)
  *2 + 1 *(SS)(S)
  1 + 2 S(SS)
  1 + 1 (S+S)
3. Words with no length choices

• Words with no length choices are not subject to length pattern preferences

• For example, 手套 and 喜歡 have no monosyllabic form

• [N N] 1+2 [皮 手套] ‘leather glove’
  Cf. *1+2 [煤 商店] ‘coal store’

• [V O] 2+1 [喜歡 錢] ‘love money’
  Cf. *2+1 [种植 蒜] ‘plant garlic’

• However, they are believed to be uncommon.
4. Predictions

- The metrical analysis makes strong predictions
- For [N N]
  - $2 + 2$ is frequent
  - $2 + 1$ is frequent
  - $1 + 2$ is infrequent, except for words with no flexible length
  - $1 + 1$ is frequent
- For [V O]
  - $2 + 2$ is frequent
  - $2 + 1$ is infrequent, except for words with no flexible length
  - $1 + 2$ is infrequent
  - $1 + 1$ is frequent
- The predictions remain to be verified.
Questions about predictions

- How frequent is frequent?
- How many words in Chinese have no flexible length?
- Is (2+2) better than (2+1) for [N N]?
- Is (1+2) better than (1+1) for [V O]?
- Etc.
A ‘reasonable’ expectation

• For [N N]
  2 + 2  25%
  2 + 1  25%
  1 + 2  5%  (or less)
  1 + 1  25%
  Others 5%  (e.g. 2+3, 3+3, 3+1, 4+2, etc.)

• For [V O]
  2 + 2  25%
  2 + 1  5%  (or less)
  1 + 2  25%
  1 + 1  25%
  Others 5%
5. Corpus examination

- Lancaster Corpus of Mandarin Chinese (LCMC, McEnery & Xiao 2004)
- Examine all [N N] units
- Examine all [V O] units
LCMC

- [http://ota.ahds.ac.uk/headers/2474.xml](http://ota.ahds.ac.uk/headers/2474.xml)
- Balanced selection of modern texts of various styles
- Automatic POS tagging with hand-check; accuracy is reported to be 98%
- **Overall statistics**
  - Words (词数) 839,006 (84%)
  - Punctuations 162,820 (16%)
  - Characters (字数) 1,341,628 (89%)
  - Punctuation graphs 166,861 (11%)
  - All graphs 1,508,489
  - Average word length 1.6 characters
或许，严同已在忙碌中，开始算清了为那九十万斤粮票。
LCMC coding

<w POS="nr">严</w> name as ‘nr’; can be counted as N
<w POS="nr">同</w> 已 name as ‘nr’; can be counted as N
<w POS="p">在</w>
<w POS="r">这</w>
<w POS="a">忙碌</w> labeled as adjective; should it be N?
<w POS="f">中</w> locative: should it count as N?
<c POS="w">，</c> punctuation
<w POS="v">开始</w> disyllabic V
<w POS="v">算清</w>
<w POS="u">了</w>
<w POS="p">为</w>
<w POS="r">那</w>
<w POS="rm">九十万</w> numeral not labeled as N
<w POS="q">斤</w>
<w POS="n">粮票</w> disyllabic N
Types of Texts in LCMC

- LCMC has 15 different types of texts
- For example:
  - A: News reportage
  - D: Religion
  - G: Biographies and essays
  - H: Reports and official documents
  - J: Science: academic prose
  - R: Humor
[N N] units

- Corpus
- Initial results
- Revised results
Corpus for [N N]

- Corpus used: Biographies and essays, 77 texts
- Overall statistics (typical of LCMC)
  
<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words (词数)</td>
<td>130,309</td>
<td>(84%)</td>
</tr>
<tr>
<td>Punctuations</td>
<td>24,031</td>
<td>(16%)</td>
</tr>
<tr>
<td>Characters (字数)</td>
<td>202,178</td>
<td>(89%)</td>
</tr>
<tr>
<td>Punctuation graphs</td>
<td>24,895</td>
<td>(11%)</td>
</tr>
<tr>
<td>All graphs</td>
<td>227,073</td>
<td></td>
</tr>
<tr>
<td>Average word length:</td>
<td>1.55 characters</td>
<td></td>
</tr>
</tbody>
</table>
What is a noun?

