brain+cognitive sciences

PEKING UNIVERSITY

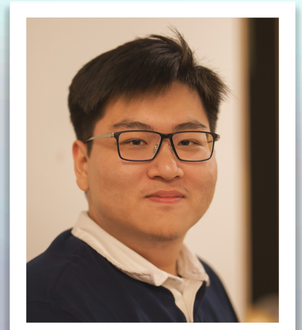
Full Paper

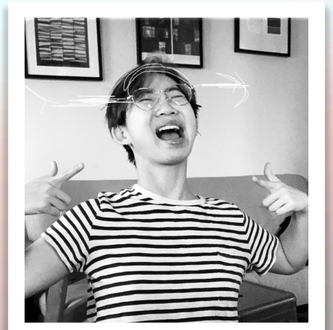
COGSCI 2024 DYNAMICS OF COGNITION July 24-27 • Rotterdam

Sayan Gul Award
Best Undergrad Paper

# Finding structure in logographic writing with library learning

**Guangyuan Jiang**, Matthias Hofer, Jiayuan Mao, Lionel Wong, Josh Tenenbaum, and Roger Levy



# Human language is deeply structured

- Human language is deeply structured — a **universal** trait of human communication systems.



see *How Efficiency Shapes Human Language,* Gibson et al., 2019 for a review

# Human language is deeply structured

- Human language is deeply structured — a **universal** trait of human communication systems.

$$\begin{bmatrix} +\text{stop} \\ +\text{consonantal} \\ +\text{alveolar} \end{bmatrix} \rightarrow [\text{ɾ}] \;/\; \begin{bmatrix} +\text{vowel} \\ +\text{stressed} \end{bmatrix} \_\_\_\_ \begin{bmatrix} +\text{vowel} \\ -\text{stressed} \end{bmatrix}$$

phonetics and phonology



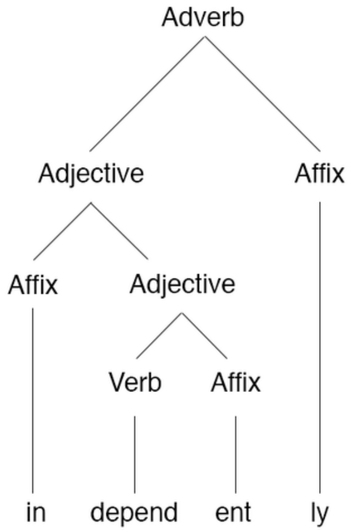
see *How Efficiency Shapes Human Language,* Gibson et al., 2019 for a review

# Human language is deeply structured

- Human language is deeply structured — a **universal** trait of human communication systems.

$$\begin{bmatrix} +\text{stop} \\ +\text{consonantal} \\ +\text{alveolar} \end{bmatrix} \rightarrow [\text{ɾ}] \;/\; \begin{bmatrix} +\text{vowel} \\ +\text{stressed} \end{bmatrix} \_\_\_\_ \begin{bmatrix} +\text{vowel} \\ -\text{stressed} \end{bmatrix}$$

phonetics and phonology




morphology and syntax



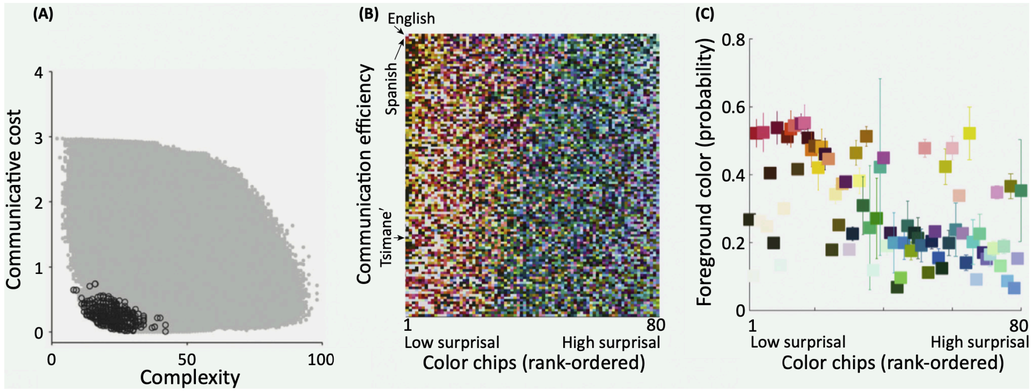
see *How Efficiency Shapes Human Language,* Gibson et al., 2019 for a review

# Human language is deeply structured

- Human language is deeply structured — a **universal** trait of human communication systems.

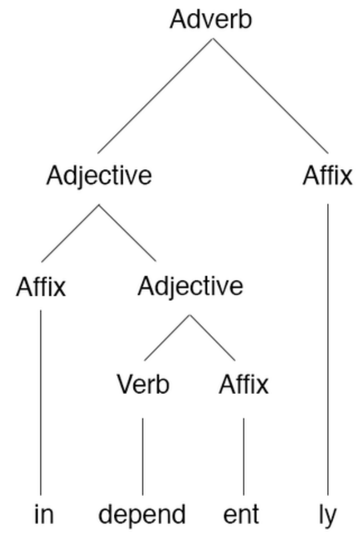
$$\begin{bmatrix} +\text{stop} \\ +\text{consonantal} \\ +\text{alveolar} \end{bmatrix} \rightarrow [\text{ɾ}] \;/\; \begin{bmatrix} +\text{vowel} \\ +\text{stressed} \end{bmatrix} \_\_\_\_ \begin{bmatrix} +\text{vowel} \\ -\text{stressed} \end{bmatrix}$$

phonetics and phonology




semantic space: kinship, color

(Kemp & Regier, 2012; Zaslavsky et al., 2018)




morphology and syntax

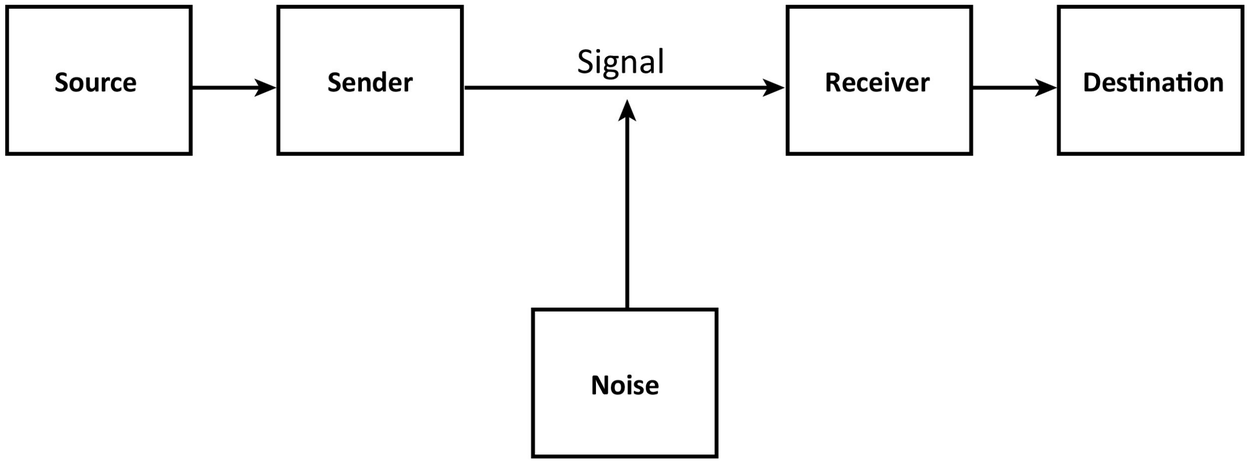
see *How Efficiency Shapes Human Language,* Gibson et al., 2019 for a review

# Why structured? Shaped by efficiency



Figures from: Gibson et al., 2019; Chater & Vitányi, 2003; Smith, 2018; Hawkins et al., 2023.

