The Chinese Script as a Self-regulating System

Applying Köhler's Basic Model of Synergetic Linguistics to Simplified Chinese Characters

Dr. Cornelia Schindelin

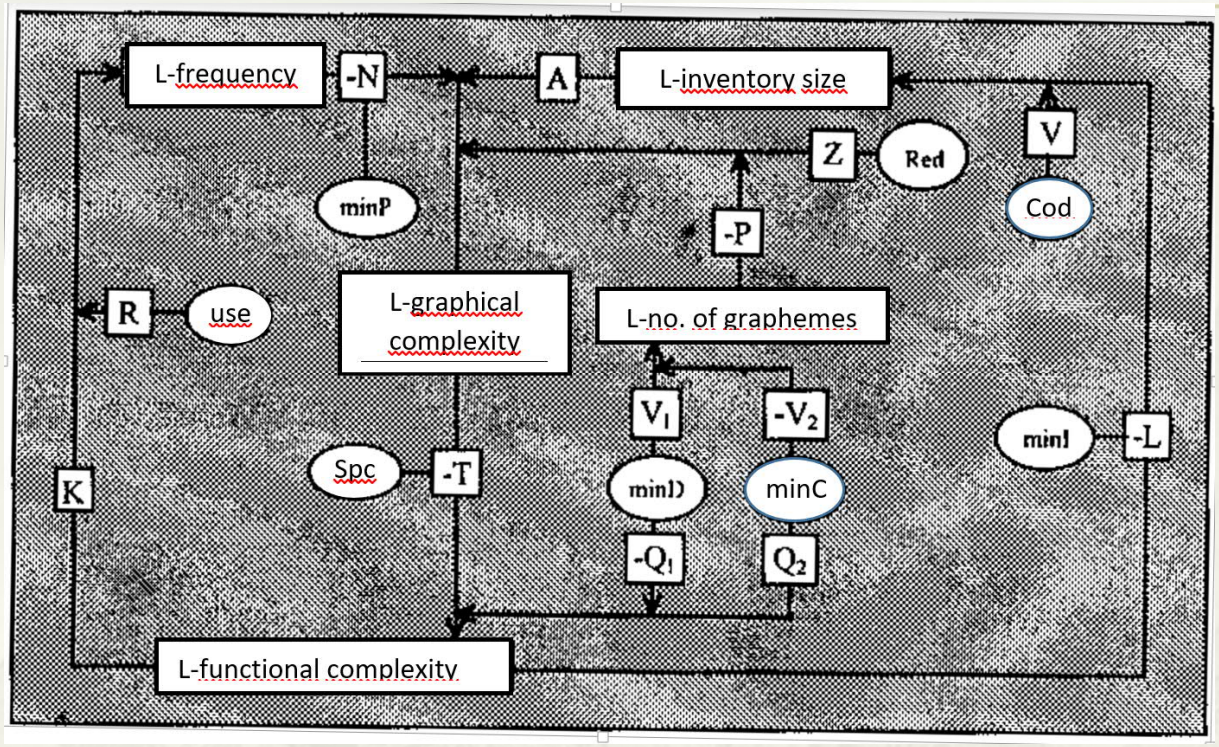
Center of Modern East Asian Studies
Chinese as a Foreign Language
Göttingen University

Faculty 06: Translation, Linguistics and Culture Studies
Dept. Of Chinese
Mainz University

What‘s it about?

- Köhler‘s Basic Model of Synergetic Linguistics
  - Variables, needs, relationships/dependencies
  - Direct and indirect dependencies → functions
- The Data
- Six Hypotheses
  - Three Hypotheses about direct functional dependencies
  - Three Hypotheses about indirect / mediated functional dependencies
- How did things come out?
- Any Conclusions?

Köhler‘s Basic Model of Synergetic Linguistics



- L- : logarithmized variable
- use: need to use a character
- minP: need to minimize producion effort
- Cod: need to encode
- Spc: need for specification
- Red: need for redundancy
- minC: need to minimize coding effort (writer)
- minD: need to minimize decoding effort (reader)
- minI: need to keep inventory size small/limited

- Inventory size: number of characters (types) used in the text corpus
- Number of (component) graphemes: number of different components available to make up the characters
- Graphical complexity: measured in a) strokes; b) components; c) weighted strokes→writing effort
- (Text or Token) Frequency: number of occurrences of each character in the corpus
- Functional complexity: number of different words the character is used in in the corpus

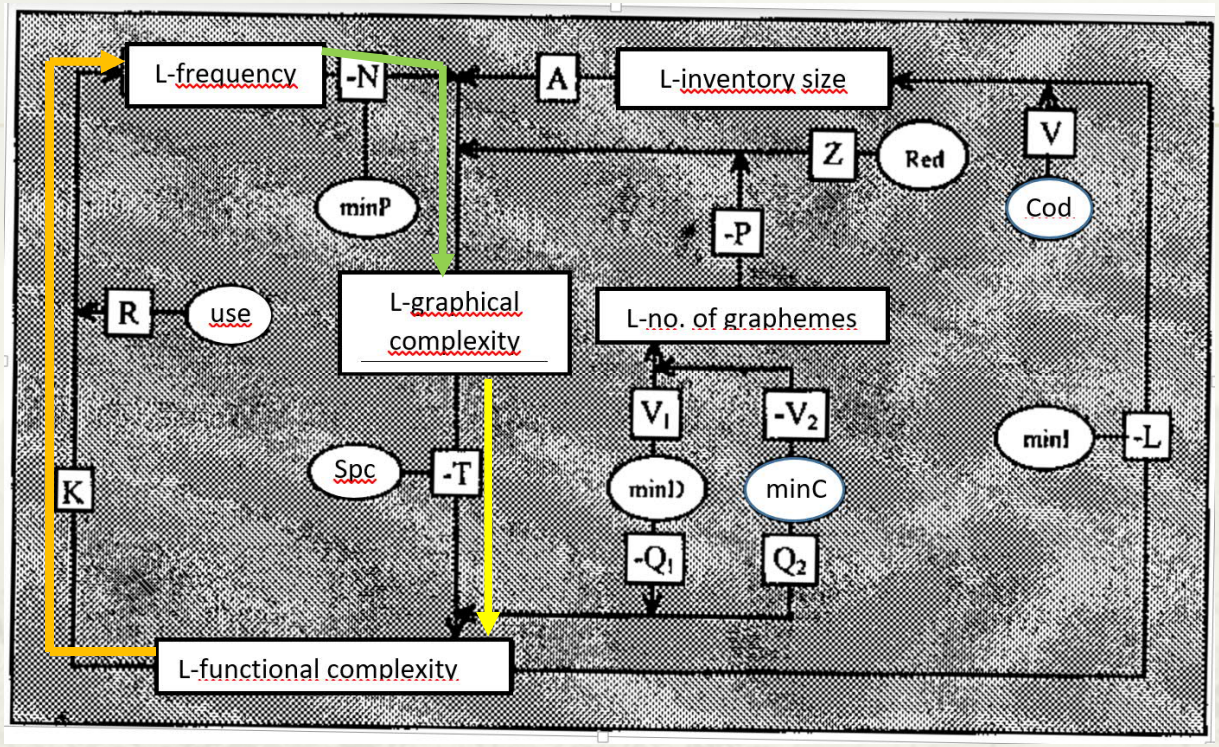
The Data: Source and Corpus

- [Frequency Dictionary of the Modern Chinese Language] Xiandai Hanyu pinlü cidian 现代汉语频率词典 (Beijing 1986)
- Factual prose (about 40 %), drama, fictional prose and essays as well as fairy-tales
- Corpus size in characters (token total): 1,808,114
- recruited from an inventory of 4,574 character types

Characters and Words

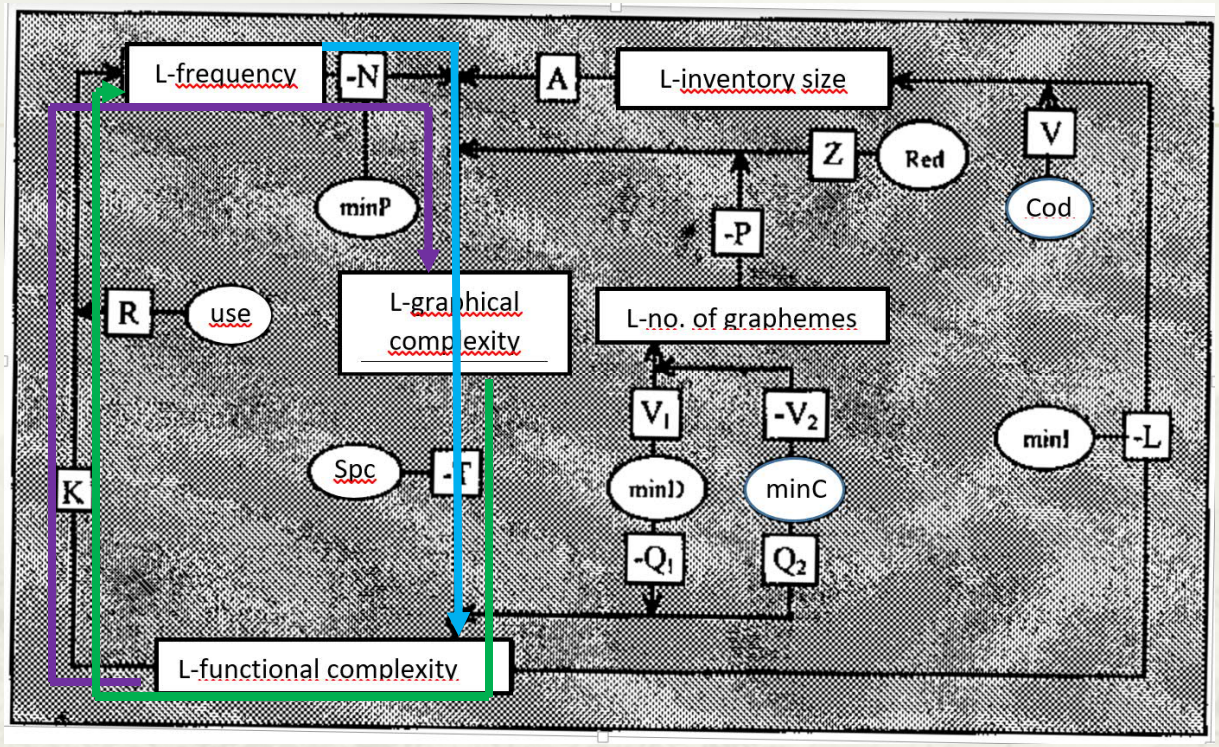
- 217 characters (=4.7 %) only write monosyllabic words
- 1,620 characters (=35.5 %) only write di- or polysyllabic words,
  - 519 only ever occur at the beginning of words,
  - 39 exclusively in the “middle” (which is not further specified) of words,
  - 433 exclusively at the end of words, and
  - 168 can appear in all three positions.
- 2,737 characters (= 59.8 %) appear in texts as representations of monosyllabic words as well as parts of longer words

