A quantitative probe into the hierarchical structure of written Chinese

Heng Chen, Guangdong University of Foreign Studies
Haitao Liu, Zhejiang University

Outline

- Problems
- Materials and Methods
- Results
- Discussions
- Conclusions

Problems

- Language units
  - Saussure: language entities or language units
- Language levels
  - American descriptive linguistics
  - multi-level system
- The boundaries between language levels
  - not clear
  - different linguistic schools different definitions

Materials and Methods

- microscopic scale VS. the system level
- Authentic language data
- simultaneously on all levels
- an orderly hierarchy of levels

Materials

Language units	scale
Character (tokens)	1,314,058
Character (types)	4,705
Clauses (types)	126,455
Sentence (types)	45,969
Word (types)	847,521

Investigations of several levels in one text

Lancaster Corpus of Mandarin Chinese

Methods

- Menzerath-Altmann‘s law （short for MA law）
  - the longer a word (measured in number of syllables), the shorter its syllables (measured in number of phonemes)

- Altmann (1980), two generalizations (in two directions)
  - first, not only for **words and syllables**, but also for other language units **(clause - word**, **sentence - clause**)
  - second, monotonicity is not required, the mean size of constituents is a function of the size of the construct

Methods

MA law

(1) $y(x) = ax^b$

- parameter b is negative, decreasing function

(2) $y(x) = ax^b e^{cx}$

- this function can attain its maximum not only for x=1, but also in other points
- y – mean size of constituent, x – construct size

Formula (1) is in many aspects more simple and „nicer“, but it does not fit data sufficiently well in some cases..

We say the result is accepted for $R^2 > 0.75$, good for R
0.80, and very good for $R^2 > 0.90$.

Methods

- Language units in written Chinese
  - Sentence, Clause, Word, Character, Component, Stroke
- Sentences
  - separated from one another by using special marks of punctuation (full-stop, question-mark, exclamation-mark).
- Clause
  - Lu (2006) claims that the constituents just between two punctuations (comma and period) can be defined as clauses roughly.
  - But we need to state that, since in LCMC sentences are tagged, we choose comma and semicolon as our marks of clause boundaries.

word "语言"("yǔ yán", which means nguage") consists of two characters " , 言 " ("yǔ, yán", which means nguage, parole"), and the two racters have nine strokes 丶 乛 — 丨 — 丨 ㄱ — ， and seven strokes 丶 —,— 丨,ㄱ,—respectively, eleven in total. ""言" have three components "讠"" "口"and one component"言" (means role"), respectively.