# Why structured? Shaped by efficiency

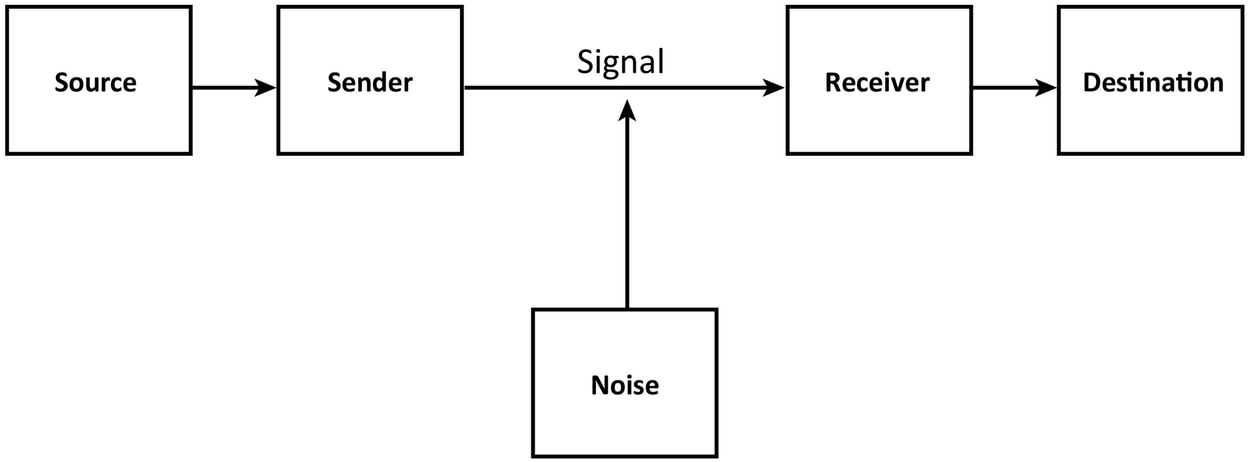



communicative efficiency

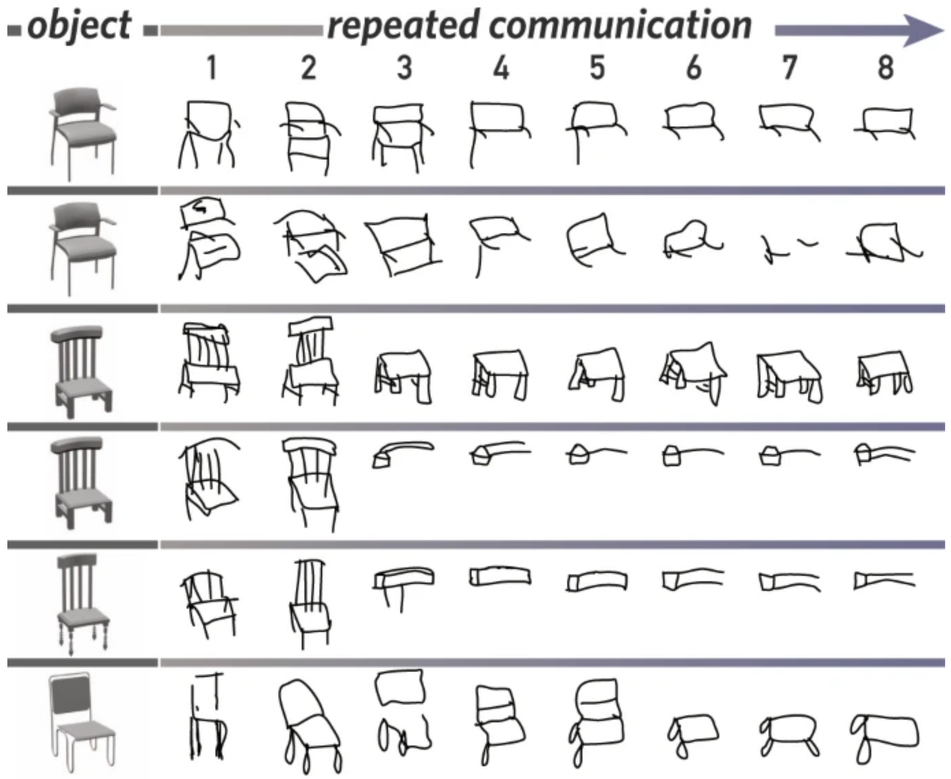


Figures from: Gibson et al., 2019; Chater & Vitányi, 2003; Smith, 2018; Hawkins et al., 2023.

# Why structured? Shaped by efficiency



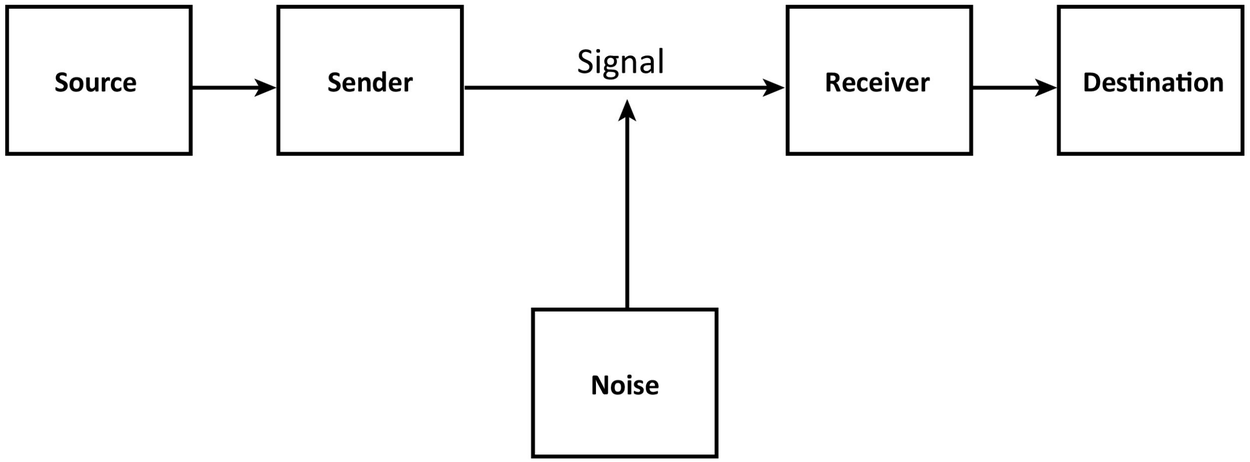

communicative efficiency



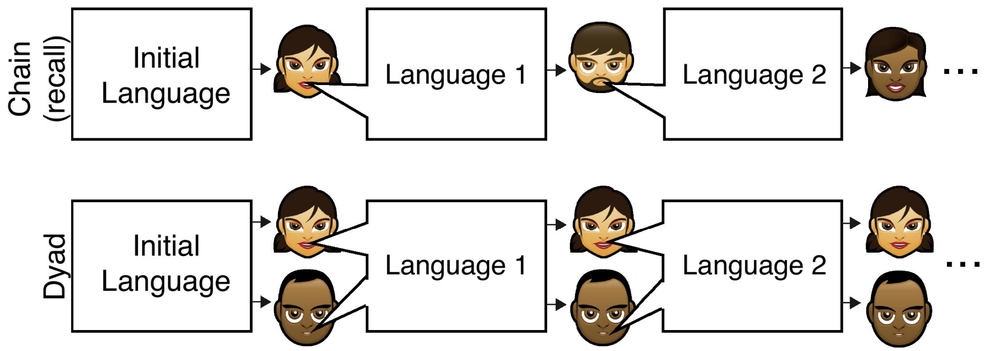

representational efficiency

Figures from: Gibson et al., 2019; Chater & Vitányi, 2003; Smith, 2018; Hawkins et al., 2023.

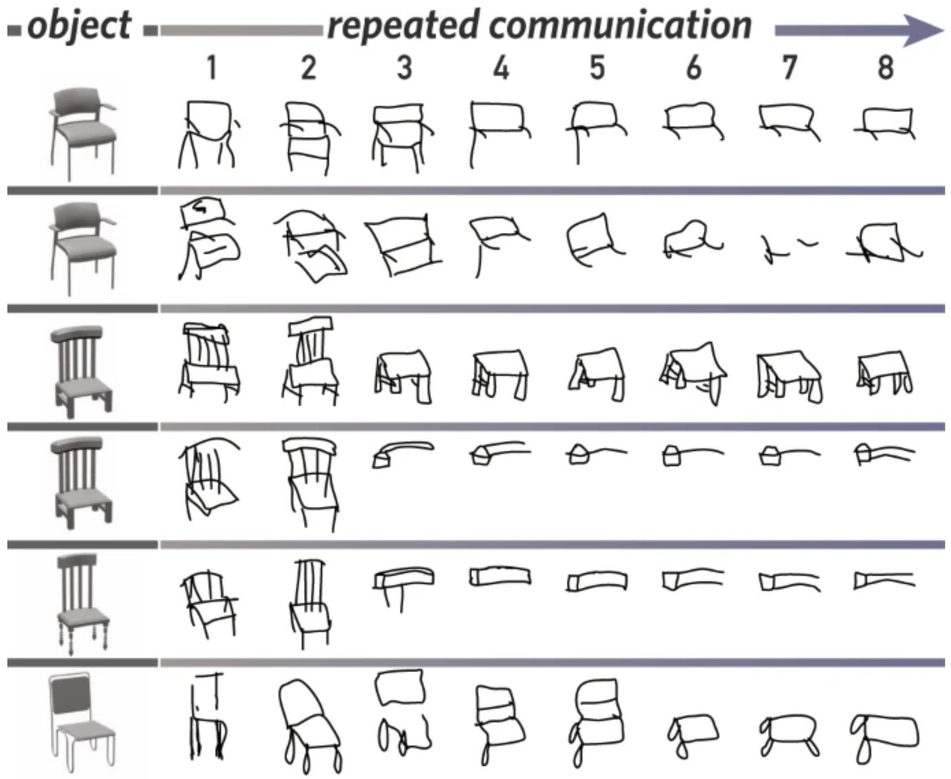
# Why structured? Shaped by efficiency




communicative efficiency




learnability




representational efficiency



Figures from: Gibson et al., 2019; Chater & Vitányi, 2003; Smith, 2018; Hawkins et al., 2023.