Three Hypotheses about Direct Dependencies



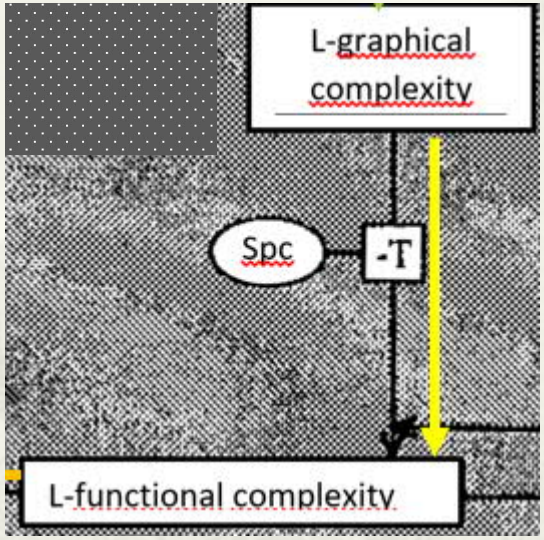
- H 1: The **functional complexity** of Chinese characters is directly a function of their **graphical complexity**. (yellow)
- H 2: The **text frequency** of Chinese characters is a function of their **functional complexity**. (orange)
- H 3: The **graphical complexity** of Chinese characters is a function of their **text frequency**. (green)

Three Hypotheses about Indirect Dependencies



- H 4: The **graphical complexity** of Chinese characters is indirectly a function of its **functional complexity**, mediated by frequency. (purple)
- H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity. (blue)
- H 6: The **text frequency** of characters is indirectly a function of their **graphical complexity**, mediated by functional complexity. (green)

Direct H 1: The **functional complexity** of Chinese characters is directly a function of their **graphical complexity.**



- 其 qí (pronoun; winnowing basket)
  - 箕 qí (winnowing basket)
- 來 lái (kind of wheat; Verb: to come)
  - 萊 lái (kind of wheat)

L-functional complexity = ln A + B * L-graphical complexity,
where B is expected to be negative

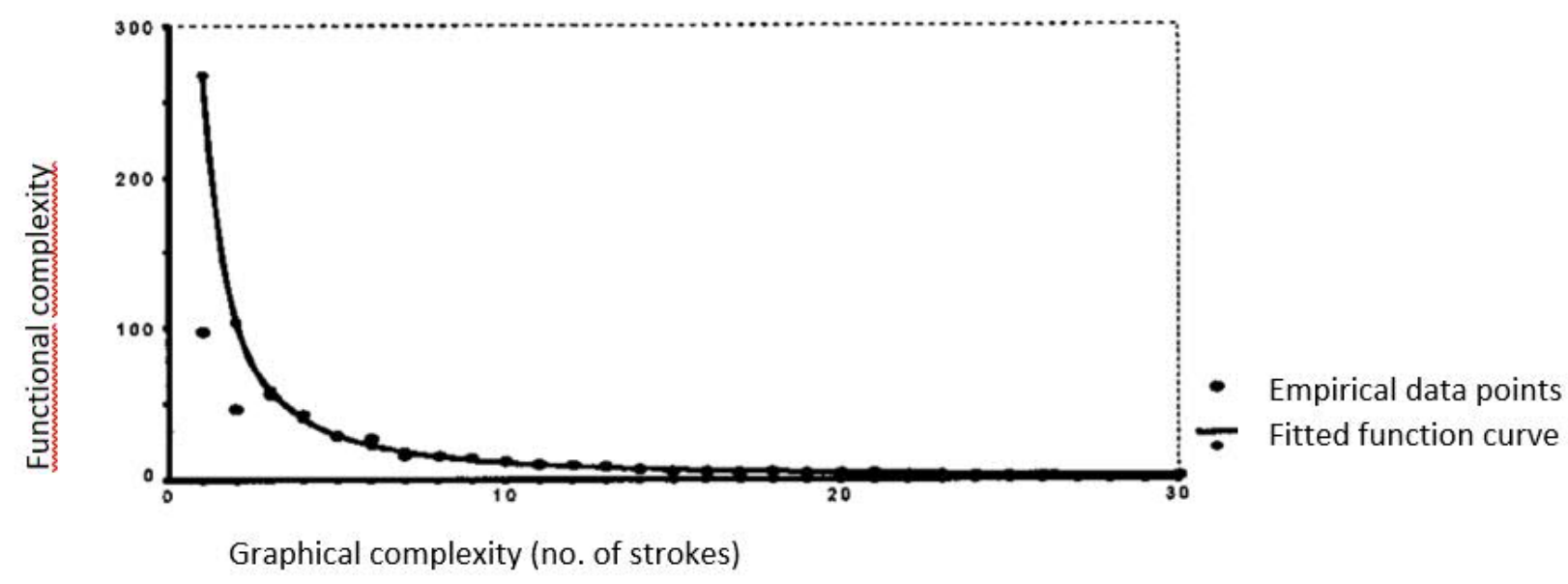
Power function: Functional complexity = A * Graphical complexity^{B}

Direct H 1: The **functional complexity** of Chinese characters is directly a function of their **graphical complexity.**

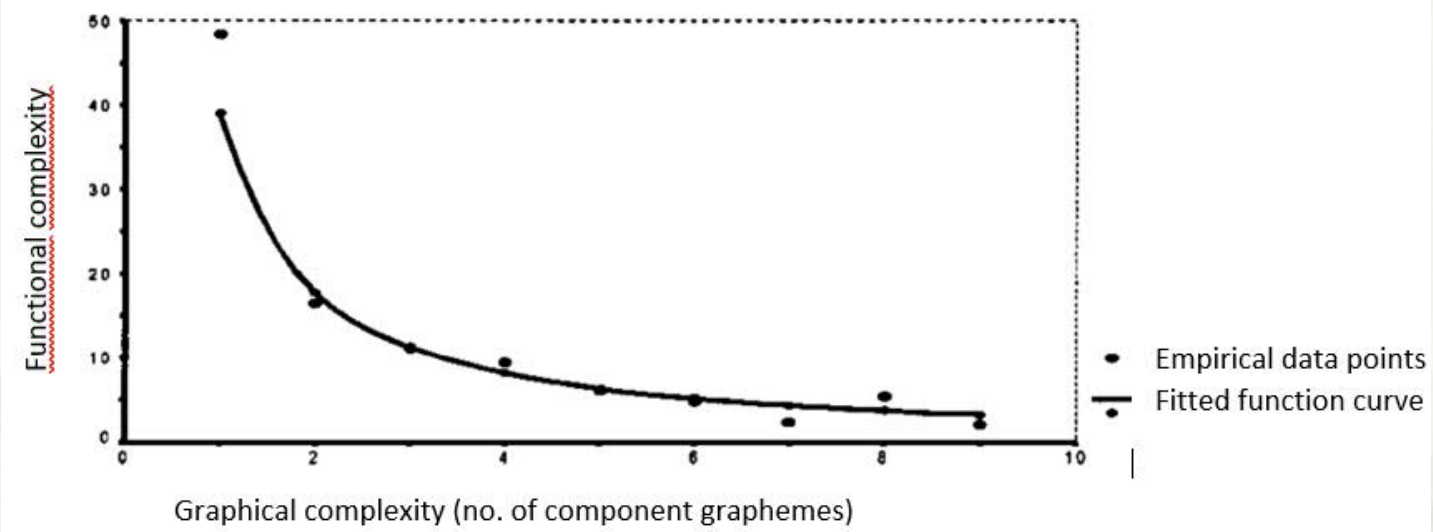
* L-functional complexity = ln A + B * L-graphical complexity,
  * where B is expected to be negative

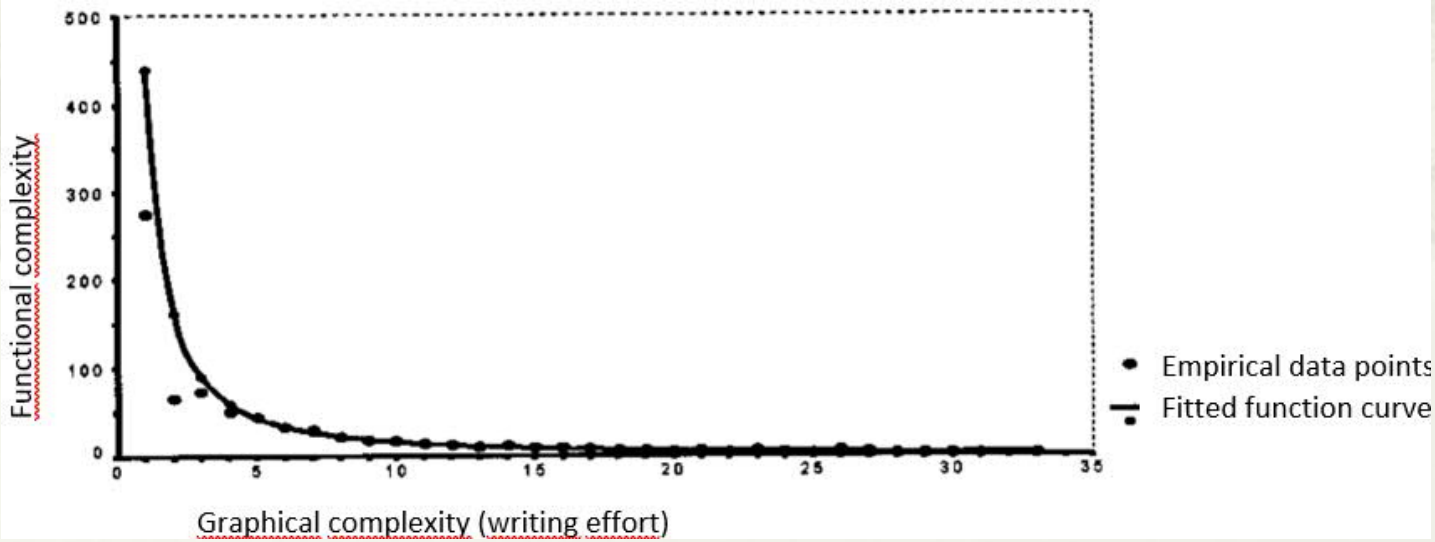
* Power function: Functional complexity = A * graphical complexity^{B}

a) Number of strokes: D = 0.956 A = $e^{5.59}$ = 268.12 B = - 1.373
b) Number of graphemes: D = 0.953 A = $e^{3.666}$ = 39.09 B = - 1.133
c) Writing effort: D = 0.95 A = $e^{6.086}$ = 439.72 B = - 1.44

