The More Frequent a Character Word Is, the Simpler Form It Has

QUALICO 2018
July 5th 2018
Yuan-Lu Chen, University of Arizona, cheny@email.arizona.edu
Hsuan-Ying Liu, University of North Dakota
Link to the current slides: https://tinyurl.com/QUALICO-2018-Chen

Road Map

1. Background and Research Question: Golden Pattern in Natural Languages
2. Finding the golden pattern in Chinese: a corpus study
3. Discussion and Conclusion

Link to the current slides: https://tinyurl.com/QUALICO-2018-Chen

Road Map

1. **Background and Research Question: Golden Pattern in Natural Languages**
2. Finding the golden pattern in Chinese: a corpus study
3. Discussion and Conclusion

Background: information theory and language

Cross-linguistically the **length of a word** is predicted by its **frequency** and the amount of **information content** it has: the more frequent and with less information content the shorter a word is (Sigurd et. al. 2004, Piantadosi et al. 2011).

{more frequent, less information content} → short

{less frequent, more information content}→ long

Background: What is information content?

- Information content defined as how unexpected it is for a word to occur in a certain context.
- Zero information content: “United States of **America** ”
- High information content: “the **Unicorn**”

Background: information theory and language

{more frequent, less information content} → short

{less frequent, more information content}→ long

Human language lexical systems result from an optimization of communicative pressures. This pattern is found across different languages (Czech, Dutch, English, French, German, Italian, Polish, Portuguese, Romanian, Spanish, and Swedish) (Piantadosi et al. 2011).

Figure adapted from Piantadosi et al. (2011:2): the correlations in English

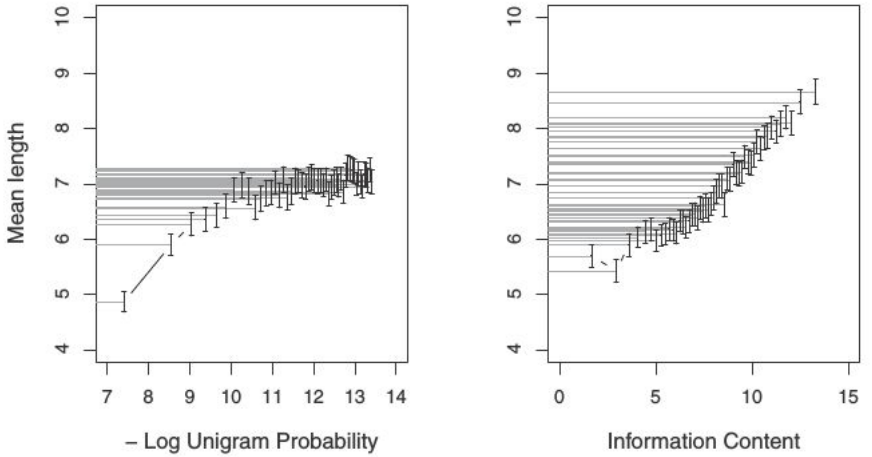

Fig. 2. Relationship between frequency (negative log unigram probability) and length, and information content and length. Error bars represent SEs and each bin represents 2% of the lexicon.

Road Map

1. ~~Background and Research Question: Golden Pattern in Natural Languages~~
2. **Finding the golden pattern in Chinese: a corpus study**
3. Discussion and Conclusion

The Puzzles

- How do we define the length/complexity of a Chinese word?

  Two proposed measurements:

  a. Amount of characters:
     - character_length("邑") = 1 ; character_length("国都") = 2 (both means 'capital')
     - potential problem:
       - character_length("龘") = 1; character_length("龍龍") = 1 ("龘" reads as 'dá', 'flying dragon')

  b. Total Stroke Number:
     - stroke_length("邑") = 7; stroke_length("国都") = 18; stroke_length("龘") = 1; stroke_length("龍龍") = 48

- The golden pattern of "the more frequent, the shorter" holds in alphabet-based languages. How does the optimization pattern extend to Chinese, a character based language?