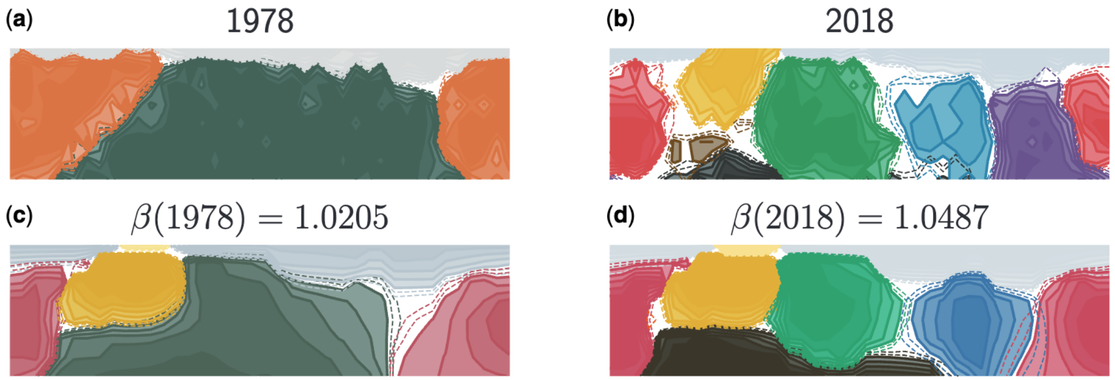
# Human language shaped by efficiency

## …and how languages evolve over time

# Human language shaped by efficiency

## …and how languages evolve over time

In the wild



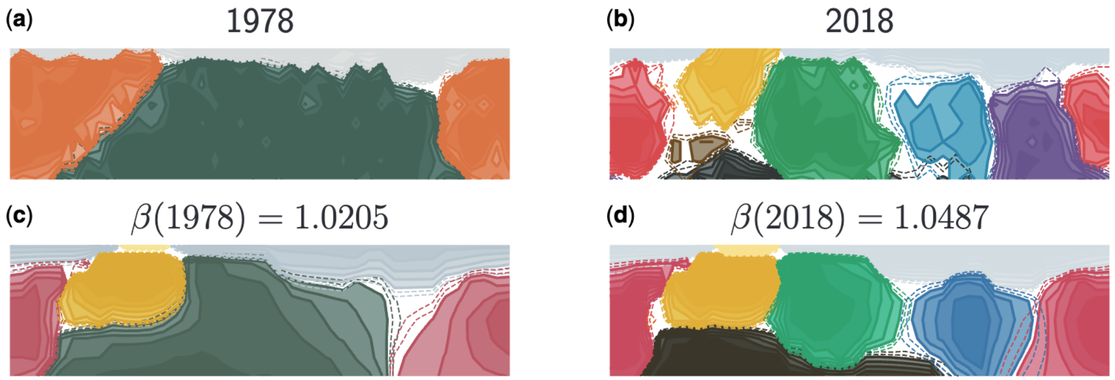

(Zaslavsky et al., 2022)

Chinese characters

# Human language shaped by efficiency

## …and how languages evolve over time

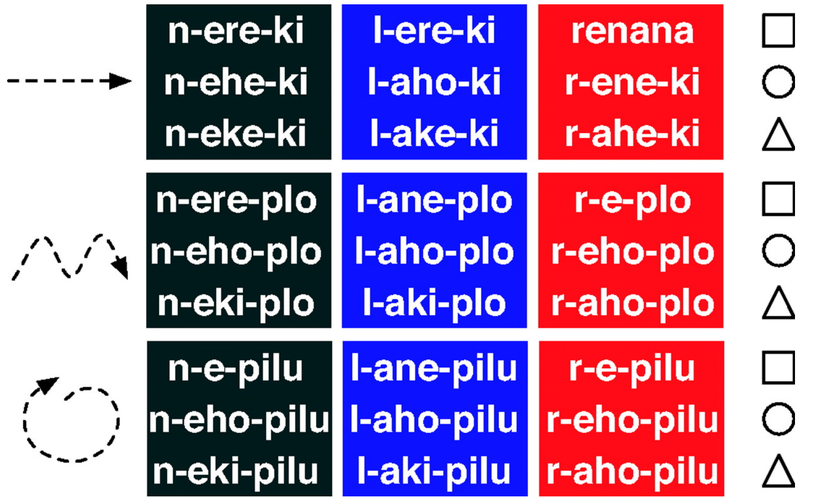
In the wild




(Zaslavsky et al., 2022)

Chinese characters

In the lab



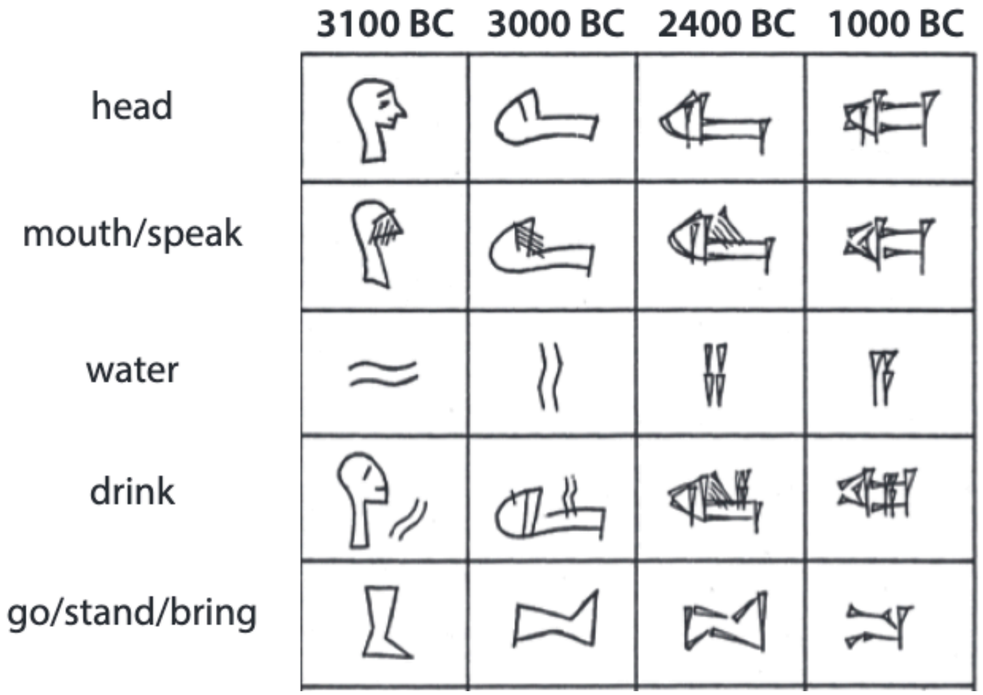

(Kirby, Cornish, & Smith, 2008)

# Human language shaped by efficiency

## Languages’ earliest records —writing

# Human language shaped by efficiency

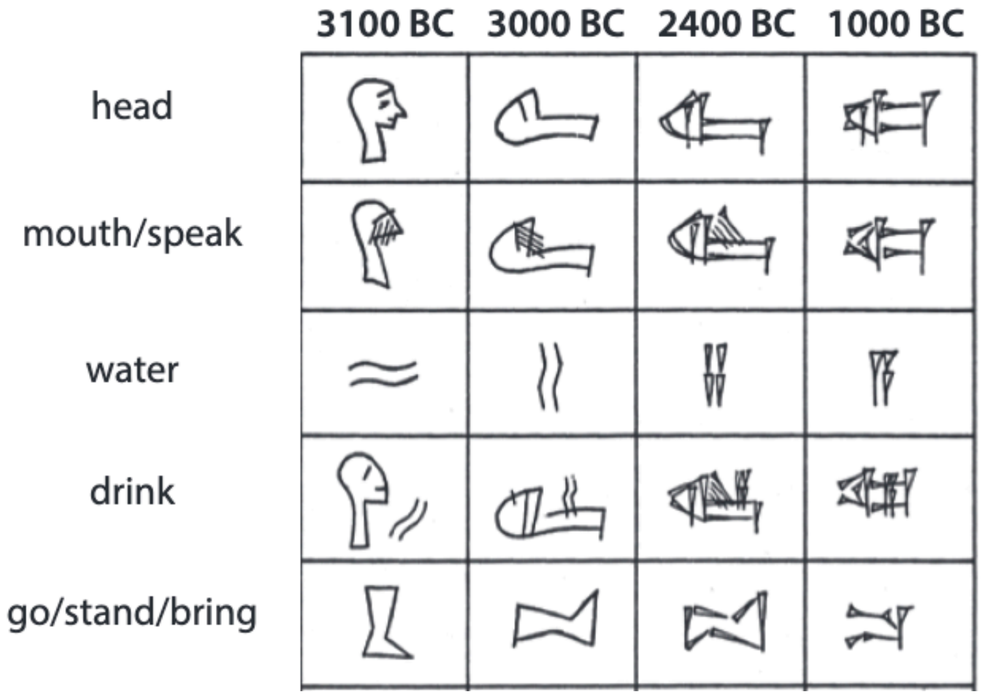
## Languages' earliest records —writing



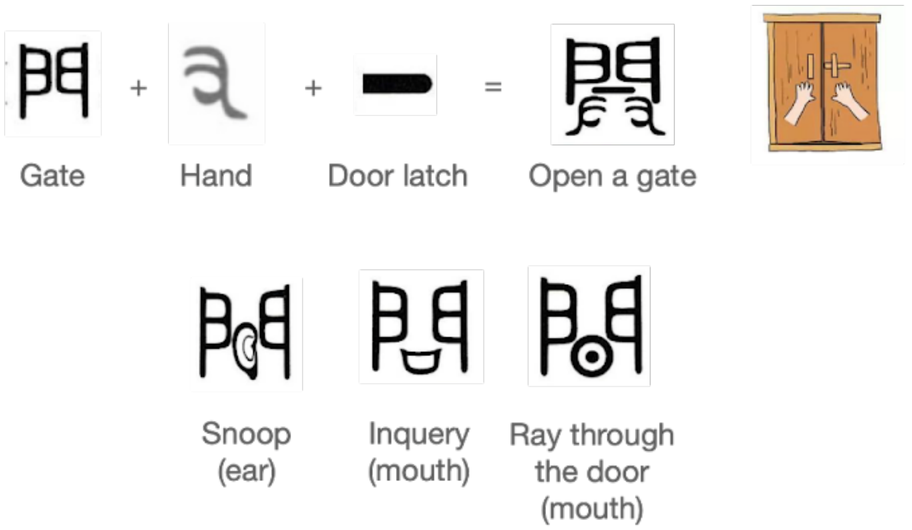

Sumerian Cuneiform (Sampson, 1985)