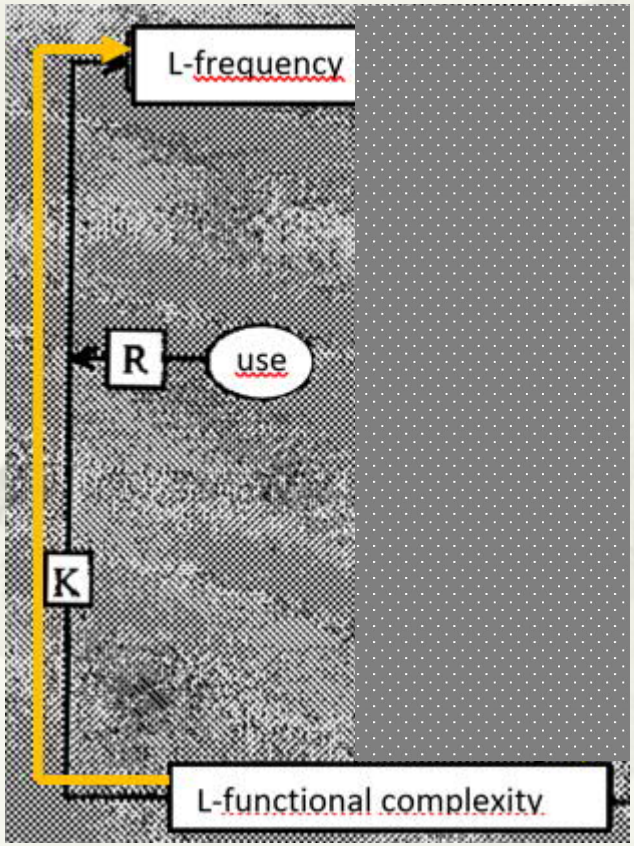


Direct H 1: The **functional complexity** of Chinese characters is directly a function of their **graphical complexity.**





Direct H 2: The **text frequency** of Chinese characters is a function of their **functional complexity.**



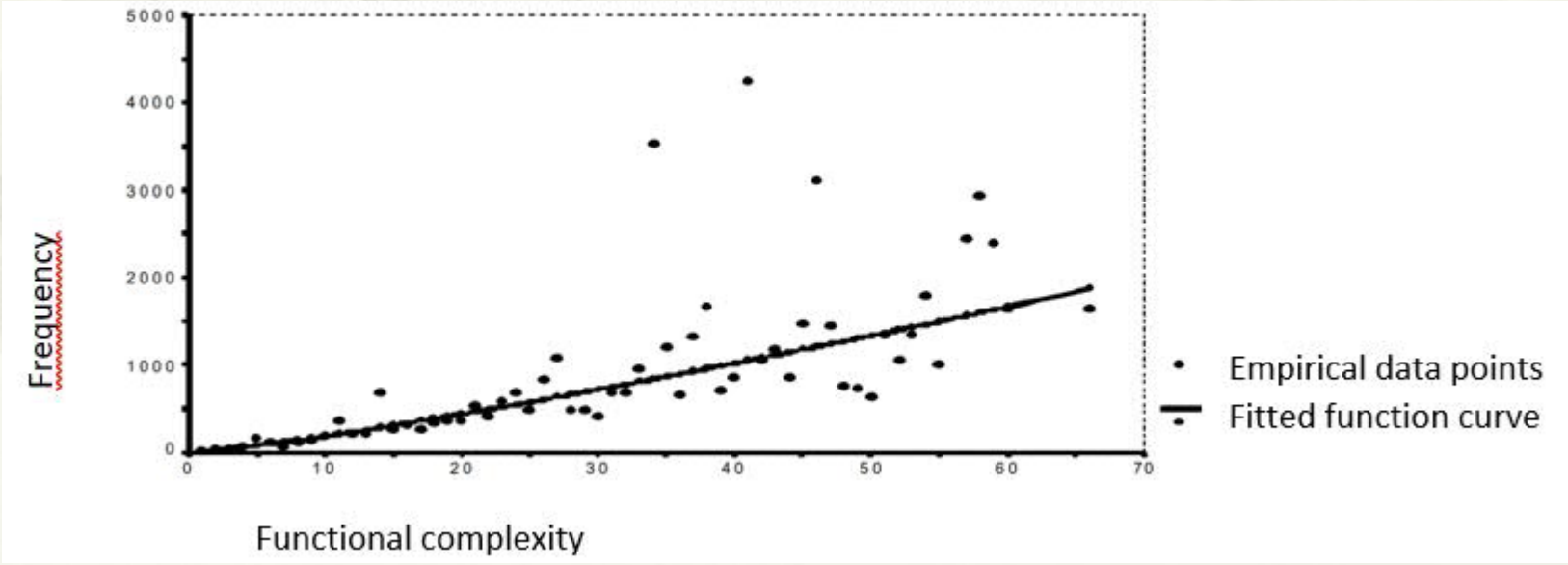
* L-frequency = ln A + B * L-functional complexity

* Power function:
  Frequency = A * Functional complexity^{B}

Direct H 2: The **text frequency** of Chinese characters is a function of their **functional complexity.**

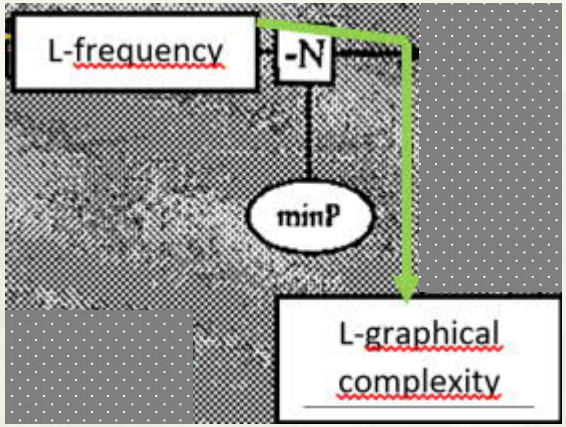
* L-frequency = ln A + B * L-functional complexity
* Power function: Frequency = A * Functional complexity^{B}

D = 0.958 A = $e^{2.444}$ = 11.52 B = 1,215



(Only data points with weights > 5 included.)

Direct H 3: The **graphical complexity** of Chinese characters is a function of their **text frequency**.



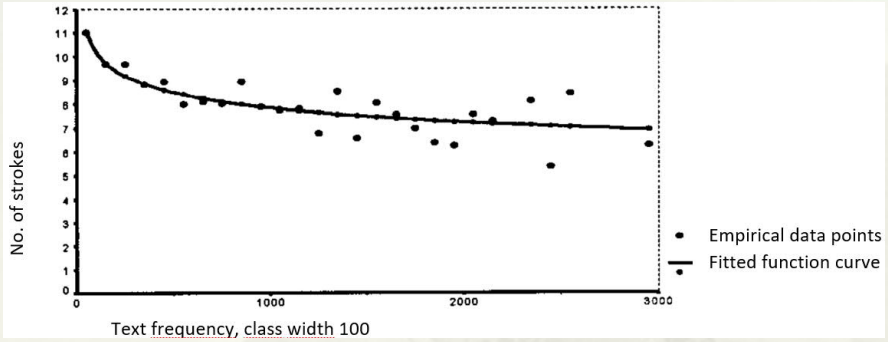
- L-graphical complexity = ln A + B * L-frequency
  - A negative value for B is expected
- Power function: Graphical complexity = A * Frequency^{B}

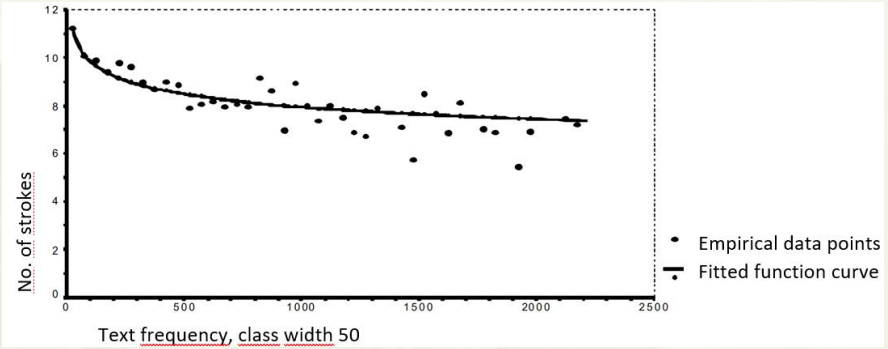
Direct H 3: The **graphical complexity** of Chinese characters is a function of their **text frequency**.