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>23,407</td>
<td>common noun</td>
</tr>
<tr>
<td>ng</td>
<td>1,365</td>
<td>noun morpheme</td>
</tr>
<tr>
<td>nr</td>
<td>7,279</td>
<td>personal name</td>
</tr>
<tr>
<td>ns</td>
<td>2,452</td>
<td>place name</td>
</tr>
<tr>
<td>nt</td>
<td>92</td>
<td>organization name</td>
</tr>
<tr>
<td>nz</td>
<td>294</td>
<td>other proper noun</td>
</tr>
<tr>
<td>vn</td>
<td>2,581</td>
<td>verb with nominal function</td>
</tr>
<tr>
<td>an</td>
<td>494</td>
<td>adjective with nominal function</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>37,964</strong></td>
<td></td>
</tr>
</tbody>
</table>
What to count?

• Including [N N]

• Excluding [N N N], e.g. <春节><联欢><晚会>

• Excluding [N N N N], e.g. <中央> <农村> <文化> <工作队>
Initial results on [N N]: **Surprise!**

Preliminary counts of [N N] in biographies and essays

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2+2</td>
<td>2,382</td>
<td>37%</td>
</tr>
<tr>
<td>2+1</td>
<td>462</td>
<td>7%</td>
</tr>
<tr>
<td>1+2</td>
<td>2,005</td>
<td>31%</td>
</tr>
<tr>
<td>1+1</td>
<td>920</td>
<td>14%</td>
</tr>
<tr>
<td>Others</td>
<td>636</td>
<td>10%</td>
</tr>
<tr>
<td>All</td>
<td>6,405</td>
<td>100%</td>
</tr>
</tbody>
</table>
Understanding 2+1 [N N]

• Why is 2+1 so low (7%, hoping for 20%)?

• Locatives not included
  – E.g. <生活><中> is [N F]

• Some 2+1 are parsed as N=3
  – E.g. 联欢节,工作队, 解放区, 保险箱, etc.

• Solutions:
  – Count locatives as N
  – Check N=3 for 2+1 and 1+2
Understanding 1+2 [N N]

• Why is 1+2 so high (31%, hoping for 5%)?

• Most of them involve personal names
  – E.g. <李><所长>, <林><伯渠>, <徐><特立>

• Solution
  – Exclude personal names
Understanding 1+1 [N N]

- Why is 1+1 so low (14%, hoping for 20+%)?
- Most of them are parsed as 2
- Solution
  - Check N=2; distinguish 1+1 from true N=2
Revising the count

• Exclude personal names
• Include locatives as N
• Divide single N=3 into 1+2 and 2+1
• Divide single N=2 into 1+1 and true 2
New result on [N N]

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2+2</td>
<td>2,379</td>
<td>18.0%</td>
</tr>
<tr>
<td>2+1</td>
<td>2,256</td>
<td>17.1%</td>
</tr>
<tr>
<td>1+2</td>
<td>59</td>
<td>0.4%</td>
</tr>
<tr>
<td>1+1</td>
<td>7,794</td>
<td>59.0%</td>
</tr>
<tr>
<td>Others</td>
<td>724</td>
<td>5.5%</td>
</tr>
<tr>
<td>All</td>
<td>13,211</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Notes:

a. 2+1 includes 75% of N=3
b. Of 296 1+2, only 20% are real [N N]
c. 1+1 includes 481 1+1 plus 50% of N=2
New result on [N N], excluding 1+1

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2+2</td>
<td>2,379</td>
<td>43.9%</td>
</tr>
<tr>
<td>2+1</td>
<td>2,256</td>
<td>41.6%</td>
</tr>
<tr>
<td>1+2</td>
<td>59</td>
<td>1.1%</td>
</tr>
<tr>
<td>Others</td>
<td>724</td>
<td>13.4%</td>
</tr>
<tr>
<td>All</td>
<td>5,418</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Length Pattern of NN, Excluding 1+1 (prediction: *1+2)
Summary on [N N]

- Corpus data requires proper interpretation
- Quantitative data from written corpora support previous observations of preferred length combinations of [N N]
- Quantitative data also supports the predictions of the metrical analysis
Corpus for [V N]

- Same as that for [N N]: Biographies and essays
- Overall statistics (typical of LCMC)
  
  | Words (词数) | 130,309 |
  | Characters (字数) | 202,178 |
Initial result on [V N]: **Surprise?**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2+2</td>
<td>1,261</td>
<td>35%</td>
</tr>
</tbody>
</table>
| 2+1     | 220   | 6% | (a bit too high?)
| 1+2     | 1,185 | 33%|
| 1+1     | 520   | 15%| (too low)       |
| Others  | 383   | 11%|
| All     | 3,569 |    |
Understanding 2+1 [V N]