# Human language shaped by efficiency

## Languages' earliest records —writing




Sumerian Cuneiform (Sampson, 1985)




Chinese oracle bone scripts

# Efficiency-based structure discovery

## Morphology discovery (Goldsmith, 2001)

*laughed*
*laughing*
*laughs*
*walked*
*walking*
*walks*
*jumped*
*jumping*
*jumps*

total letter count: 57

# Efficiency-based structure discovery

## Morphology discovery (Goldsmith, 2001)

*laugh***ed**
*laugh***ing**
*laugh***s**
*walk***ed**
*walk***ing**
*walk***s**
*jump***ed**
*jump***ing**
*jump***s**

total letter count: **57**

# Efficiency-based structure discovery

## Morphology discovery (Goldsmith, 2001)

*laugh**ed***
*laugh**ing***
*laugh**s***
*walk**ed***
*walk**ing***
*walk**s***
*jump**ed***
*jump**ing***
*jump**s***

total letter count: **57**

# Efficiency-based structure discovery

## Morphology discovery (Goldsmith, 2001)

*laugh***ed**
*laugh***ing**
*laugh***s**
*walk***ed**
*walk***ing**
*walk***s**
*jump***ed**
*jump***ing**
*jump***s**

{ *laugh* *walk* *jump* }{ **ed** **ing** **s** }

total letter count: **57**

total letter count: **19**

# Efficiency-based structure discovery

## Morphology discovery (Goldsmith, 2001)

*laughed*
*laughing*
*laughs*
*walked*
*walking*
*walks*
*jumped*
*jumping*
*jumps*

{ *laugh* *walk* *jump* } { **ed** **ing** **s** }

*Efficient representation leads to morphological structure*

total letter count: **57**

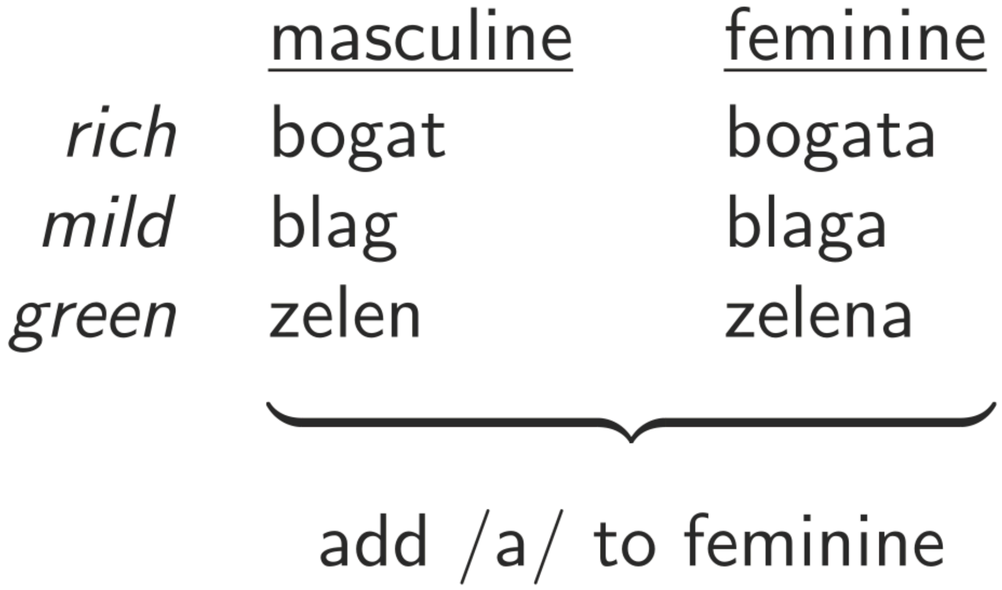
total letter count: **19**

# Efficiency-based structure discovery

## Learning syntax/morphological rules (Kim, Dyer, & Rush, 2019; Ellis et al., 2022)

| PCFG Rule | DMV parameter |
|---|---|
| $S \rightarrow Y_h$ | $P_{root}(h)$ |
| $Y_h \rightarrow L_h^0 \ R_h^0$ | $1$ |
| $L_h^0 \rightarrow h_l$ | $P_{stop}(\text{Stop}\|h, \leftarrow, \text{no\_dep})$ |
| $L_h^0 \rightarrow L_h^{'}$ | $P_{stop}(\neg\text{Stop}\|h, \leftarrow, \text{no\_dep})$ |
| $L_h^{'} \rightarrow Y_d \ L_h$ | $P_{choose}(d\|h, \leftarrow)$ |
| $L_h \rightarrow h_l$ | $P_{stop}(\text{Stop}\|h, \leftarrow, \text{one\_dep})$ |
| $L_h \rightarrow L_h^{'}$ | $P_{stop}(\neg\text{Stop}\|h, \leftarrow, \text{one\_dep})$ |

learning syntax/grammar
(Kim, Dyer, & Rush, 2019)




morpho-phonology discovery
(Ellis et al., 2022)



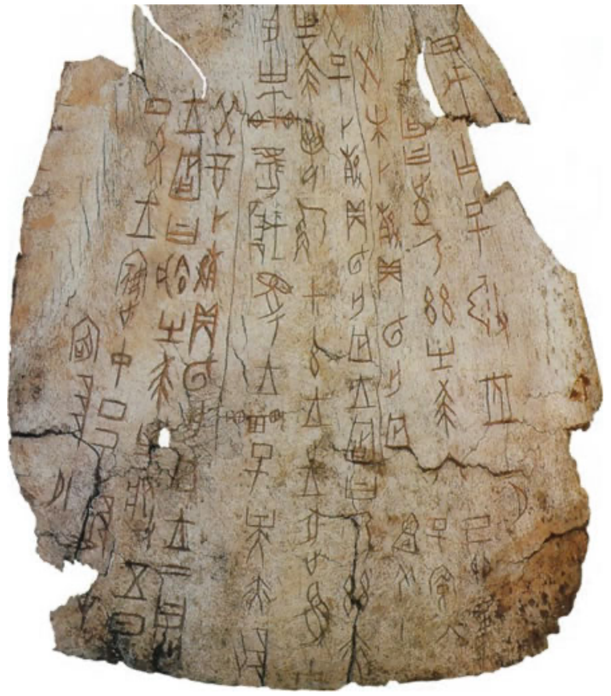
Figure: Pate & Johnson, 2016

# Our hypotheses

- If language has been shaped by pressure for efficiency,
  - then search for representational efficiency should recover its combinatorial elements.
- Furthermore, we should see an increase in efficiency over time.

# Our domain: Chinese Orthography

## Earliest records — oracle bone scripts (~1500-1050 BC)




*source: omniglot.com*

# Our domain: Chinese Orthography

| | oracle | seal | traditional | simplified |
|---|---|---|---|---|
| *sink* |  |  | 瀋 | 沈 |
| *float* |  |  | 漂 | 漂 |
| *color* |  |  | 綵 | 彩 |
| *insect* |  |  | 崑 | 昆 |
| *orange* |  |  | 桔 | 桔 |
| *peace* |  |  | 寧 | 宁 |