Research Questions

Given a Chinese Word	
Length/Complexity	Information measurement
Character Length	Frequency
Total Stroke Number	Average Information Content

- Does Chinese manifest the golden patten of “the more frequent, the shorter/less complex”?
- Which Chinese written system, simplified Chinese or traditional Chinese, is more efficient in term of communication (i.e. is closer to the golden pattern)?
- Piantadosi et al. (2011) found that the average information content of a word is a better predictor than frequency to predict its length. Does the same pattern hold in Chinese (i.e. average information content is better than frequency)?

Research Materials

- A collection of Chinese classic novels (Vierthaler, 2016). This corpus consists of texts written in Chinese during the Ming and Qing dynasties, spanning roughly 1368 to the early 20th century (the newest text was written in 1916).
- Pre-process: The whole text is tokenized into sentences by using punctuations as delimiter. Each sentence is then tokenized into words by using *Polyglot* word tokenizer (Al-Rfou, 2017).
- 1,720,988 sentences, 7,291,097 words, and 9,125,189 characters.

Link to the current slides: https://tinyurl.com/QUALICO-2018-Chen10

An Example

- Text: “看见纸张白亮，图书鲜红，...”
  - Sentences:
    - “看见纸张白亮”
      - Words:
        - “看见”, “纸张”, “白”, “亮”
    - “图书鲜红”
      - Words:
        - “图书”, “鲜”, “红”

- Text: “看見紙張白亮，圖書鮮紅，...”
  - Sentences:
    - “看見紙張白亮”
      - Words:
        - “看見”, “紙張”, “白”, “亮”
    - “图书鲜红”
      - Words:
        - “圖書”, “鮮”, “紅”

Research Methods

The four key pieces of information that need be extracted from our data are:

- Complexity measurements:
  1. total stroke number (A python program that counts stroke number: https://github.com/lucien0410/count_stroke)
     a. In the form of simplified Chinese, e.g. **证实** zheng4 shi2 'verify'; total stroke number = 15

     b. In the form of traditional Chinese, e.g. **證實** zheng4 shi2 'verify'; total stroke number = 33
  2. Character length
- Information measurements:
  3. average information content of each word
  4. frequency of each word

Measuring Average Information Content: an example

Definition of Average information content (Piantadosi et al. (2011)):

$$Average\ information\ content(w) = -\frac{1}{N}\sum_{i=1}^{N} log_2 P(w|c_i) \qquad (1)$$

where c_i is the context for the ith occurrence of w and N is the total frequency of w in the corpus.

A python program that computes avergae information content of each token in a corpus
https://github.com/lucien0410/average_information_content_calculator

Measuring Average Information Content: an example “纸张” ‘paper’

1. First of all we need to find all the contexts where “纸张” ‘*paper*’ occurs. In this example the target word occurs in two contexts (N=2):

   (a) c_1 看见 纸张 ‘*saw*’ ‘*paper*’

   (b) c_2 笔砚 纸张 ‘*writing brush and ink stone*’ ‘*paper*’

2. Now we need to find the unigram probability of the context word and bigram probability of the context word followed by the target word (here the probabilities are estimated using relative frequency):

   (a) $p(\text{‘}saw\text{’})$
   $= 0.00021293629609574254$

   (b) $p(\text{‘}saw\text{’}, \text{‘}paper\text{’})$
   $= 1.3715357236366488e-07$

   (c) $p(\text{‘}writing\ brush\ and\ ink\ stone\text{’})$
   $= 1.7532013956814655e-05$

   (d) $p(\text{‘}writing\ brush\ and\ ink\ stone\text{’}, \text{‘}paper\text{’})$
   $= 1.3715357236366488e-07$

3. By plugging in the unigram and bigram probabilities, the conditional probabilities are calculated:

   (a) $p(\text{‘}paper\text{’}|\text{‘}saw\text{’})$
   $= \frac{p(\text{‘}saw\text{’}, \text{‘}paper\text{’}))}{p(\text{‘}saw\text{’})}$
   $= \frac{2b}{2a}$
   $= 0.0006441061241245434$