- 2+1 seems a bit too high (6%, hoping for less)
- Some should be [A N] or [N N]
  - E.g. [通电稿] ‘telegraph draft’, [发行权] ‘publication right’
- Some should be [[… V] X]
  - E.g. [访问时], [混杂着]
- Some are [V [N N]] but labeled as [V N F], e.g. [回到 [房里]]
- Only a small minority are real 2+1, e.g. [打开书]
- Conclusion: real 2+1 is about 1-2%, as expected
Understanding 1+1 [V N]

• Why is 1+1 so low (15%, hoping for 20+%)?

• Most of them are parsed as V=2

• Solution
  – Check V=2 for 1+1 and true V=2
Revising the count

• Divide single $V=2$ into $[V O] 1+1$ and true $V=2$
• (Also, use new definitions of $N$)
### New result on \([V N]\)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2+2</td>
<td>1,237</td>
<td>28%</td>
</tr>
<tr>
<td><strong>2+1</strong></td>
<td><strong>74</strong></td>
<td><strong>2%</strong></td>
</tr>
<tr>
<td>1+2</td>
<td>1,150</td>
<td>26%</td>
</tr>
<tr>
<td>1+1</td>
<td>1,589</td>
<td>36%</td>
</tr>
<tr>
<td>Others</td>
<td>344</td>
<td>8%</td>
</tr>
<tr>
<td>All</td>
<td>4,394</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Notes:** based on 100 samples each

a. In 2+1, only 20% of 371 are real \([V N]\)

b. 1+1 \([V N]\) includes 615 1+1 plus 10% of \(V=2\)
Length Patterns of VN (prediction: *2+1)
Summary on [V N]

• Again, quantitative data from written corpora support previous observations of preferred length combinations of [V N]

• Quantitative data also supports the predictions of the metrical analysis
6. Variation among corpora

• Do we expect length patterns to differ among different styles of writing?

• If a text has many proper names or special actions, we may expect more use of long words, or a higher rate or $N = 2$, $V = 2$, and $2 + 2$ in NN and VN.

• In contrast, in more colloquial speech, we expect more shorter words and a lower rate of $N = 2$, $V = 2$, a lower rate of $2 + 2$, and a higher rate of $1 + 1$. 
Some Text Types in LCMC

- LCMC has 15 different types of texts

- Some examples:

<table>
<thead>
<tr>
<th>Style</th>
<th>Possible property</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: News</td>
<td>neutral</td>
</tr>
<tr>
<td>D: Religion</td>
<td>neutral</td>
</tr>
<tr>
<td>G: Essays (biography)</td>
<td>neutral</td>
</tr>
<tr>
<td>H: Official documents</td>
<td>longer words (formal)</td>
</tr>
<tr>
<td>J: Science</td>
<td>longer words (formal)</td>
</tr>
<tr>
<td>R: Humor stories</td>
<td>shorter words (colloquial)</td>
</tr>
</tbody>
</table>
Length patterns of NN

- 2+2
- 2+1
- 1+2
- 1+1
- Other NN

Categories: News, Religion, Essays, Documents, Science, Humor, All
NN without 1+1

- 2+2
- 2+1
- 1+2
- Other NN

Categories:
- News
- Religion
- Essays
- Documents
- Science
- Humor
- All
VN without 1+1

Graph showing the percentage distribution across different categories: News, Religion, Essays, Documents, Science, Humor, and All. The categories are represented by different colors: 2+2, 2+1, 1+2, and Others.

- News: Blue line (2+2)
- Religion: Red line (2+1)
- Essays: Green line (1+2)
- Documents: Purple line (Others)
- Science: Blue line (2+2) with a sharp increase
- Humor: Red line (1+2) with a peak
- All: Blue line (2+2) with a gradual decrease
Summary

• Different styles of writing are well reflected in word length patterns

• Technical texts use more disyllabic N and V, more 2+2, and fewer 1+1

• Colloquial texts use fewer disyllabic N and V, fewer 2+2, and more 1+1
7. Conclusions

• Word length preferences are real generalizations in Chinese phonology
• Phonological generalizations should and can be verified by quantitative data
• Word length preferences in Chinese are strong and supported by corpus data
• Even written texts reflect rhythmic requirements
• It would be a mystery if the word length problem is only found in Chinese
• It would be interesting to find out whether the word length problem is found in other languages
References