1500 BC 1050 BC 200 AD 1950 AD

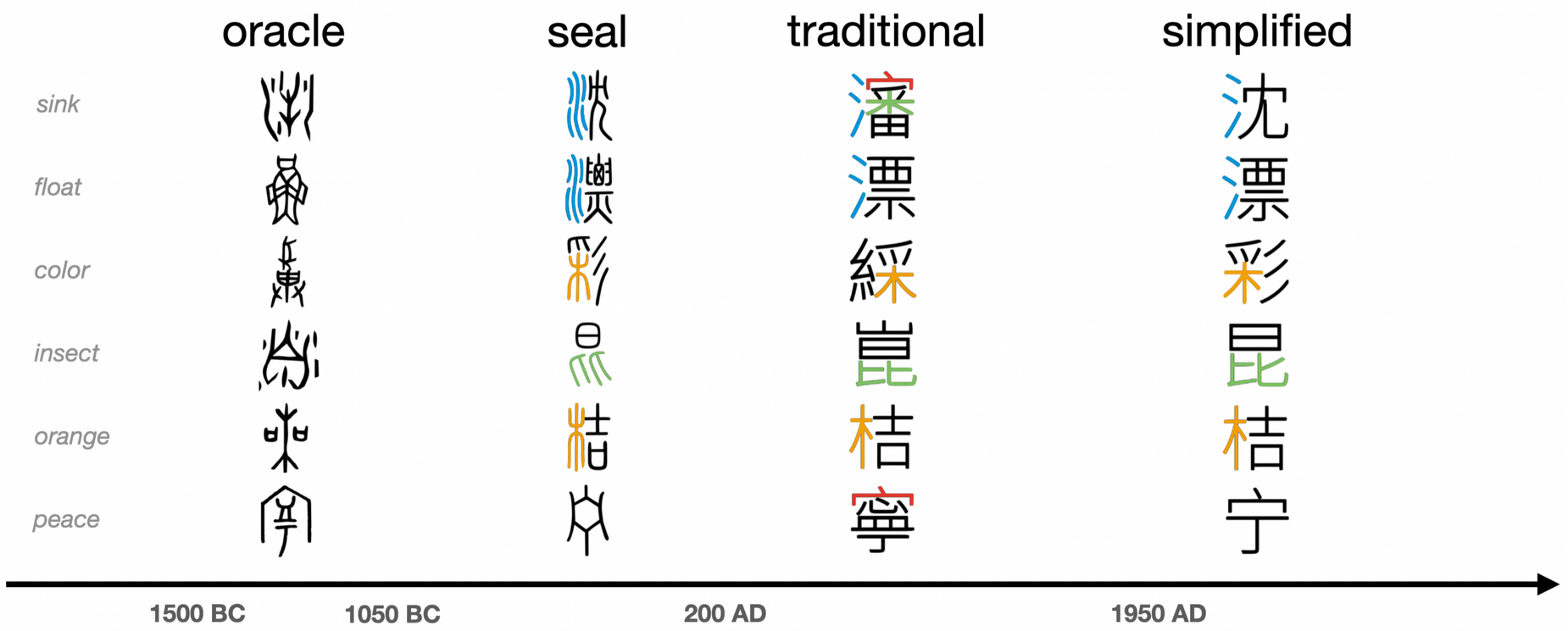
# Our domain: Chinese Orthography

| | oracle | seal | traditional | simplified |
|---|---|---|---|---|
| *sink* |  |  | 瀋 | 沈 |
| *float* |  |  | 漂 | 漂 |
| *color* |  | 彩 | 綵 | 彩 |
| *insect* |  |  | 崑 | 昆 |
| *orange* |  | 桔 | 桔 | 桔 |
| *peace* |  |  | 寧 | 宁 |

1500 BC 1050 BC 200 AD 1950 AD

# Chinese 101

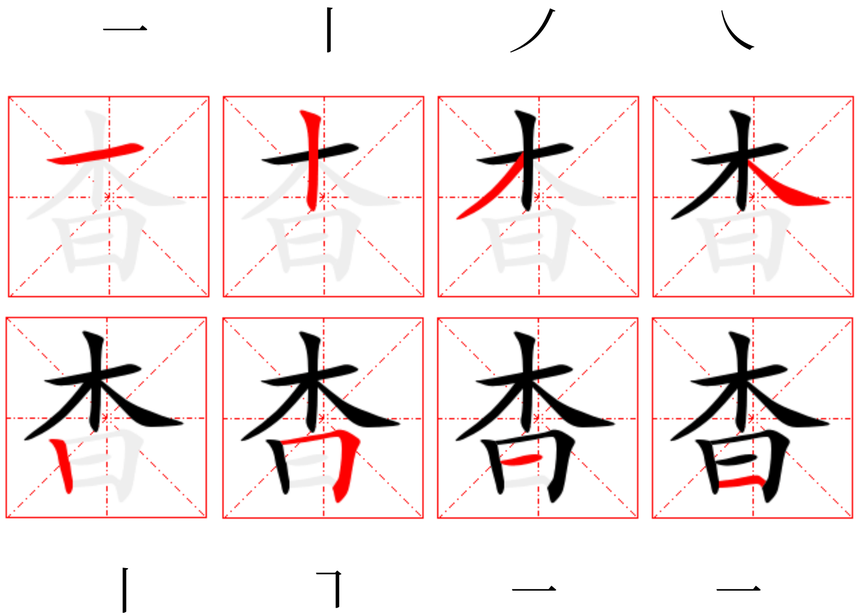
## Long evolutionary history (of over 3,000 years).




- **Frequent reuse** of graphical elements **within** individual **characters** and **across** the **writing system**.
- **Unique opportunity** for studying combinatorial structure.

# Chinese 101

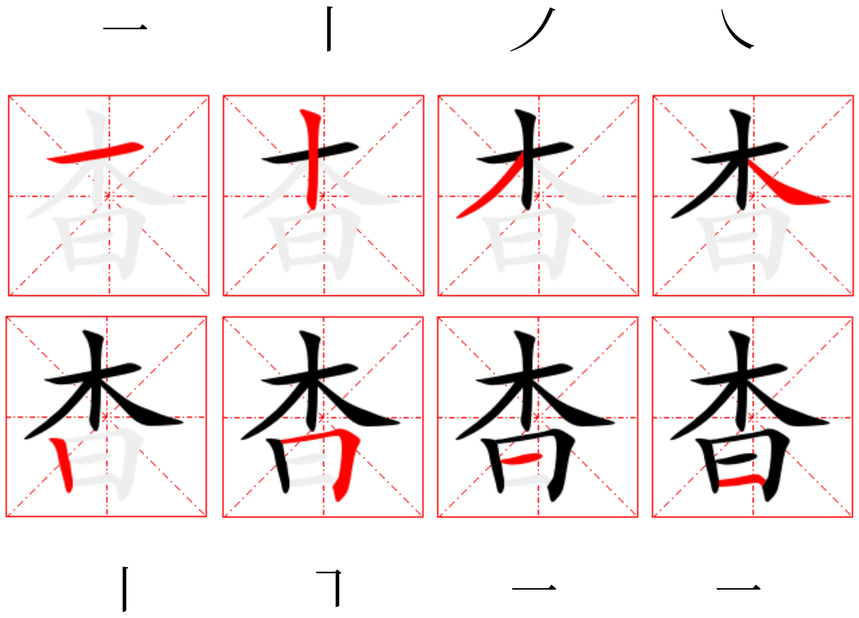
## Characters, radicals, and strokes



*Chinese characters are made up of strokes*

# Chinese 101

## Characters, radicals, and strokes



一 丨 丿 丶

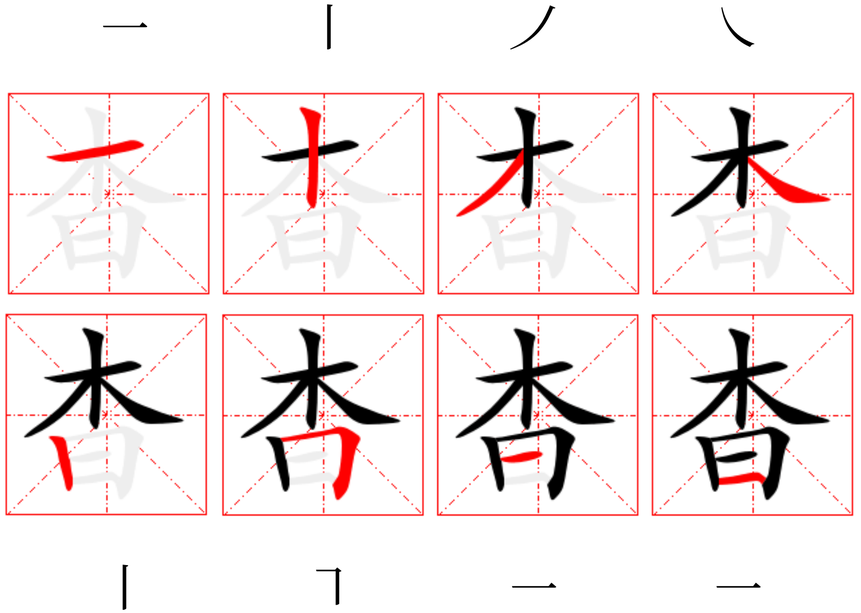
丨 ㇆ 一 一

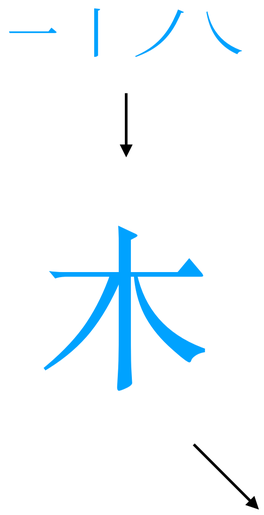
*strokes (primitives)*

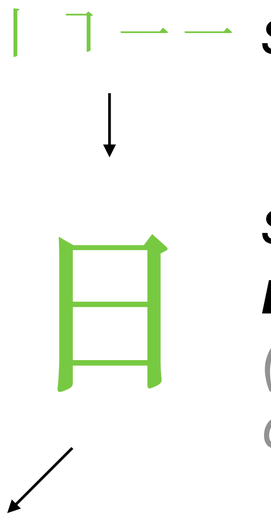
*Chinese characters are made up of strokes*