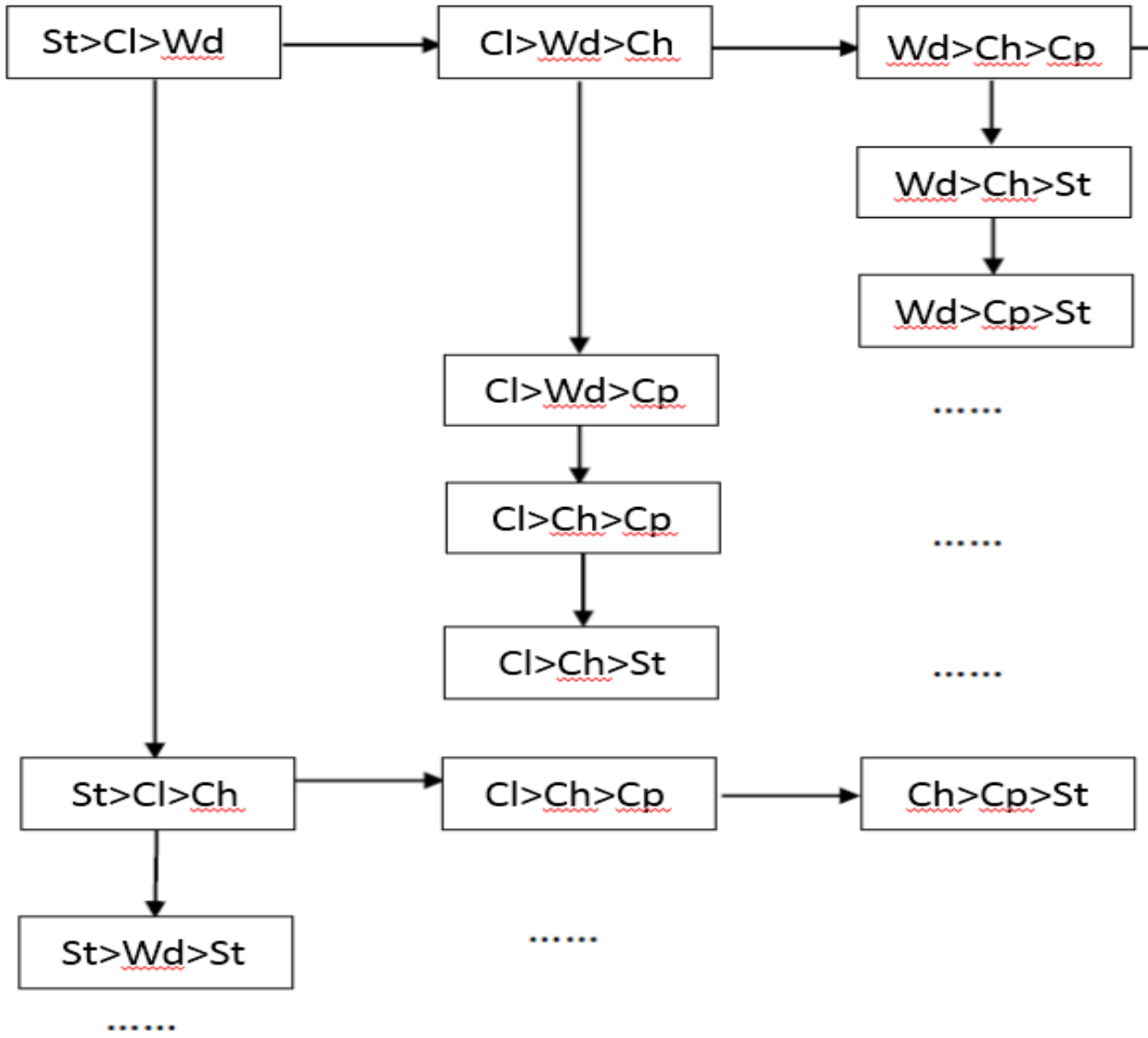
20902 Characters code CJK character se

Methods

- Why no Phrase?

- Phrase is not the basic language unit.
  - it is difficult to segment a sentence into several phrase sequences
  - Two phrases can be composed into one phrase.

