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 production 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**.

\[ \text{L-functional complexity} = \ln A + B \times \text{L-graphical complexity}, \]

where \( B \) is expected to be negative.

**Power function:**

\[ \text{Functional complexity} = A \times \text{Graphical complexity}^B \]
Direct H 1: The **functional complexity** of Chinese characters is directly a function of their **graphical complexity**.

- L-functional complexity = \( \ln A + B \times \text{L-graphical complexity} \),
  - where \( B \) is expected to be negative

- Power function: Functional complexity = \( A \times \text{graphical complexity}^B \)
  a) Number of strokes: \( D = 0.956 \quad A = e^{5.59} = 268.12 \quad B = -1.373 \)
  b) Number of graphemes: \( D = 0.953 \quad A = e^{3.666} = 39.09 \quad B = -1.133 \)
  c) Writing effort: \( D = 0.95 \quad A = e^{6.086} = 439.72 \quad B = -1.44 \)
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 = \( \ln A + B \times \text{L-functional complexity} \)

- Power function:
  \( \text{Frequency} = A \times \text{Functional complexity}^B \)
Direct H 2: The **text frequency** of Chinese characters is a function of their **functional complexity**.

- L-frequency = \( \ln A + B \times \text{L-functional complexity} \)
- Power function: Frequency = \( A \times \text{Functional complexity}^B \)

\[
D = 0.958 \quad A = e^{2.444} = 11.52 \quad B = 1,215
\]

(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 \times L-frequency
  - A negative value for B is expected

- **Power function**: Graphical complexity = A \times \text{Frequency}^8
Direct H 3: The **graphical complexity** of Chinese characters is a function of their **text frequency**.

- L-graphical complexity = $\ln A + B \times \text{L-frequency}$
  - A negative value for $B$ is expected

- Graphical complexity = $A \times \text{Frequency}^B$

<table>
<thead>
<tr>
<th>Way of measurement</th>
<th>Class width 100</th>
<th>Class width 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Number of strokes</td>
<td>$D = 0.94$</td>
<td>$D = 0.93$</td>
</tr>
<tr>
<td></td>
<td>$A = e^{2.846} = 17.22$</td>
<td>$A = e^{2.72} = 15.18$</td>
</tr>
<tr>
<td></td>
<td>$B = -0.114$</td>
<td>$B = -0.094$</td>
</tr>
<tr>
<td>a) Number of graphemes</td>
<td>$D = 0.95$</td>
<td>$D = 0.897$</td>
</tr>
<tr>
<td></td>
<td>$A = e^{1.51} = 4.53$</td>
<td>$A = e^{1.4} = 4.066$</td>
</tr>
<tr>
<td></td>
<td>$B = -0.0958$</td>
<td>$B = -0.078$</td>
</tr>
<tr>
<td>a) Writing effort</td>
<td>$D = 0.946$</td>
<td>$D = 0.92$</td>
</tr>
<tr>
<td></td>
<td>$A = e^{3.057} = 21.28$</td>
<td>$A = e^{2.94} = 18.88$</td>
</tr>
<tr>
<td></td>
<td>$B = -0.11$</td>
<td>$B = -0.09$</td>
</tr>
</tbody>
</table>
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.²
Indirect H 4: The **graphical complexity** of Chinese characters is indirectly a function of its **functional complexity**, mediated by frequency.

- **L-graphical complexity** = \( \ln X + Y \times \text{L-functional complexity} \).
  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**
    - \( \text{L-graphical complexity}_{a1} = 2.72 - 0.094 \times (2.444 + 1.215 \times \text{L-functional complexity}) \)
    - \( = 2.49 - 0.114 \times \text{L-functional complexity} \)
    - and
    - \( \text{L-graphical complexity}_{a2} = 2.85 - 0.114 \times (2.444 + 1.215 \times \text{L-functional complexity}) \)
    - \( = 2.57 - 0.138 \times \text{L-functional complexity} \)

  - **Graphical complexity measured in number of component graphemes**
    - \( \text{L-graphical complexity}_{b1} = 1.4 - 0.078 \times (2.444 + 1.215 \times \text{L-functional complexity}) \)
    - \( = 1.2 - 0.095 \times \text{L-functional complexity} \)
    - and
    - \( \text{L-graphical complexity}_{b2} = 1.51 - 0.096 \times (2.444 + 1.215 \times \text{L-functional complexity}) \)
    - \( = 1.277 - 0.116 \times \text{L-functional complexity} \)

  - **Graphical complexity measured in writing effort**
    - \( \text{L-graphical complexity}_{c1} = 2.94 - 0.09 \times (2.444 + 1.215 \times \text{L-functional complexity}) \)
    - \( = 2.72 - 0.109 \times \text{L-functional complexity} \)
    - and
    - \( \text{L-graphical complexity}_{c2} = 3.06 - 0.109 \times (2.444 + 1.215 \times \text{L-functional complexity}) \)
    - \( = 2.79 - 0.13 \times \text{L-functional complexity} \)
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: $D = 0.73 \quad A = e^{2.55} = 12.82 \quad B = -0.116$
  b) Number of graphemes: $D = 0.60 \quad A = e^{1.25} = 3.49 \quad B = -0.092$
  c) Writing effort: $D = 0.75 \quad A = e^{2.78} = 16.19 \quad B = -0.114$
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