# Chinese 101

## Characters, radicals, and strokes







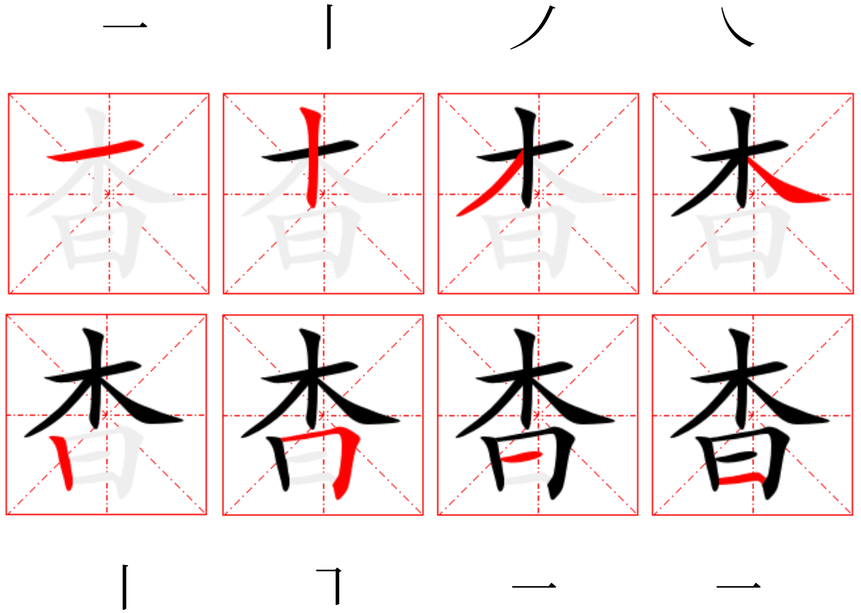
*strokes (primitives)*

*strokes group to* ***radicals*** *(graphical components)*

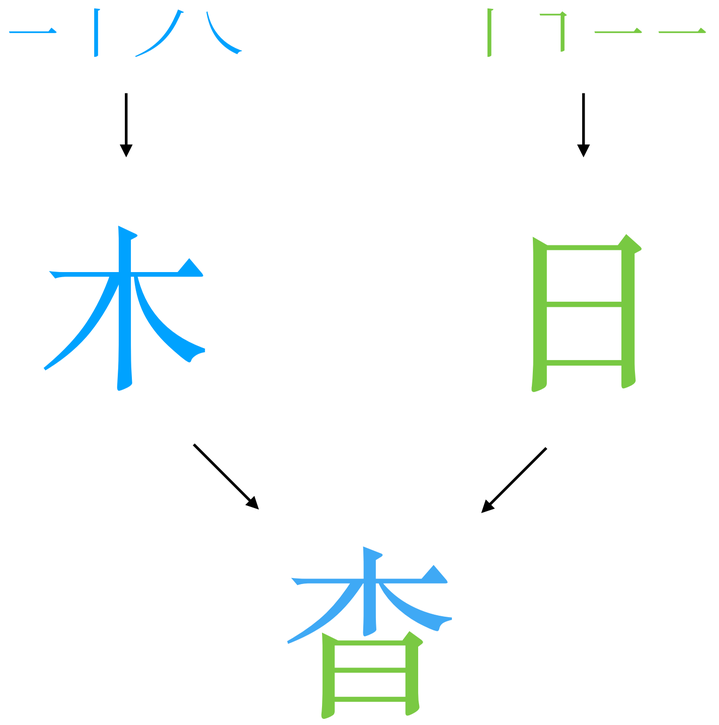
*Chinese characters are made up of strokes*

# Chinese 101

## Characters, radicals, and strokes



*Chinese characters are made up of strokes*



*strokes (primitives)*

*strokes group to* ***radicals*** *(graphical components)*

*radicals group to* ***characters***

# Chinese 101

## Characters, radicals, and strokes

*empirically, radicals can be combinatorially reused*

木

木

日

木

日

# Chinese 101

## Characters, radicals, and strokes

*empirically, radicals can be combinatorially reused*

木 x 3 =

木 + 日 =

日 x 3 =

木 + 月 =

日 + 月 =

# Chinese 101

## Characters, radicals, and strokes

*empirically, radicals can be combinatorially reused*

木 x 3 = 森

木 + 日 = 杳

日 x 3 = 晶

木 + 月 = 枴

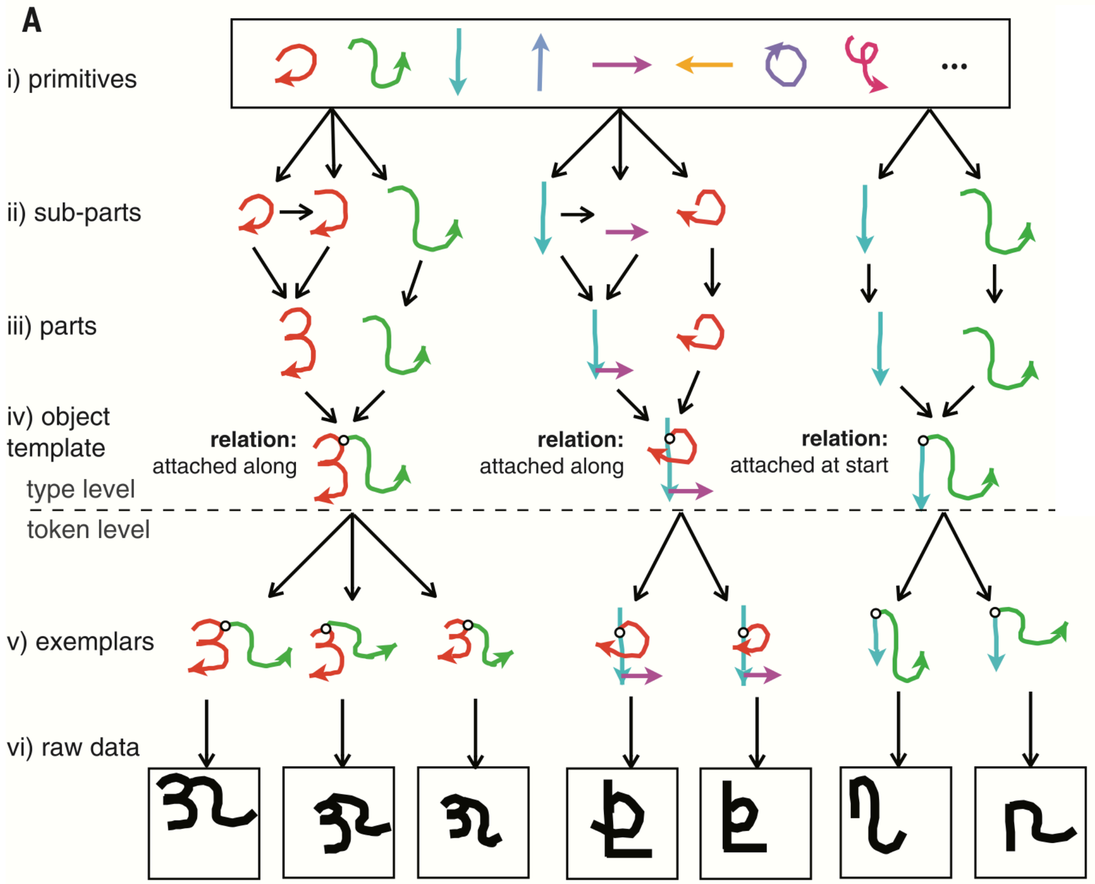
日 + 月 = 明

# Measuring representational efficiency

## Inferring motor programs from images with bayesian inference

# Measuring representational efficiency

## Inferring motor programs from images with bayesian inference

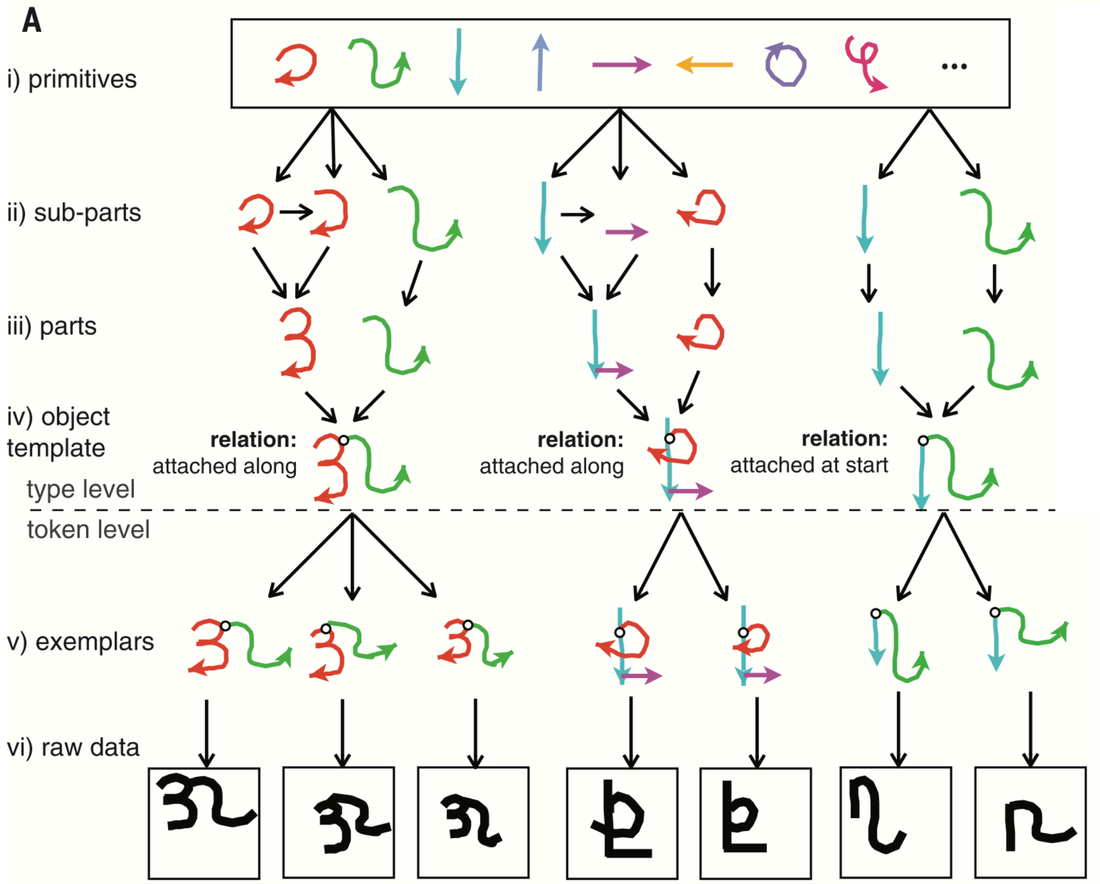



handwritten characters

(Lake, Salakhutdinov & Tenenbaum, 2015)