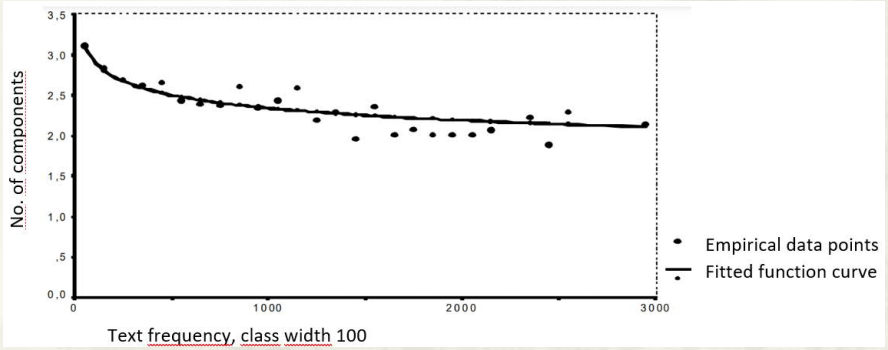
- L-graphical complexity = ln A + B * L-frequency
  - A negative value for B is expected
- Graphical complexity = A * Frequency^{B}

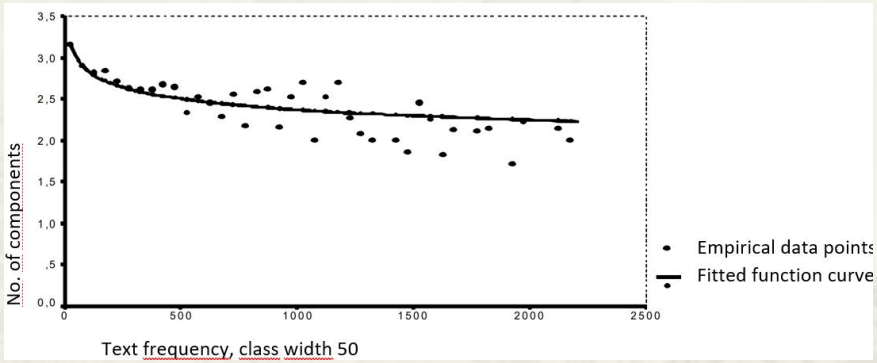
Way of measurement	Class width 100	Class width 50
a) Number of strokes	D = 0.94	D = 0.93
	A = $e^{2.846}$ = 17.22	A = $e^{2.72}$ = 15.18
	B = - 0.114	B = - 0.094
a) Number of graphemes	D = 0.95	D = 0.897
	A = $e^{1.51}$ = 4.53	A = $e^{1.4}$ = 4.066
	B = - 0.0958	B = - 0.078
a) Writing effort	D = 0.946	D = 0.92
	A = $e^{3.057}$ = 21.28	A = $e^{2.94}$ = 18.88
	B = - 0.11	B = - 0.09

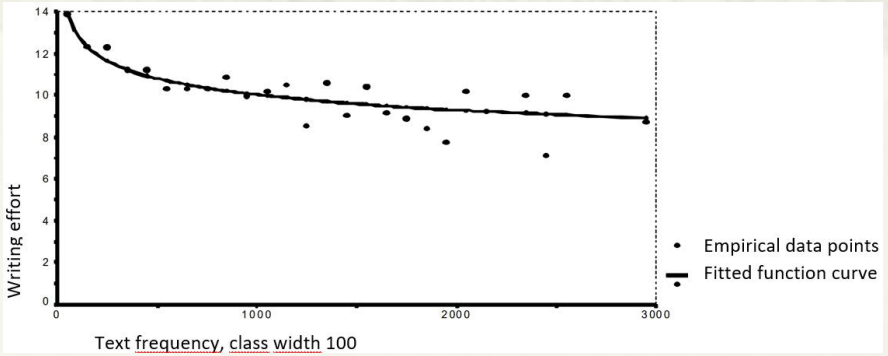
Direct H 3: The **graphical complexity** of Chinese characters is a function of their **text frequency**.

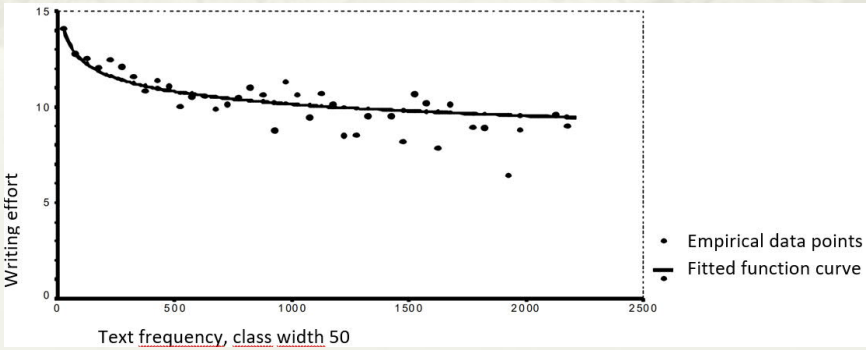




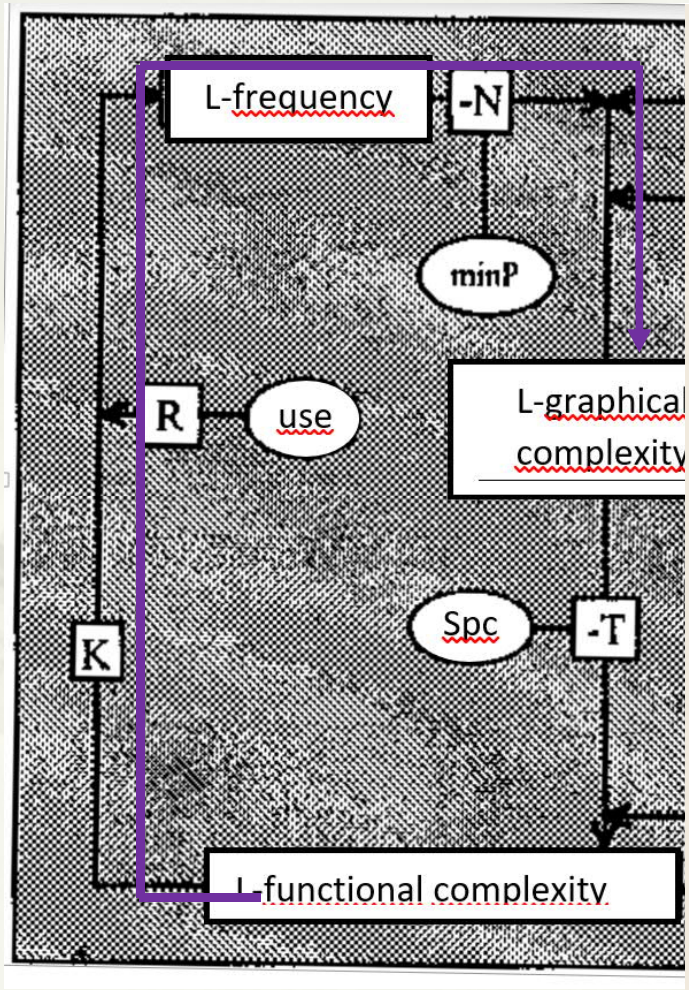








Indirect H 4: The **graphical complexity** of Chinese characters is indirectly a function of its **functional complexity**, mediated by frequency.



- L-graph. comp. = ln X + Y * L-funct. comp.
- Power function:

  Graph. comp. = A * funct. comp.^{B}

Indirect H 4: The **graphical complexity** of Chinese characters is indirectly a function of its **functional complexity**, mediated by frequency.

- L-graphical complexity = ln X + Y * L-functional complexity.

  As graphical complexity was measured in three ways and there were two class widths for frequency, we get six theoretical models:

- Graphical complexity measured in number of strokes
  - L-graphical complexity$_{a1}$ = 2.72 - 0.094 * (2.444 + 1.215 * L-functional complexity)
    = 2.49 - 0.114 * L-functional complexity

    and
  - L-graphical complexity$_{a2}$ = 2.85 - 0.114 * (2.444 + 1.215 * L-functional complexity)
    = 2.57 - 0.138 * L-functional complexity
- Graphical complexity measured in number of component graphemes
  - L-graphical complexity$_{b1}$ = 1.4 - 0.078 * (2.444 + 1.215 * L-functional complexity)
    = 1.2 - 0.095 * L-Functional complexity

    and
  - L-graphical complexity$_{b2}$ = 1.51 - 0.096 * (2.444 + 1.215 * L-functional complexity)
    = 1.277 - 0.116 * L-functional complexity
- Graphical complexity measured in writing effort
  - L-graphical complexity$_{c1}$ = 2.94 – 0.09 * (2,444 + 1,215 * L- functional complexity)
    = 2.72 – 0.109 * L-functional complexity

    and
  - L-graphical complexity$_{c2}$ = 3.06 - 0,109 * (2,444 + 1,215 * L- functional complexity)
    = 2.79 – 0.13 * L- functional complexity

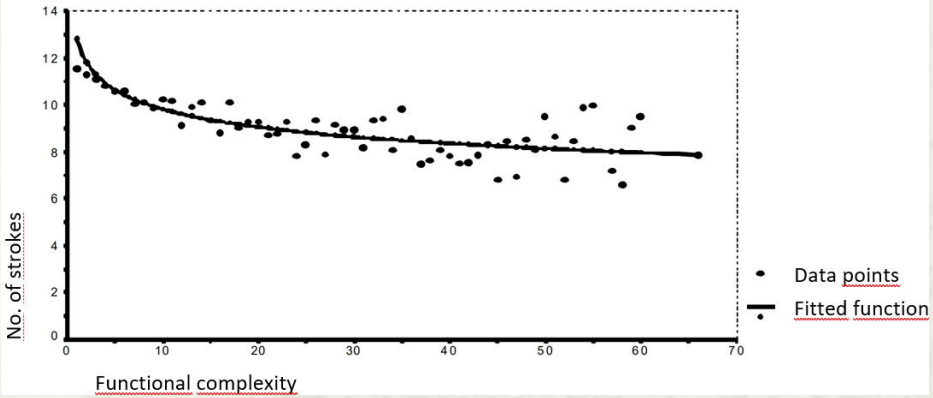
Indirect H 4: The **graphical complexity** of Chinese characters is indirectly a function of its **functional complexity**, mediated by frequency.