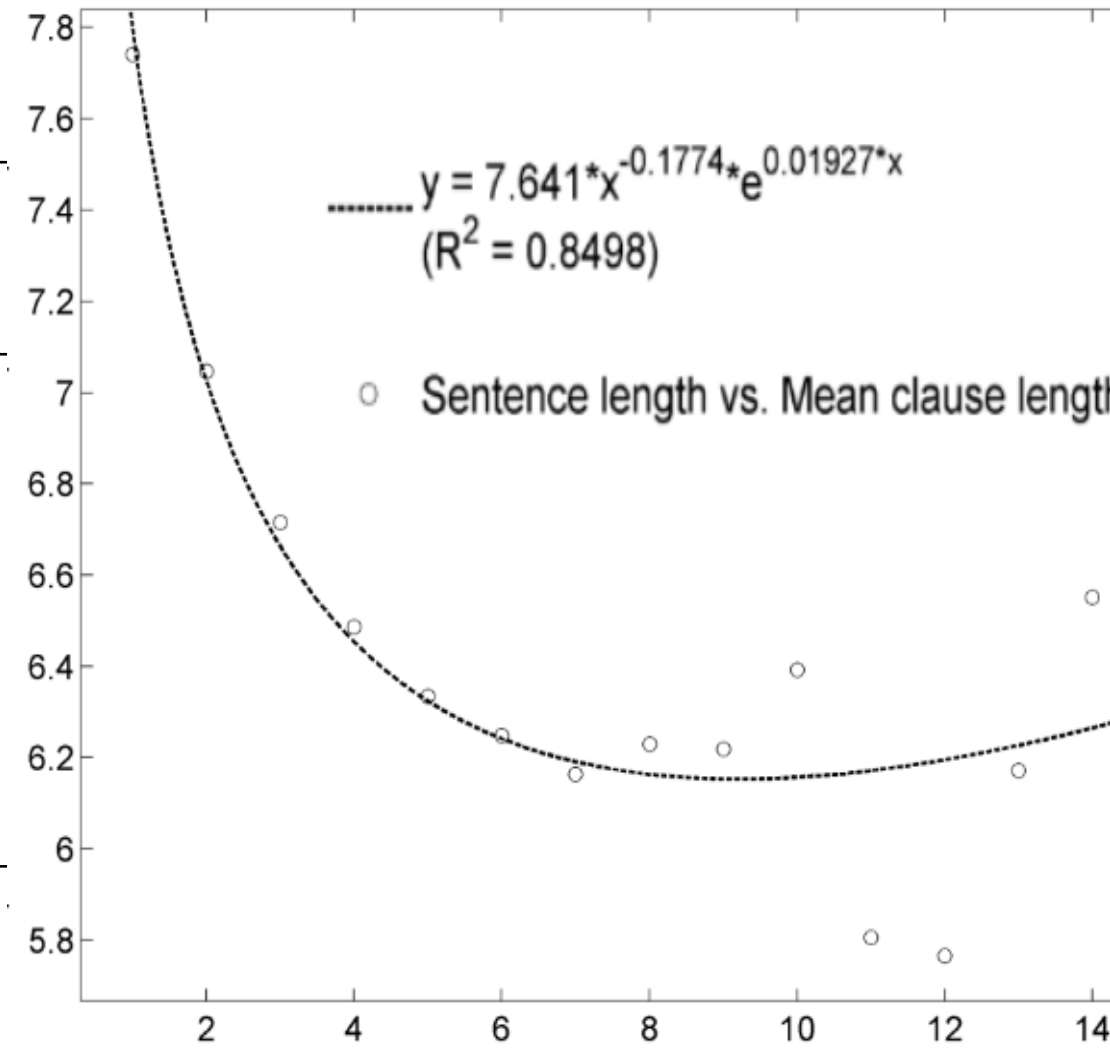
Methods



St-Sentence
Cl-Clause
Wd-Word
Ch-Character
Cp-Component
St-Stroke

Results: (1) Sentence> Clause >Word

ntence gth(in use)	Mean clause length(in word)	Sentence length(in clause)	Mean clause length(in word)
	7.7407	9	6.2194
	7.0465	10	6.3932
	6.7162	11	5.8068
	6.4866	12	5.7661
	6.3357	13	6.1723
	6.2485	14	6.5510
	6.1646	15	6.4500
	6.2296		