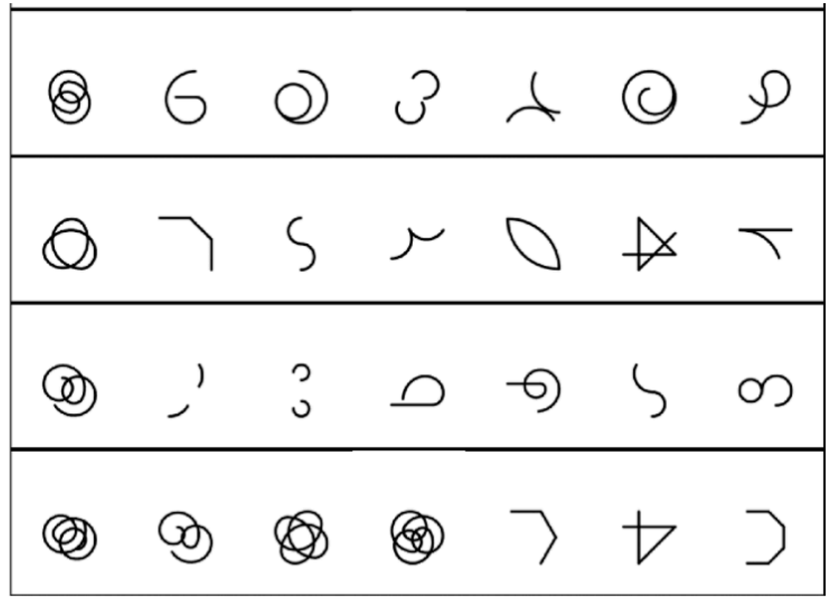
# Measuring representational efficiency

## Inferring motor programs from images with bayesian inference




handwritten characters

(Lake, Salakhutdinov & Tenenbaum, 2015)



geometric shapes

(Sablé-Meyer et al., 2022)

Shape perception formulated as searching programs with MDL.

# Contributions

# Contributions

- We **develop a library learning model** that allows us to jointly discover an underlying inventory of **higher order graphical forms** and evaluate the **MDL** of writing system represented with that set of abstract components.

# Contributions

- We **develop a library learning model** that allows us to jointly discover an underlying inventory of **higher order graphical forms** and evaluate the **MDL** of writing system represented with that set of abstract components.

- Our findings:

# Contributions

- We **develop a library learning model** that allows us to jointly discover an underlying inventory of **higher order graphical forms** and evaluate the **MDL** of writing system represented with that set of abstract components.

- Our findings:
  - Our model **rediscovers widely recognized theories** of combintorial structure in the Chinese orthography.

# Contributions

- We **develop a library learning model** that allows us to jointly discover an underlying inventory of **higher order graphical forms** and evaluate the **MDL** of writing system represented with that set of abstract components.

- Our findings:
  - Our model **rediscovers widely recognized theories** of combintorial structure in the Chinese orthography.
  - We extend this analysis **diachronically**, investigating the **evolution of Chinese scripts** over several representative historical stages.

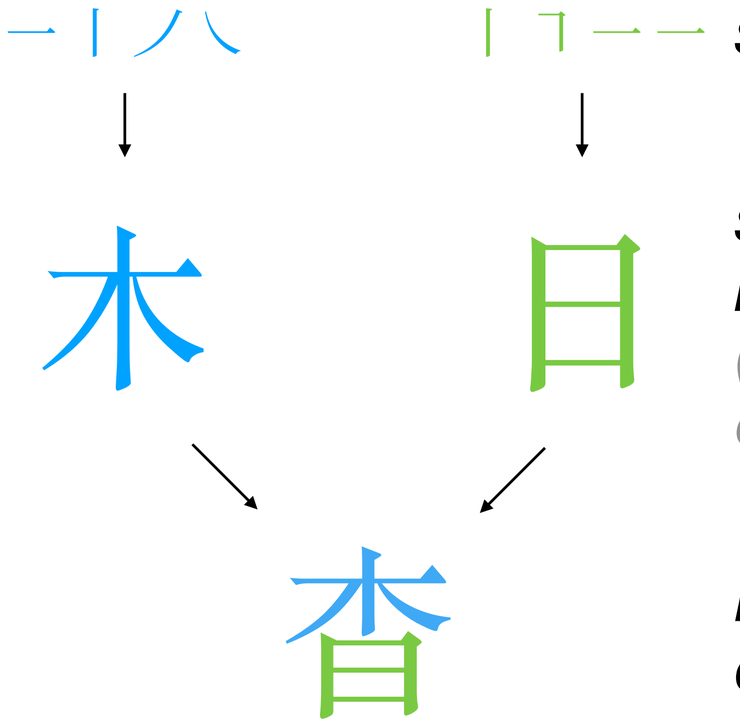
# Contributions

- We **develop a library learning model** that allows us to jointly discover an underlying inventory of **higher order graphical forms** and evaluate the **MDL** of writing system represented with that set of abstract components.

- Our findings:
  - Our model **rediscovers widely recognized theories** of combintorial structure in the Chinese orthography.
  - We extend this analysis **diachronically**, investigating the **evolution of Chinese scripts** over several representative historical stages.
  - And yield an interesting diachronic finding about the relationship **between** two **modern Chinese scripts**.

# Our Approach: Strokes

## Structure discovery with library learning



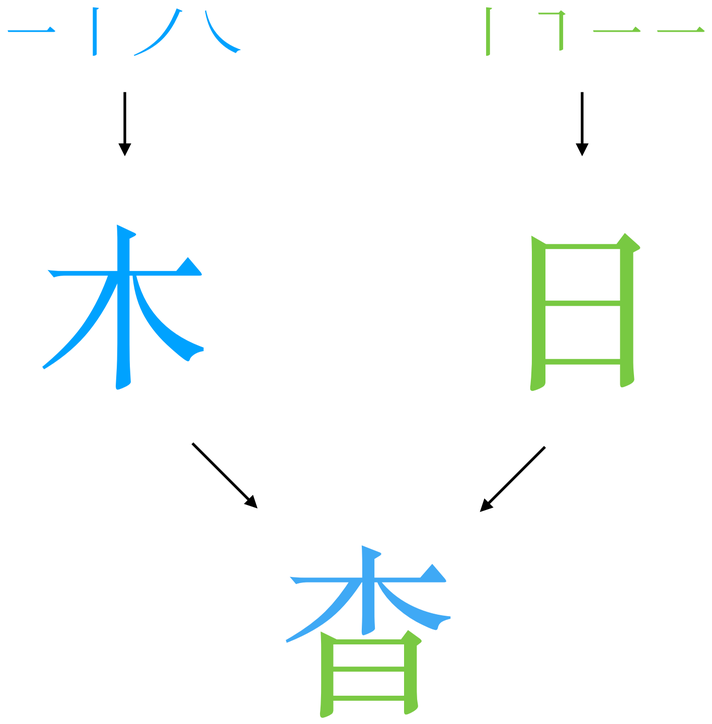
*strokes (primitives)*

*strokes group to* ***radicals*** *(graphical components)*

*radicals group to* ***characters***

# Our Approach: Strokes

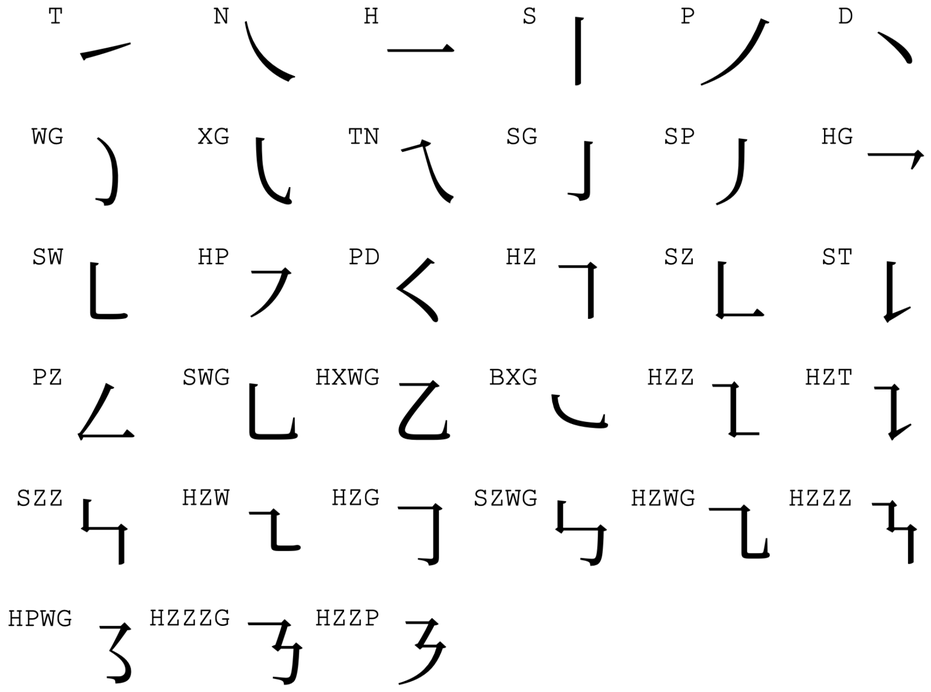
## Structure discovery with library learning