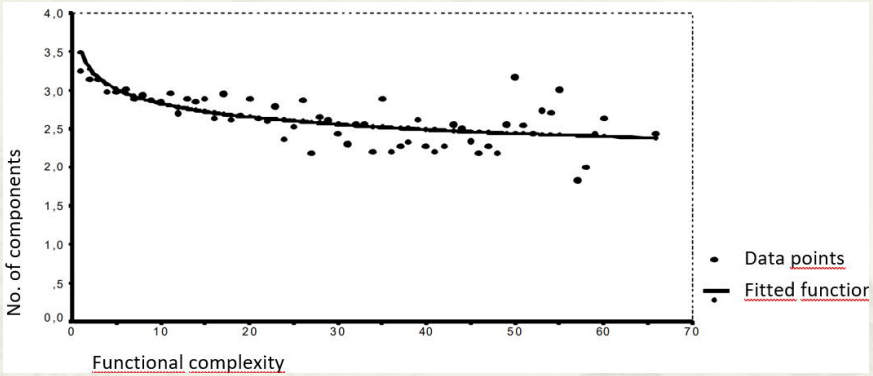
* The results of regression on the actual data were:

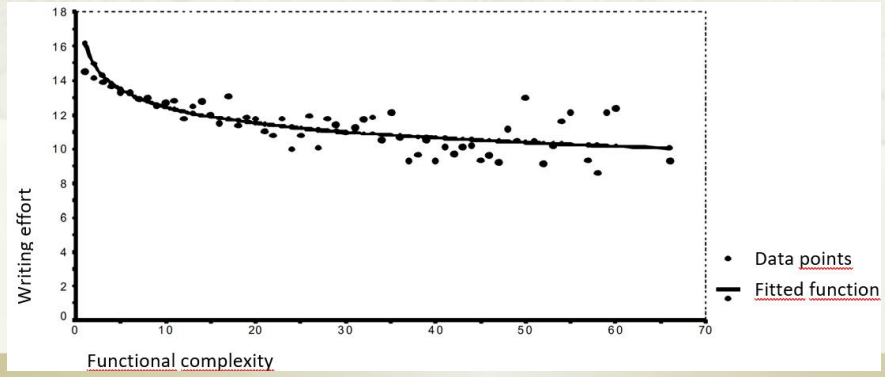
  a) Number of strokes: D = 0.73 $A = e^{2.55} = 12.82$ B = - 0.116

  b) Number of graphemes: D = 0.60 $A = e^{1.25} = 3.49$ B = - 0.092

  c) Writing effort: D = 0.75 $A = e^{2.78} = 16.19$ B = - 0.114





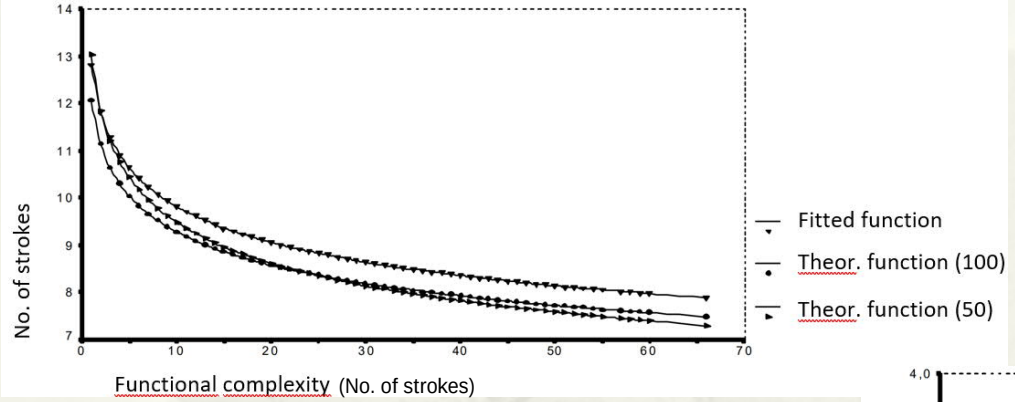


Indirect H 4: The **graphical complexity** of Chinese characters is indirectly a function of its **functional complexity**, mediated by frequency.

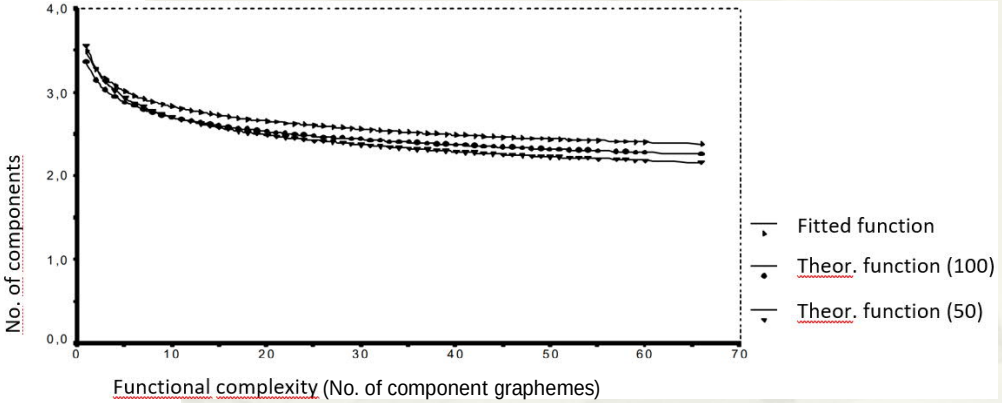
- Comparison between fitted functions and theoretical functions
- (“funct. comp.“ = functional complexity)
- Power function: Graphical complexity = A * functional complexity^{B}

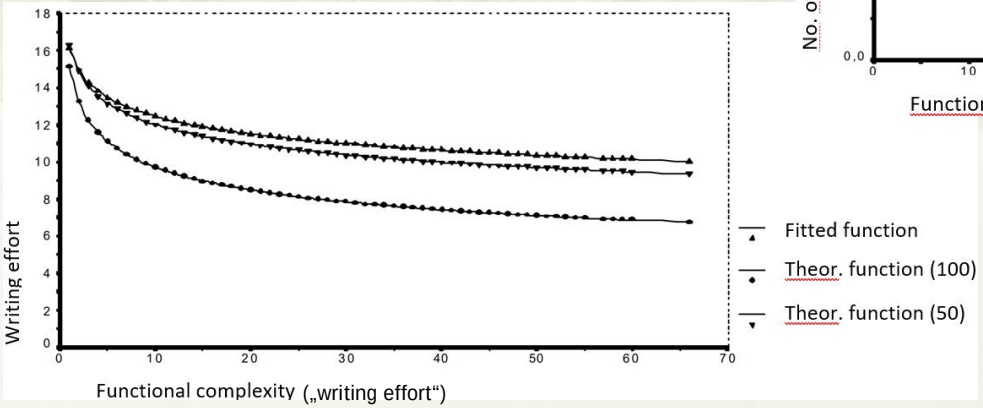
Way of measurement	Theoretical function	Empirical function
a) Number of strokes	Graph.comp.$_{a1}$ = 12.06 * funct. comp.$^{-0.114}$	Graph.comp.$_{ae}$ = 12.82 * funct. comp.$^{-0.116}$
	Graph.comp.$_{a2}$ = 13.04 * funct. comp.$^{-0.138}$	
a) Number of graphemes	Graph.comp.$_{b1}$ = 3.36 * funct. comp.$^{-0.095}$	Graph.comp.$_{be}$ = 3.49 * funct. comp.$^{-0.092}$
	Graph.comp.$_{b2}$ = 3.59 * funct. comp.$^{-0.116}$	
a) Writing effort	Graph.comp.$_{c1}$ = 15.16 * funct. comp.$^{-0.109}$	Graph.comp.$_{ce}$ = 16.19 * funct. comp.$^{-0.114}$
	Graph.comp.$_{c2}$ = 16.3 * funct. comp.$^{-0.13}$	

Indirect H 4: The **graphical complexity** of Chinese characters is indirectly a function of its **functional complexity**, mediated by frequency.

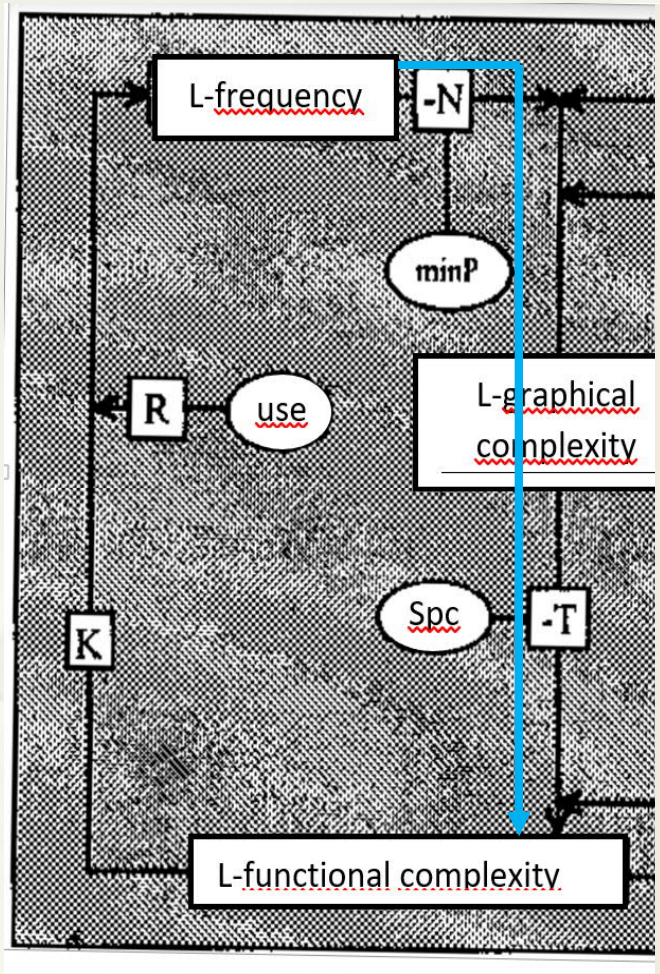


* Comparing fitted and theoretical functions (power functions)





Indirect H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity.



- L-functional complexity = ln X + Y * L-frequency
- Power function:

  Funct. Comp. = A * Frequency^{B}

Indirect H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity.