<table>
<thead>
<tr>
<th>Way of measurement</th>
<th>Theoretical function</th>
<th>Empirical function</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Number of strokes</td>
<td>Graph.comp.a1 = 12.06 * funct. comp.^-0.114</td>
<td>Graph.comp.ae = 12.82 * funct. comp.^-0.116</td>
</tr>
<tr>
<td>a) Number of graphemes</td>
<td>Graph.comp.b1 = 3.36 * funct. comp.^-0.095</td>
<td>Graph.comp.be = 3.49 * funct. comp.^-0.092</td>
</tr>
<tr>
<td>a) Writing effort</td>
<td>Graph.comp.c1 = 15.16 * funct. comp.^-0.109</td>
<td>Graph.comp.ce = 16.19 * funct. comp.^-0.114</td>
</tr>
<tr>
<td>a) Writing effort</td>
<td>Graph.comp.c2 = 16.3 * funct. comp.^-0.13</td>
<td></td>
</tr>
</tbody>
</table>
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:
  
  \[
  \text{Funct. Comp.} = A \times \text{Frequency}^B
  \]
Indirect H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity.

\[
\text{L-functional complexity} = \ln X + Y \times \text{L-frequency}
\]

\[
\text{L-functional complexity}_{a1} = 5.59 - 1.373 \times (2.85 - 0.114 \times \text{L-frequency})
\]
\[= 1.68 + 0.156 \times \text{L-frequency} \]

and

\[
\text{L-functional complexity}_{a2} = 5.59 - 1.373 \times (2.72 - 0.094 \times \text{L-frequency})
\]
\[= 1.85 + 0.13 \times \text{L-frequency} \]

\[
\text{L-functional complexity}_{b1} = 3.666 - 1.133 \times (1.51 - 0.096 \times \text{L-frequency})
\]
\[= 1.95 + 0.108 \times \text{L-frequency} \]

and

\[
\text{L-functional complexity}_{b2} = 3.666 - 1.133 \times (1.4 - 0.078 \times \text{L-frequency})
\]
\[= 2.076 + 0.088 \times \text{L-frequency} \]

\[
\text{L-functional complexity}_{c1} = 6.086 - 1.441 \times (3.06 - 0.109 \times \text{L-frequency})
\]
\[= 1.68 + 0.157 \times \text{L-frequency} \]

and

\[
\text{L-functional complexity}_{c2} = 6.086 - 1.441 \times (2.94 - 0.09 \times \text{L-frequency})
\]
\[= 1.85 + 0.13 \times \text{L-frequency} \]
Indirect H 5: **Functional complexity** indirectly is a function of **text frequency**, mediated by graphical complexity.

- Power function: \( \text{Funct. Comp.} = A \times \text{Frequency}^B \)
- Regression results:
  - Class width 100: \( D = 0.969 \quad A = e^{-1.649} = 0.192 \quad B = 0.804 \)
  - Class width 50: \( D = 0.97 \quad A = e^{-1.173} = 0.31 \quad B = 0.74 \)
Indirect H 5: **Functional complexity** indirectly is a function of text frequency, mediated by graphical complexity.

- Comparison between fitted functions and theoretical functions

<table>
<thead>
<tr>
<th>Class width</th>
<th>Theoretical functions</th>
<th>Empirical function</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>FUNCT. COMP. A1 = 5.37 * FREQ. 0.156</td>
<td>FUNCT. COMP. E1 = 0.192 * FREQ. 0.804</td>
</tr>
<tr>
<td></td>
<td>FUNCT. COMP. B1 = 7.05 * FREQ. 0.108</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FUNCT. COMP. C1 = 5.36 * FREQ. 0.157</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>FUNCT. COMP. A2 = 6.36 * FREQ. 0.13</td>
<td>FUNCT. COMP. E2 = 0.31 * FREQ. 0.74</td>
</tr>
<tr>
<td></td>
<td>FUNCT. COMP. B2 = 7.98 * FREQ. 0.088</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FUNCT. COMP. C2 = 6.36 * FREQ. 0.13</td>
<td></td>
</tr>
</tbody>
</table>
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** = \( \ln X + Y \times \text{L-graphical complexity} \)

- **Power function**:

  \[ \text{Frequency} = A \times \text{Graphical complexity}^8 \]
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 \( L-freq_a = 2.444 + 1.215 \times (5.59 - 1.373 \times L\text{-graph.comp.}) \)
  \[ = 9.24 - 1.67 \times L\text{-graph.comp.} \]
  b) No. of graphemes \( L-freq_b = 2.444 + 1.215 \times (3.666 - 1.133 \times L\text{-graph.comp.}) \)
  \[ = 6.9 - 1.377 \times L\text{-graph.comp.} \]
  c) Writing effort \( L-freq_c = 2.444 + 1.215 \times (6.086 - 1.441 \times L\text{-graph.comp.}) \)
  \[ = 9.84 - 1.75 \times L\text{-graph.comp.} \]

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

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

<table>
<thead>
<tr>
<th></th>
<th>Theoretically</th>
<th>Empirically</th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>( Freq_a = 10,287.14 \times \text{Komp}^{1.67} )</td>
<td>( Freq_{ea} = 64,690.26 \times \text{Komp}^{2.466} )</td>
</tr>
<tr>
<td>b)</td>
<td>( Freq_b = 992.27 \times \text{Komp}^{1.377} )</td>
<td>( Freq_{eb} = 2,058.5 \times \text{Komp}^{1.98} )</td>
</tr>
<tr>
<td>c)</td>
<td>( Freq_c = 18,797.89 \times \text{Komp}^{1.75} )</td>
<td>( Freq_{ec} = 117,557.75 \times \text{Komp}^{2.47} )</td>
</tr>
</tbody>
</table>
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！