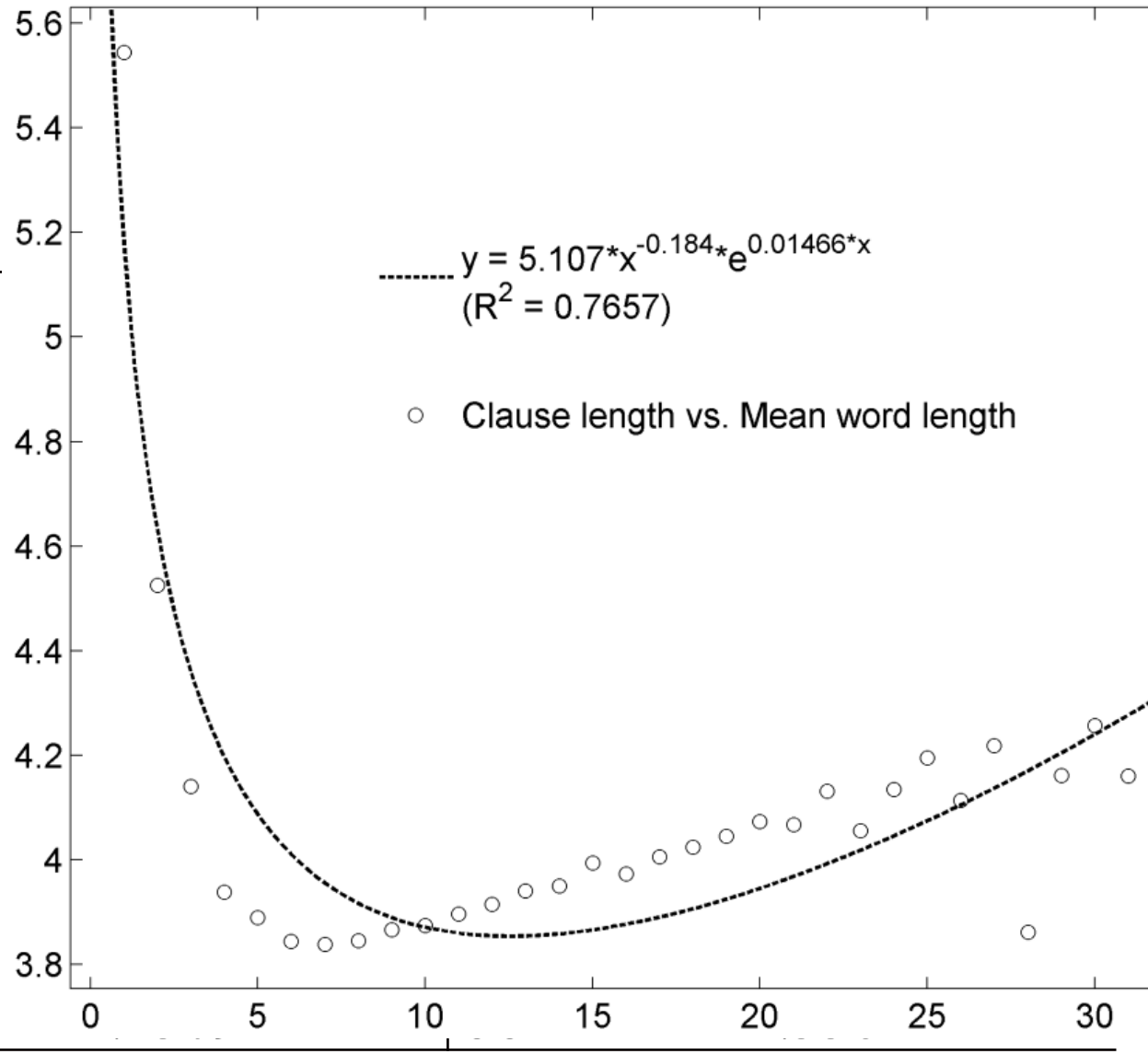
Results: (2) Clause>Word>Character

小句长（基于词）	平均词长（基于字）	小句长（基于词）	平均词长（基于字）
1	2.1777	26	1.5940
2	1.7501	27	1.6427
3	1.6281	28	1.5235
4	1.5565	29	1.6098
5	1.5378	30	1.6535
6	1.5189	31	1.5742
7	1.5170	32	1.5717
8	1.5187	33	1.6061
9	1.5258	34	1.6471
10	1.5263	35	1.4714
11	1.5326	36	1.8426
12	1.5381	37	2.0766
13	1.5441	38	1.7579