   (b) $p(\text{‘}paper\text{’}|\text{‘}writing\ brush\ and\ ink\ stone\text{’})$
   $= \frac{p(\text{‘}writing\ brush\ and\ ink\ stone\text{’}, \text{‘}paper\text{’})}{p(\text{‘}writing\ brush\ and\ ink\ stone\text{’})}$
   $= \frac{2d}{2c}$
   $= 0.00782303577338607$

4. The final step is to find the log values of the conditional probabilities and average them. By plugging the values in (3), the average information content of “纸张” ‘paper’ is calculated:
   $= -\frac{1}{2}(log_2(3a) + log_2(3b))$
   $= -\frac{1}{2}(-10.600413970503666 + -6.99805572248539)$
   $= 8.799234846494528$

Result

	A	B	C	D	E	F	G	H	I	J
1	sim_char	trad_char	sim_unicode	trad_unicode	sim_stroke	trad_stroke	average_surpisal	len	freq	log_freq
2	平定	平定	u'\u5e73\u5b9a'	u'\u5e73\u5b9a'	13	13	11.750016411	2	109	6.7681843248
3	不学无术	不學無術	u'\u4e0d\u5b66\u65e0\u672f'	u'\u4e0d\u5b78\u7121\u8853'	21	43	3.8641999758	4	1	0
4	奸险	奸險	u'\u5978\u9669'	u'\u5978\u96aa'	15	21	11.1459867939	2	22	4.4594316186
5	蟹螯	蟹螯	u'\u87f9\u87af'	u'\u87f9\u87af'	35	35	10.004887756	2	1	0
6	涉	涉	u'\u6d89'	u'\u6d89'	10	10	10.5108707647	1	342	8.4178525149
7	唇吻	唇吻	u'\u8123\u543b'	u'\u8123\u543b'	18	18	12.6417194005	2	3	1.5849625007
8	第二	第二	u'\u7b2c\u4e8c'	u'\u7b2c\u4e8c'	13	13	10.4285130566	2	895	9.8057438722
9	鍊子	鍊子	u'\u934a\u5b50'	u'\u934a\u5b50'	20	20	12.5269672371	2	1	0
10	缺	缺	u'\u920c'	u'\u920c'	12	12	12.1330668269	1	1	0
11	小腿	小腿	u'\u5c0f\u817f'	u'\u5c0f\u817f'	16	16	17.833841852	2	2	1
12	第五	第五	u'\u7b2c\u4e94'	u'\u7b2c\u4e94'	15	15	11.4600998787	2	282	8.1395513524
13	外姓	外姓	u'\u5916\u59d3'	u'\u5916\u59d3'	13	13	10.1860508672	2	6	2.5849625007
14	民江	民江	u'\u6c11\u6c5f'	u'\u6c11\u6c5f'	11	11	3.8641999758	2	1	0
15	小腹	小腹	u'\u5c0f\u8179'	u'\u5c0f\u8179'	16	16	17.3873100877	2	2	1
16	薙	薙	u'\u8599'	u'\u8599'	16	16	7.4849893591	1	5	2.3219280949
17	五级	五級	u'\u4e94\u7ea7'	u'\u4e94\u7d1a'	10	13	10.2894719785	2	6	2.5849625007
18	俗话	俗話	u'\u4fd7\u8bdd'	u'\u4fd7\u8a71'	17	22	15.9232320961	2	13	3.7004397181
19	通志	通誌	u'\u901a\u5fd7'	u'\u901a\u8a8c'	17	24	7.720685694	2	31	4.9541963104
20	让座	讓座	u'\u8ba9\u5ea7'	u'\u8b93\u5ea7'	15	34	10.5936318972	2	1	0
21	小腰	小腰	u'\u5c0f\u8170'	u'\u5c0f\u8170'	16	16	8.5080561655	2	1	0
22	踏歌	踏歌	u'\u8e0f\u6b4c'	u'\u8e0f\u6b4c'	29	29	12.7471007858	2	5	2.3219280949
23	竦然	竦然	u'\u7ae6\u7136'	u'\u7ae6\u7136'	24	24	13.1165859117	2	8	3
24	战衣	戰衣	u'\u6218\u8863'	u'\u6230\u8863'	15	22	13.5605928804	2	2	1
25	俗语	俗語	u'\u4fd7\u8bed'	u'\u4fd7\u8a9e'	18	23	12.8392181572	2	105	6.7142455177
26	看住	看住	u'\u770b\u4f4f'	u'\u770b\u4f4f'	16	16	11.6784199059	2	6	2.5849625007
27	籴米	糴米	u'\u7c74\u7c73'	u'\u7cf4\u7c73'	14	28	12.2381575828	2	16	4