- L-functional complexity = ln X + Y * L-frequency

- L-functional complexity$_{a1}$ = 5.59 - 1.373 * (2.85 - 0.114 * L-frequency)
  = 1.68 + 0.156 * L-frequency

  and

  L-functional complexity$_{a2}$ = 5.59 - 1.373 * (2.72 - 0.094 * L-frequency)
  = 1.85 + 0.13 * L-frequency
- L-functional complexity$_{b1}$ = 3.666 - 1.133 * (1.51 - 0.096 * L-frequency)
  = 1.95 + 0.108 * L-frequency

  and

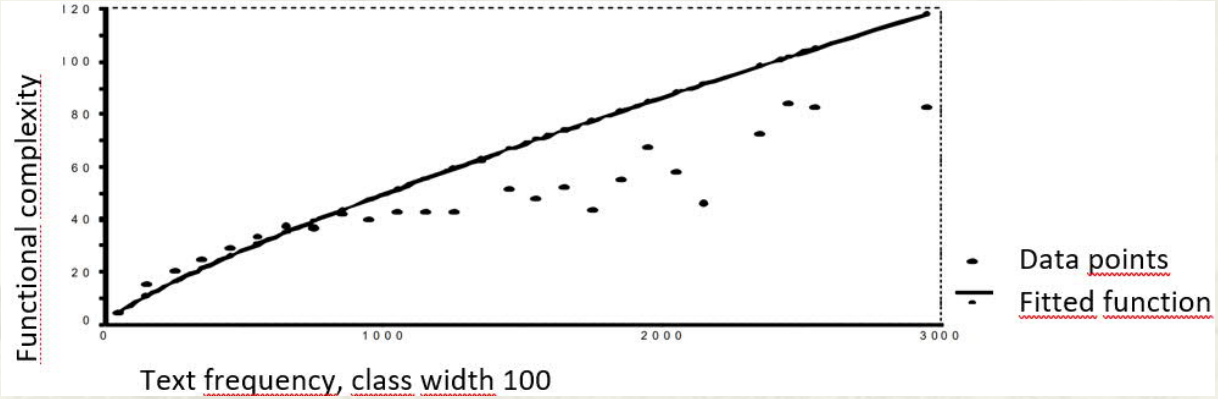
  L-functional complexity$_{b2}$ = 3.666 - 1.133 * (1.4 - 0.078 * L-frequency)
  = 2.076 + 0.088 * L-frequency
- L-functional complexity$_{c1}$ = 6.086 - 1.441 * (3.06 - 0.109 * L-frequency)
  = 1.68 + 0.157 * L-frequency

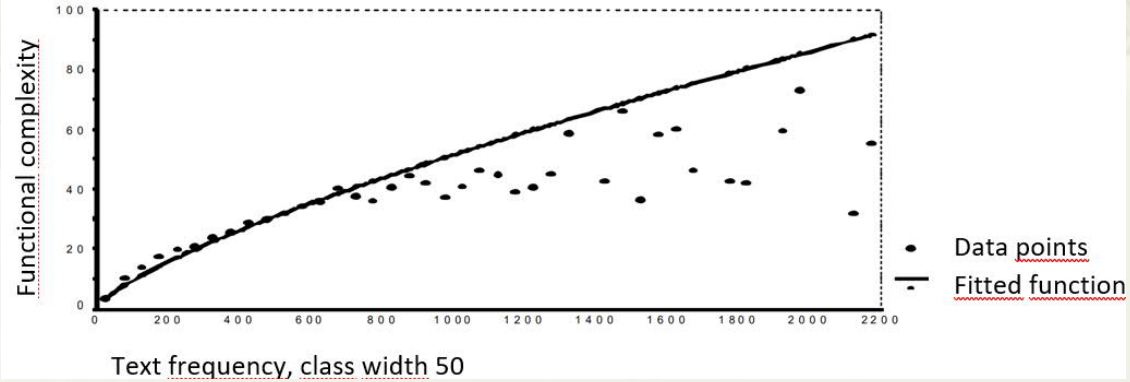
  and

  L-functional complexity$_{c2}$ = 6.086 - 1.441 * (2.94 - 0.09 * L-frequency)
  = 1.85 + 0.13 * L-frequency

Indirect H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity.

- Power function: Funct. Comp. = A * Frequency^{B}
- Regression results:
  - Class width 100: D = 0.969 A = $e^{-1.649}$ = 0.192 B = 0.804
  - Class width 50: D = 0.97 A = $e^{-1.173}$ = 0.31 B = 0.74





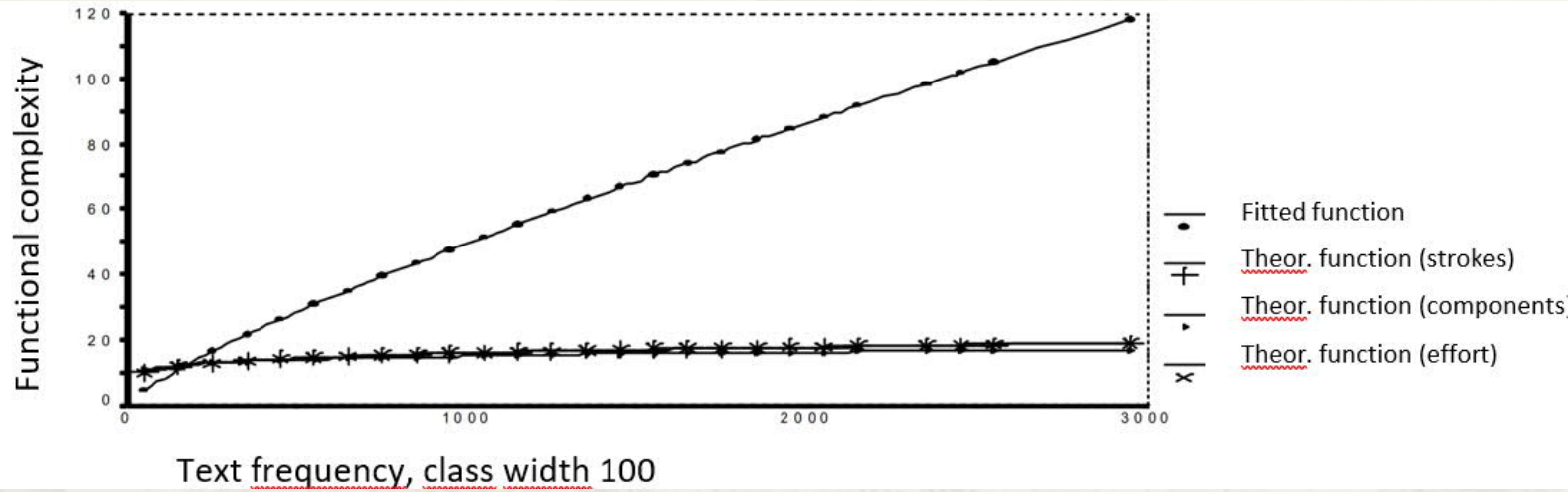
Indirect H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity.

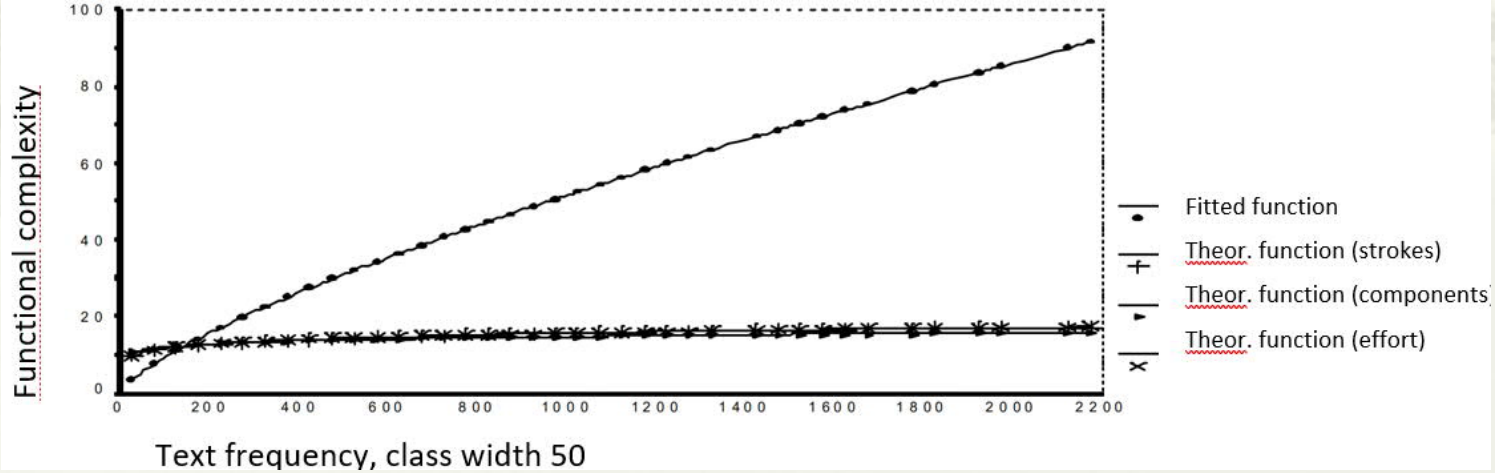
- Comparison between fitted functions and theoretical functions

Class width	Theoretical functions	Empirical function
100	funct. $comp._{a1} = 5.37 * freq.^{0.156}$	funct. $comp._{e1} = 0.192 * freq.^{0.804}$
	funct. $comp._{b1} = 7.05 * freq.^{0.108}$	
	funct. $comp._{c1} = 5.36 * freq.^{0.157}$	
50	funct. $comp._{a2} = 6.36 * freq.^{0.13}$	funct. $comp._{e2} = 0.31 * freq.^{0.74}$
	funct. $comp._{b2} = 7.98 * freq.^{0.088}$	
	funct. $comp._{c2} = 6.36 * freq.^{0.13}$	

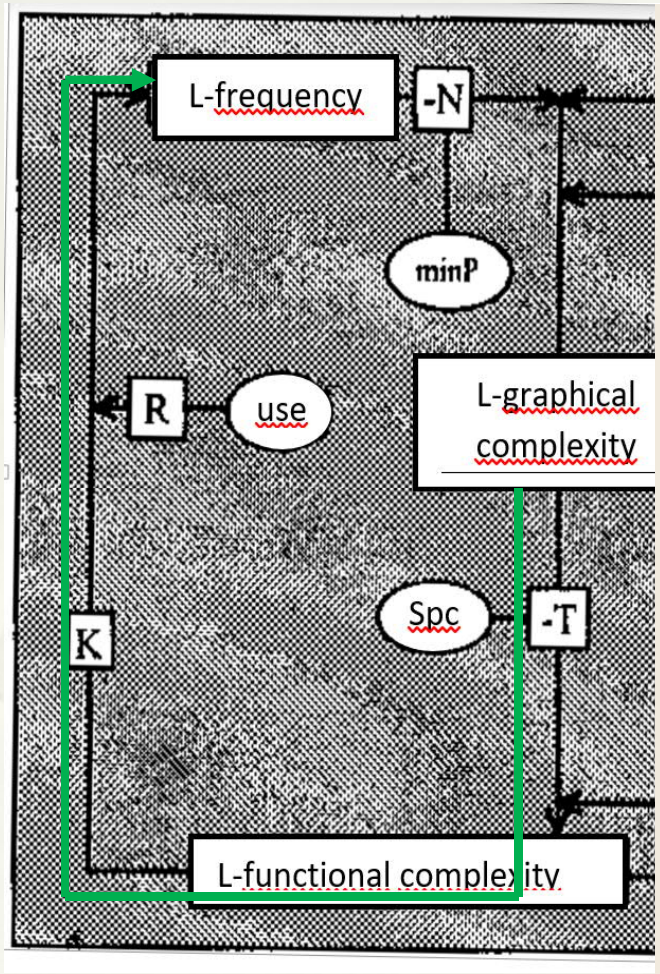
Indirect H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity.