$R^2 = 0.08993$

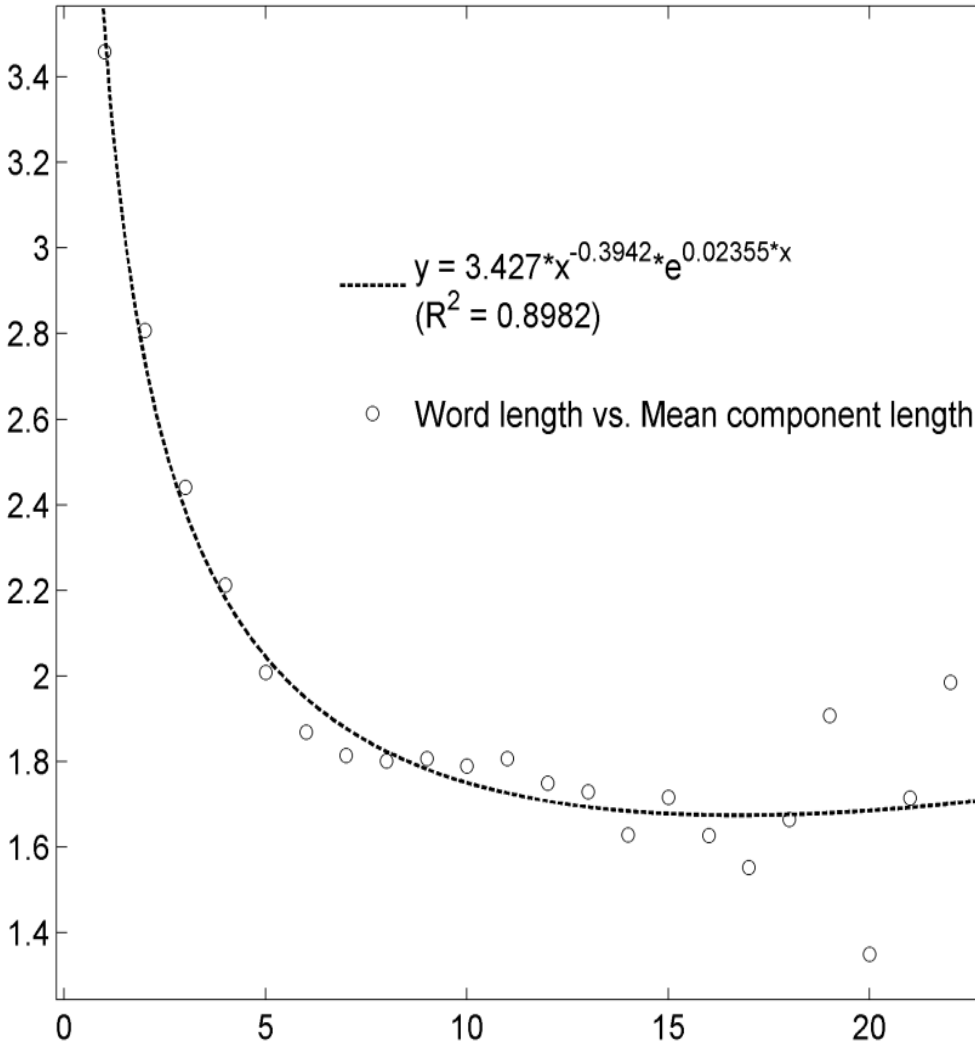
Results: (3) Clause> Word >Component

Clause length(in word)	Mean word length(in component)	Clause length(in word)
1	5.5445	12
2	4.5248	13
3	4.1405	14
4	3.9387	15
5	3.8897	16
6	3.8444	17
7	3.8383	18
8	3.8458	19
9	3.8657	20
10	3.8738	21
11	3.8966	22



Results: (4) word> component >stroke

ord gth(in mponent	Mean component length(in stroke)	Word length(in component)	Mean component length(in stroke)
	3.45959	13	1.72858
	2.80834	14	1.62894
	2.44086	15	1.71641
	2.21272	16	1.62715
	2.00806	17	1.55203
	1.86860	18	1.66435
	1.81350	19	1.90789
	1.80166	20	1.350
	1.80735	21	1.71428
	1.78970	22	1.98484
	1.80674	23	1.34782
	1.74935	25	1.960



Results: (5) word > character > component

Word length(in character)	Mean character length(in component)	Word length(in character)	Mean character length(in component)
1	2.4592	6	2.2054
2	2.5899	7	2.1860
3	2.5435	8	2.1354
4	2.5372	9	2.4222
5	2.1536	10	2.7000

R^2=0.1625

Results: (6) word > character > stroke

Word length(in character)	Mean character length(in stroke)	Word length(in character)	Mean character length(in stroke)
1	6.9359	6	6.1622
2	7.4136	7	6.2326
3	7.2189	8	6.2708
4	7.1969	9	6.5778
5	6.2356	10	6.4000

$$R^2 = 0.5009$$

Results

- The results shows that only "stroke > component > word", "component > word > clause" and "word > clause > sentence" line with Menzrath-Altmann law.

- sentence > clause > word > component > stroke

Discussions

- **Character** is an easy-to-distinguish language unit in written Chinese; **phrase** is commonly regarded as one level of language unit by grammarians. However, they are not included in the Menzerathian hierarchy.
- For **character**, the reason may be that although there are thousands of single-character words, they are not enough for communication. The combinations of characters into multi-character words makes ends meet. In **classic Chinese**, Character may be a basic language unit, however**, it is replaced by word in modern Chinese, because the classic Chinese habitually uses mono-syllable words while the modern Chinese prefers to choose multi-syllable words to express the same meaning**.

Discussions

- As for **phrase**
  - firstly, it is difficult to segment a sentence into several phrase sequences;
  - secondly, logically, two phrases can be combined into one phrase, which makes phrase not a basic language unit.

Conclusions

- That language is a system has been put forward for about 100 years, however, it has never been realized until quantification is introduced into linguistics.

- The Menzerath-Altmann law can be an efficient way of finding the basic language units in a language.

Conclusions

- some particular parameter values for some language units?

- tendency – if we go upwards in language unit hierarchy, parameter *b* (absolute value) is getting smaller.
  - b: 0.394 > 0.184 > 0.177

- In the future, we will investigate into this question from a diachronic perspective to see if the basic language units have changed with time.

THANK YOU!