Result

Table 1. the Pearson correlation coefficient of the 5 measures

Measure	mean	sd	1	2	3	4
1. Log Frequency	3.287	2.715				
2. Average Information Content	11.877	3.215	-0.269***			
3. Word Length (by Character)	1.998	0.666	-0.311***	0.295***		
4. Total Stroke Number (Simplified Chinese)	16.434	6.401	-0.324***	0.186***	0.589***	
5. Total Stroke Number (Traditional Chinese)	19.836	8.341	-0.274***	0.181***	0.590***	0.784***

N=55,744; *Note *p<.05, **p<.01, ***p<.001

Road Map

1. ~~Background and Research Question: Golden Pattern in Natural Languages~~
2. ~~Finding the golden pattern in Chinese: a corpus study~~
3. **Discussion and Conclusion**

Discussion

- As both frequency and average information content are correlated with word length and total stoke number, Chinese (simplified and traditional) does have the optimized pattern.
- Simplified Chinese is closer to the golden pattern.

Table 1. the Pearson correlation coefficient of the 5 measures

Measure	mean	sd	1	2	3	4
1. Log Frequency	3.287	2.715				
2. Average Information Content	11.877	3.215	-0.269***			
3. Word Length (by Character)	1.998	0.666	-0.311***	0.295***		
4. Total Stroke Number (Simplified Chinese)	16.434	6.401	-0.324***	0.186***	0.589***	
5. Total Stroke Number (Traditional Chinese)	19.836	8.341	-0.274***	0.181***	0.590***	0.784***

N=55,744; *Note *p<.05, **p<.01, ***p<.001

Discussion

Why is simplified Chinese more optimal than traditional Chinese?

- During the 1950's, simplified Chinese became the official written language in the People's Republic of China; however, it has been existing for hundreds of years (DeFrancis, 1986).
- Speculation: Without regulated by the law (as the official written form), simplified Chinese has more freedom to evolve.

Discussion: remaining puzzles

Which **information measurement** is better, frequency or average information content?