* Comparison between fitted function and theoretical functions





H 6: The **text frequency** of Chinese characters is indirectly a function of their **graphical complexity**, mediated by functional complexity.



- L-frequency = ln X + Y * L-graphical complexity
- Power function:

  Frequency = A * Graphical complexity^{B}

H 6: The **text frequency** of Chinese characters is indirectly a function of their **graphical complexity**, mediated by functional complexity.

* L-frequency = ln X + Y * L-graphical complexity

a) Number of strokes $L\text{-}freq_a = 2.444 + 1.215 * (5.59 - 1.373 * L\text{-}graph.comp.)$
$= 9.24 - 1.67 * L\text{-}graph.comp.$

b) No. of graphemes $L\text{-}freq_b = 2.444 + 1.215 * (3.666 - 1.133 * L\text{-}graph.comp.)$
$= 6.9 - 1.377 * L\text{-}graph.comp.$

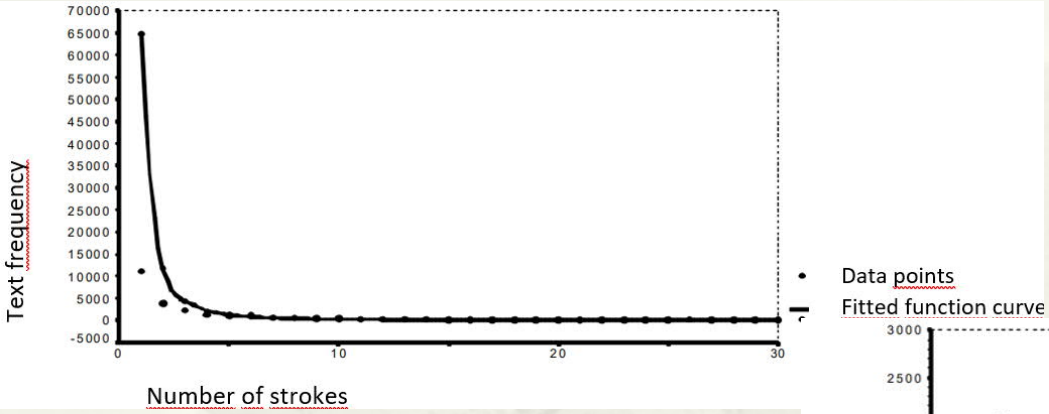
c) Writing effort $L\text{-}freq_c = 2.444 + 1.215 * (6.086 - 1.441 * L\text{-}graph.comp.)$
$= 9.84 - 1.75 * L\text{-}graph.comp.$

* Power function: Frequency = A * Graphical complexity^B

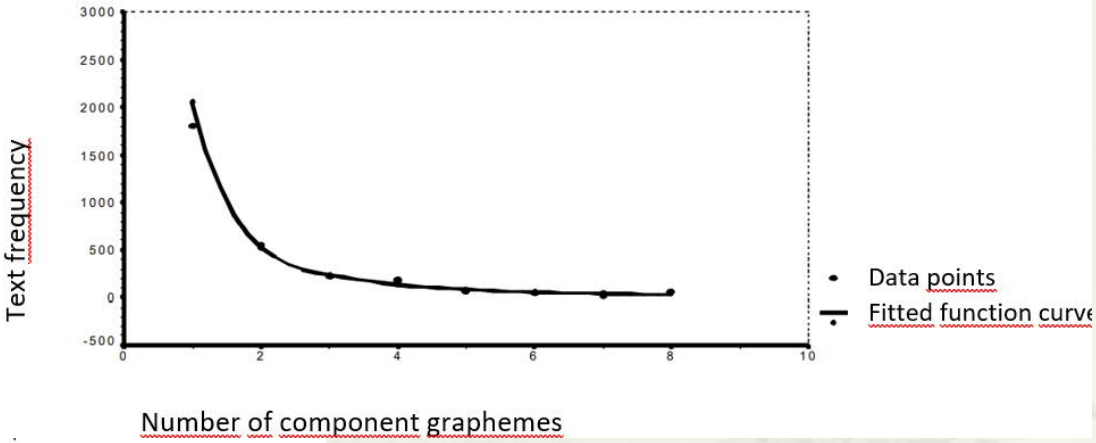
* Regression results:

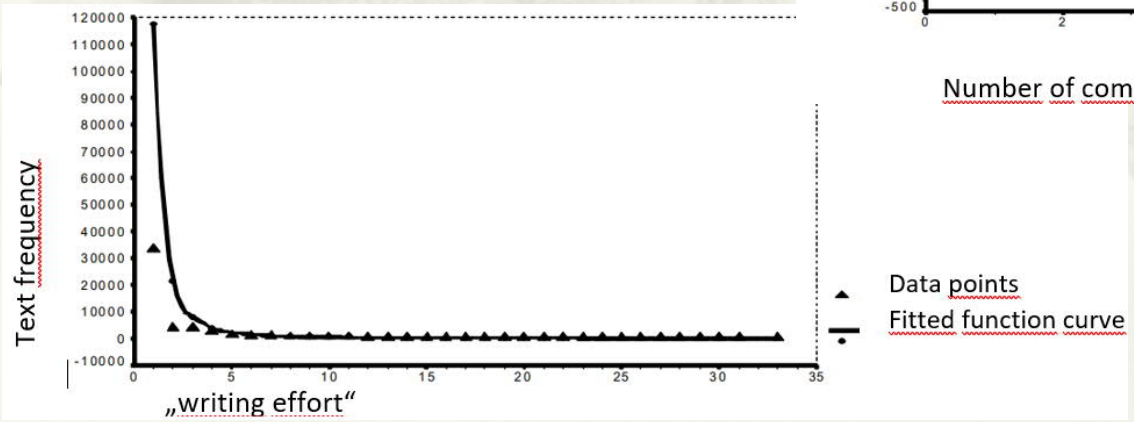
a) Number of strokes	D = 0.93	$A = e^{11.077} = 64.690.26$	B = -2.466
b) Number of graphemes	D = 0.955	$A = e^{7.63} = 2.058.5$	B = -1.98
c) Writing effort	D = 0.88	$A = e^{11.675} = 117,557.75$	B = -2.47

H 6: The **text frequency** of Chinese characters is indirectly a function of their **graphical complexity**, mediated by functional complexity.



- Fitted functions (power functions)



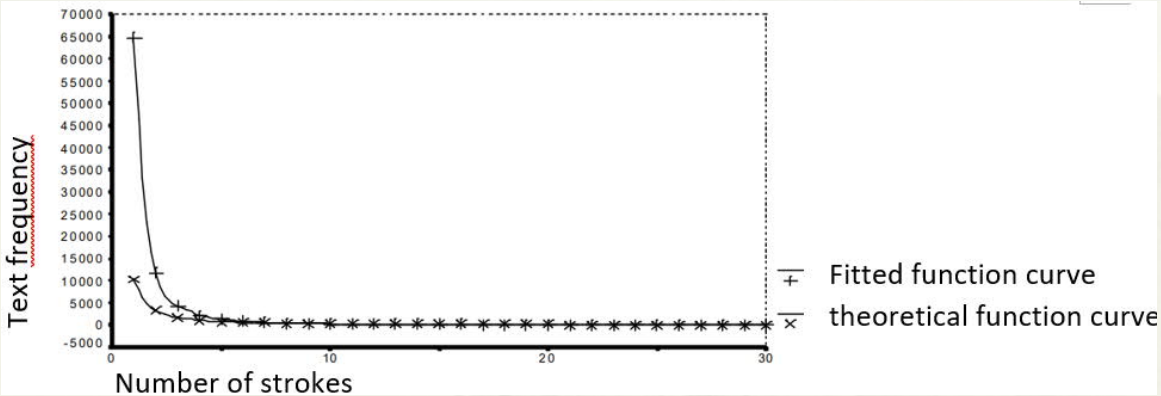


H 6: The **text frequency** of Chinese characters is indirectly a function of their **graphical complexity**, mediated by functional complexity.

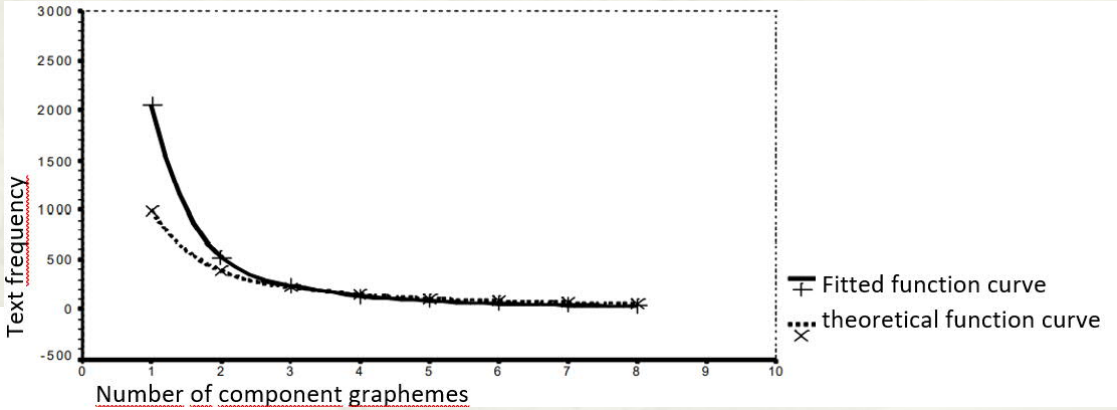
- Comparison between fitted functions and theoretical functions

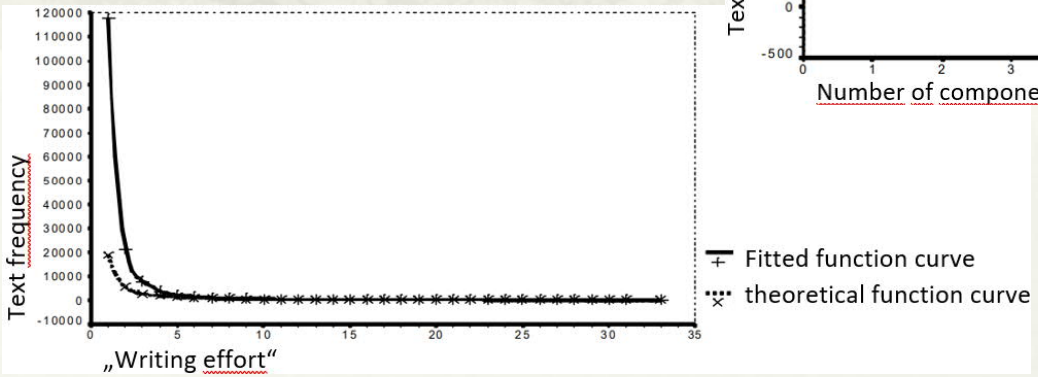
	Theoretically	Empirically
a)	$Freq_a = 10{,}287.14 * Komp^{-1.67}$	$Freq_{ea} = 64{,}690.26 * Komp^{-2.466}$
b)	$Freq_b = 992.27 * Komp^{-1.377}$	$Freq_{eb} = 2{,}058.5 * Komp^{-1.98}$
c)	$Freq_c = 18{,}797.89 * Komp^{-1.75}$	$Freq_{ec} = 117{,}557.75 * Komp^{-2.47}$

H 6: The **text frequency** of Chinese characters is indirectly a function of their **graphical complexity**, mediated by functional complexity.



- Comparison between fitted functions and theoretical functions (power functions)



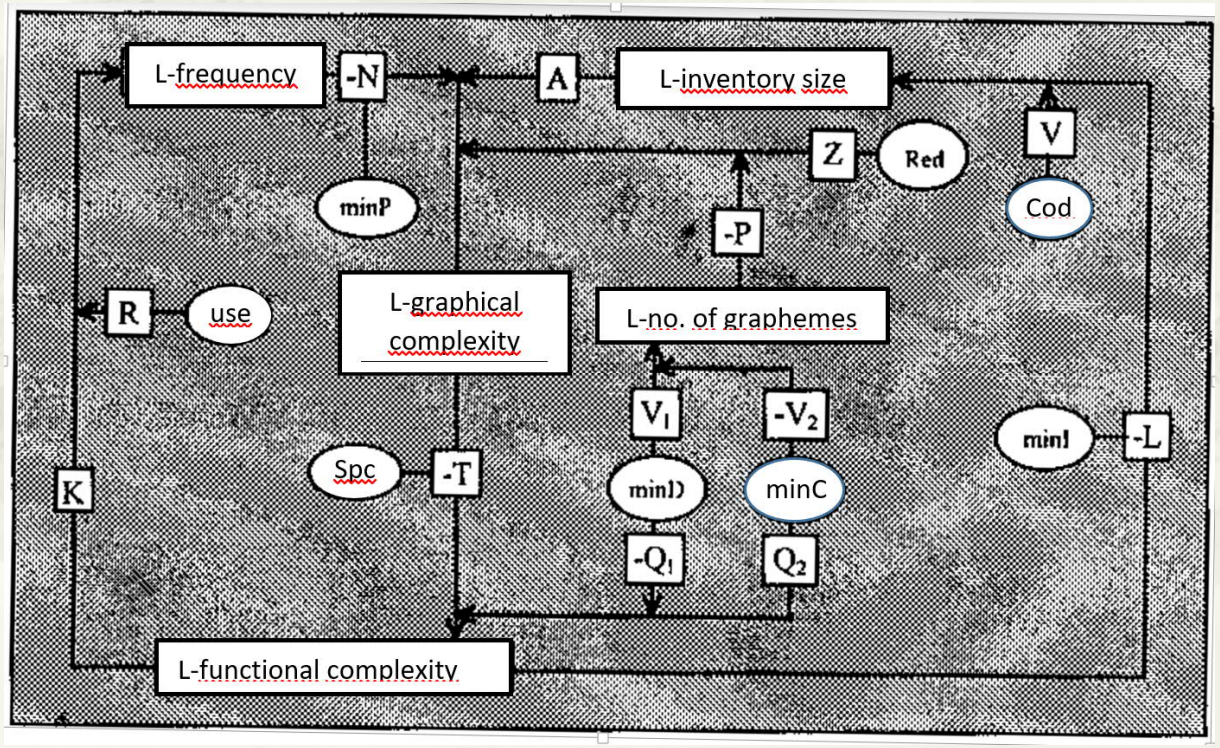


Any Conclusions?

- Three direct hypotheses:
  - Regression very good, can be accepted
- Three indirect hypotheses:
  - Only H 6 withstood testing
  - H4 and H 5 could not be validated on the data, seem to show systematic deviation. Factor involved that has not been considered, yet?
  - Step in right direction?
- Overall, relationships not very different than in the model for vocabularies.

Thank you for listening!

谢谢，请多关照！



Literature

Altmann, G. (2004). “Script Complexity”. In: Glottometrics 8. 68–74.

Altmann, G., and Köhler, R. (1996). "Language Forces? and Synergetic Modelling of Language Phenomena”. In: P. Schmidt [ed.]: Glottometrika 15. Issues in General Linguistic Theory and the Theory of Word Length. WVT. Trier. 62-76.

Best, Karl-Heinz/Zhu, Jinyang (2001). „Wortlängenverteilungen in chinesischen Texten und Wörterbüchern“. In: Karl-Heinz Best (ed.). Häufigkeitsverteilungen in Texten. Göttingen: Peust & Gutschmitdt. 101-114.

Breiter, Maria A. (1994). „Length of Chinese words in relation to their other systemic features”. In: Journal of quantitative linguistics 1(3). 224-231.

Bohn, H. (1998). Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift [Quantitative studies of modern Chinese language and script]. Hamburg: Verlag Dr. Kovač.

[Frequency Dictionary of the modern Chinese Language] Xiandai Hanyu pinlü cidian 现代汉语频率词典 (1986). Beijing: Yuyan Xueyuan chubanshe.

Grotjahn, R. (1992). “Evaluating the adequacy of regression models: Some potential pitfalls”. In: B. Rieger (ed.). Glottometrika 13 (S. 121-172). Bochum: Brockmeyer.

Hammerl, R. (1991). Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells [Studies on the structure of the lexicon: Construction of a lexical basic model]. Trier: Wissenschaftlicher Verlag Trier.

Köhler, Reinhard (1986). Zur linguistischen Synergetik. Struktur und Dynamik der Lexik [On synergetic linguistics. Structure and dynamics of the lexicon]. Bochum: Brockmeyer.

Köhler, Reinhard (1990). „Elemente der synergetischen Linguistik“ [Elements of synergetic linguistics]. In: Glottometrika 12: 179-188.

Köhler, Reinhard (2005). „Synergetic Linguistics“. In: Reinhard Köhler. Gabriel Altmann & Rajmund G. Piotrowski (eds.). Quantitative Linguistics. An International Handbook. Berlin. New York: Walter de Gruyter. 760-775.

Köhler, R., and S. Naumann (2005). „An extension of the synergetic-linguistic model and its application to word frequency”. In: P. Grzybek (ed.) Proceedings of the Workshop The Science of Language: Structures of Frequencies and Relations at Graz University.

Le, Quan Ha/Sicilia-Garcia, E. I./Ji, Ming/Smith, F.J. (2002). “Extension of Zipf's law to words and phrases”. In: Proceedings of the 19th international conference on computational linguistics (COLING-2002). Taibei, (no pages given).

Le, Quan Ha/Sicilia-Garcia, E. I./Ji, Ming/Smith, F.J. (2003): “Extension of Zipf's law to word and character n-grams for English and Chinese”. In: Computational linguistics and Chinese language processing 8(1). 77-102.

Menzel, Cornelia (2004): „Das synergetische Basismodell der Lexik und die chinesische Schrift“ [The synergetic basic model of lexics and the Chinese script]. In: Reinhard Köhler (ed.). Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik [Corpus studies for quantitative and systems theoretic linguistics]. 178-205. Trier. online. URN: https://nbn-resolving.org/urn:nbn:de:hbz:385-1-1469.

Rousseau, R[onald]., and Zhang, Qiaoqiao (1992): “Zipf's data on the frequency of Chinese words revisited”. In: Scientometrics 24(2). 201-220.

Schindelin, Cornelia (2005). "Die quantitative Erforschung der chinesischen Sprache und Schrift" [Quantitative research on the Chinese language and script]. In: Quantitative Linguistics. An International Handbook. Edited by Reinhard Köhler. Gabriel Altmann. Rajmund G. Piotrowski. Berlin. New York: Walter de Gruyter. 947-970.

Schindelin, Cornelia (2017). "Character Frequency". In: Encyclopedia of Chinese Language and Linguistics. Vol. 1. Ed. by Rint Sybesma et al. Leiden: Brill. 358-362.

Schindelin, Cornelia (2017). "Menzerath's Law". In: Encyclopedia of Chinese Language and Linguistics. Vol. 3. Ed. by Rint Sybesma et al. Leiden: Brill. 1-3.

Schindelin, Cornelia (2017). "Word Frequency". In: Encyclopedia of Chinese Language and Linguistics. Vol. 4. Ed. by Rint Sybesma et al. Leiden: Brill. 580-584.

Schindelin, Cornelia (2017). "Word Length". In: Encyclopedia of Chinese Language and Linguistics. Vol. 4. Ed. by Rint Sybesma et al. Leiden: Brill. 584-589.

Schindelin, Cornelia (2017). "Zipf's Law". In: Encyclopedia of Chinese Language and Linguistics. Vol. 4. Ed. by Rint Sybesma et al. Leiden: Brill. 723-724.

Wang, L. (2011). “Polysemy and Word Length in Chinese”. In: Glottometrics. 22. 73–84.

Wang, L. (2014). “Synergetic Studies on Some Properties of Lexical Structures in Chinese”. In: Journal of Quantitative Linguistics. 21(2). 177–197.

Wang, Yanru, and Xinying Chen (2015). “Structural Complexity of Simplified Chinese Characters”. In: Tuzzi. Arjuna. Martina Benešová. and Ján Mačutek (eds.). Recent Contributions to Quantitative Linguistics. Berlin: de Gruyter Mouton. 229-240.

Zhu, Jinyang, and Best, Karl-Heinz (1997). „Zur Modellierung der Wortlängen im Chinesischen“ [On modelling word length in Chinese]. In: Glottometrika 16. Trier: Wissenschaftl. Verlag Trier. 185-194.

Zhu, Jinyang, and Best, Karl-Heinz (1998). „Wortlängenhäufigkeiten in chinesischen Kurzgeschichten“ [Word length frequencies in Chinese short stories]. In: Asian and African Studies 7 (Bratislava). 45-51.

Zipf, George Kingsley (1932). Selected studies of the principle of relative frequency in language. Cambridge, Mass.: Harvard University Press.