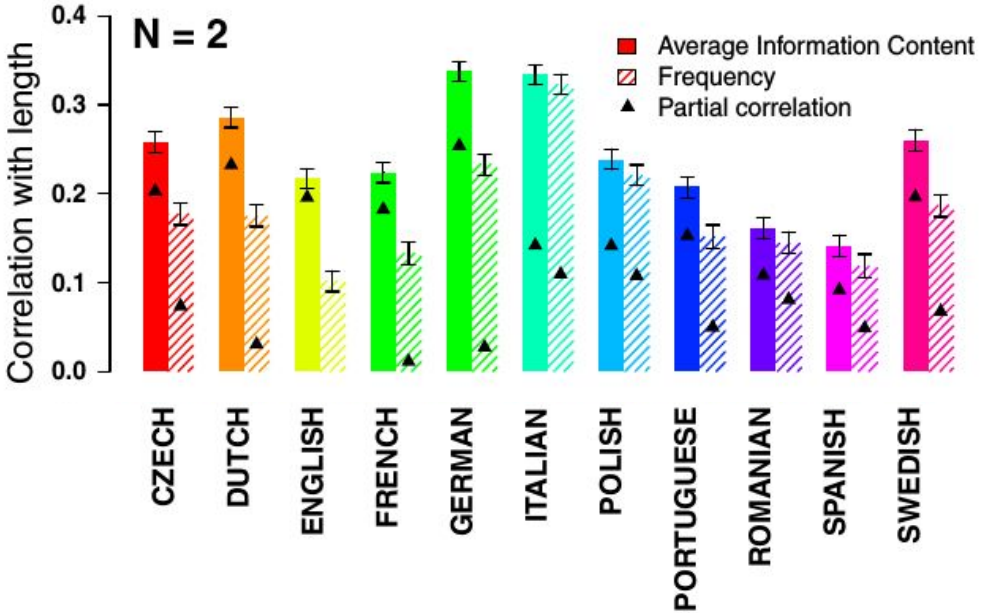

Figure from Piantadosi et al. (2011:2)

Discussion: remaining puzzles

Which **information measurement** is better, frequency or average information content?

- To our surprise, frequency is the better measurement than average information content for Chinese, contrary to the finding in Piantadosi et al. (2011).
- Chinese is different from the 11 languages in Piantadosi et al. (2011). We speculate that in Chinese the context does not provide the same amount of information as the context in other languages.

Table 1. the Pearson correlation coefficient of the 5 measures

Measure	mean	sd	1	2	3	4
1. Log Frequency	3.287	2.715				
2. Average Information Content	11.877	3.215	-0.269***			
3. Word Length (by Character)	1.998	0.666	-0.311***	0.295***		
4. Total Stroke Number (Simplified Chinese)	16.434	6.401	-0.324***	0.186***	0.589***	
5. Total Stroke Number (Traditional Chinese)	19.836	8.341	-0.274***	0.181***	0.590***	0.784***

N=55,744; *Note *p<.05, **p<.01, ***p<.001

Discussion: remaining puzzles

Which complexity measurement is better, character length or total stroke number?

- Frequency, simplified Chinese: total stroke number
- Frequency, traditional Chinese: character length
- Average information content, simplified Chinese: character length
- Average information content, traditional Chinese: character length

Table 1. the Pearson correlation coefficient of the 5 measures

Measure	mean	sd	1	2	3	4
1. Log Frequency	3.287	2.715				
2. Average Information Content	11.877	3.215	-0.269***			
3. Word Length (by Character)	1.998	0.666	-0.311***	0.295***		
4. Total Stroke Number (Simplified Chinese)	16.434	6.401	-0.324***	0.186***	0.589***	
5. Total Stroke Number (Traditional Chinese)	19.836	8.341	-0.274***	0.181***	0.590***	0.784***

N=55,744; *Note *p<.05, **p<.01, ***p<.001

Conclusion

In the current study, we found that Chinese, a character-based, has some patterns of optimizing its efficiency as alphabet-based languages.

Reference

Al-Rfou, R. (2017, Aug). *polyglot.* https://github.com/aboSamoor/polyglot.

Chao, E., Wang, E., Hsuan, M., & Matloff, N. (2009). *The chinese language, ever evolving.* https://roomfordebate.blogs.nytimes.com/2009/05/02/chinese-language-ever-evolving/. The New York Times.

Chen, H., Liang, J., & Liu, H. (2015). How does word length evolve in written chinese? *PloS one*, *10*(9), e0138567.

DeFrancis, J. (1986). *The chinese language: Fact and fantasy.* University of Hawaii Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency–zipf revisited. *Studia Linguistica*, *58*(1), 37–52.

Unicode, C. (2017, Jun). *Unicode character database.* http://www.unicode.org/Public/UCD/latest/.

Vierthaler, P. (2016). *Late imperial chinese texts: The corpus for fiction and history: Polarity and stylistic gradience in late imperial chinese literature.* Harvard Dataverse.

Yan, X., & Minnhagen, P. (2015). Maximum entropy, word-frequency, chinese characters, and multiple meanings. *PloS one*, *10*(5), e0125592.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books.