*strokes (primitives)*

*strokes group to* ***radicals*** *(graphical components)*

*radicals group to* ***characters***




base strokes

# Structure discovery with library learning

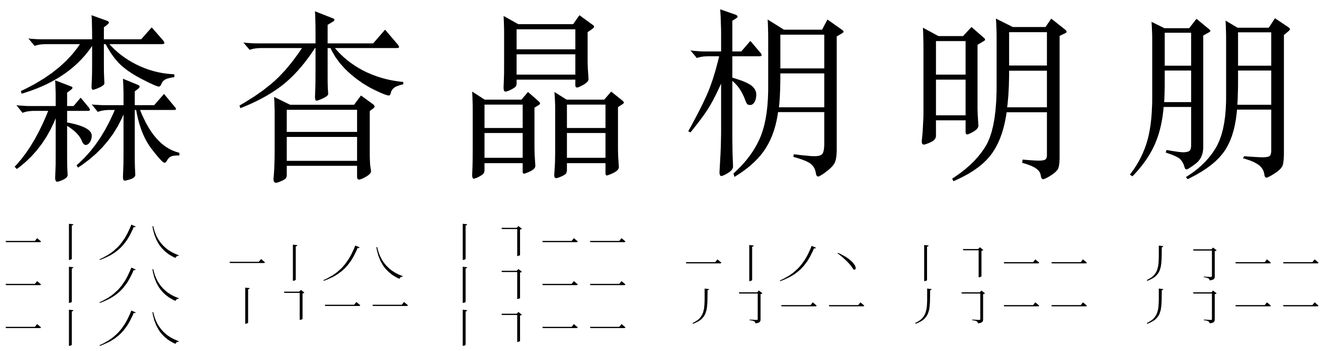
## Library learning as finding MDL (minimum description length)

森 杳 晶 枏 明 朋

$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)

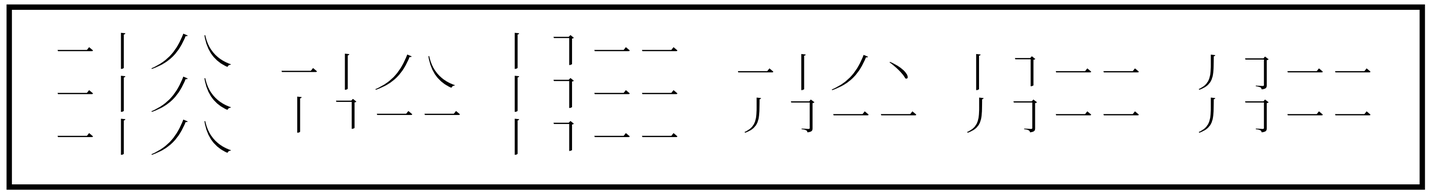


$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)

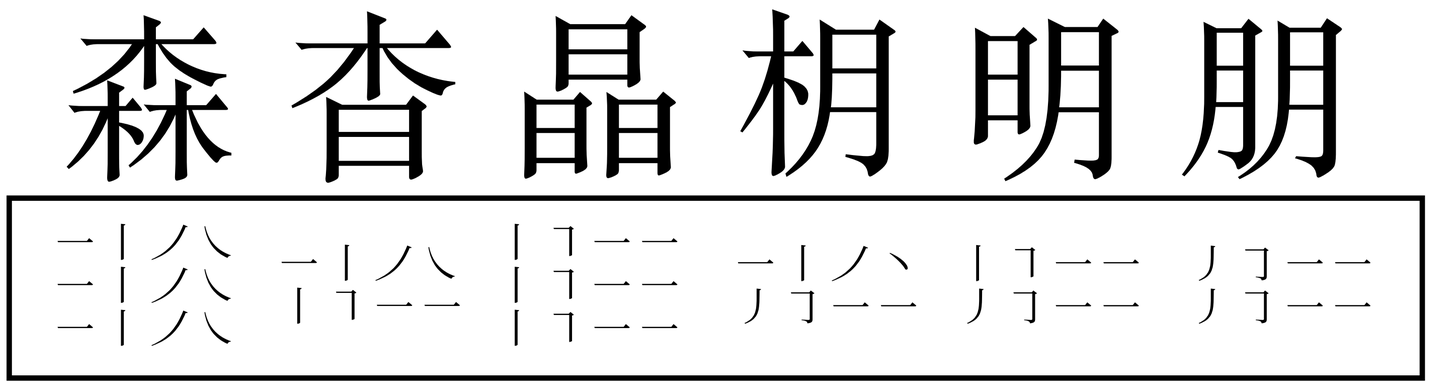
森 杳 晶 枡 明 朋



$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)



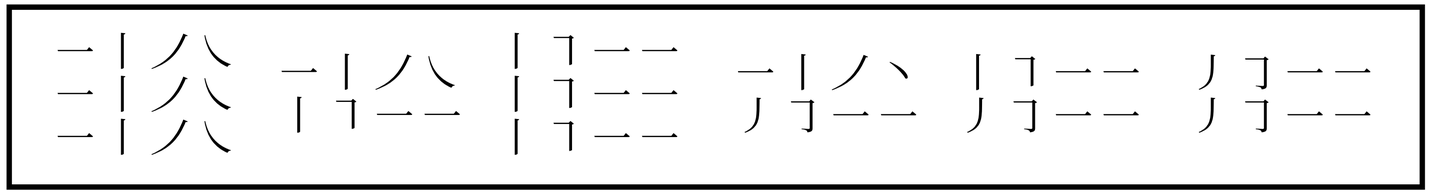
$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

**literal length** = 56
(#strokes used)

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)

森 杳 晶 栂 明 朋



**Raw DL = 56**

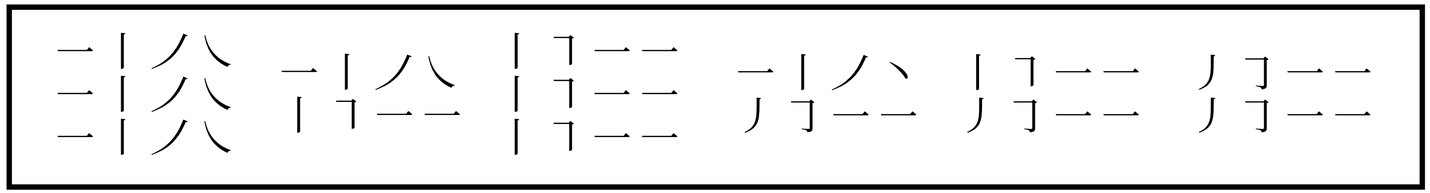
$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

**literal length** = 56
(#strokes used)

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)

森 杳 晶 枏 明 朋



**Raw DL = 56**

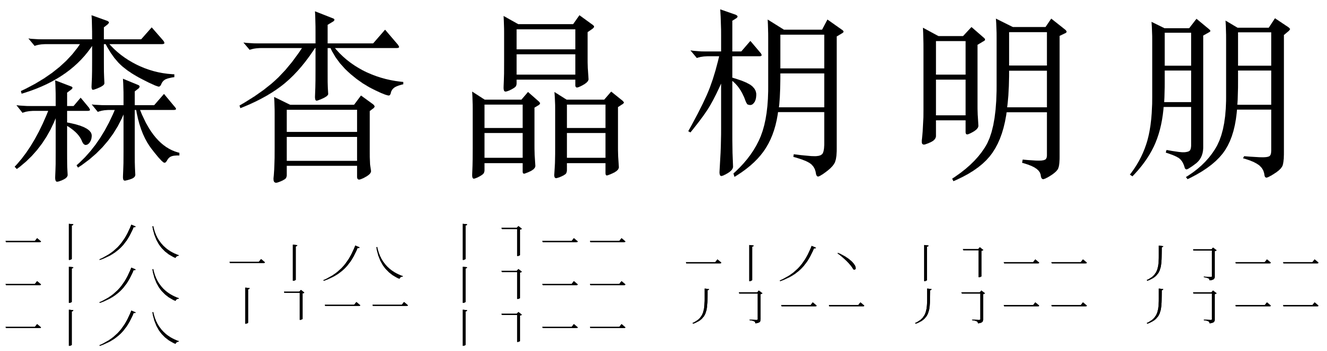
$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

**literal length** = 56
(#strokes used)

Our goal: utilize a library (or a vocabulary) of patterns to efficiently represent the stroke sequences.

# Structure discovery with library learning

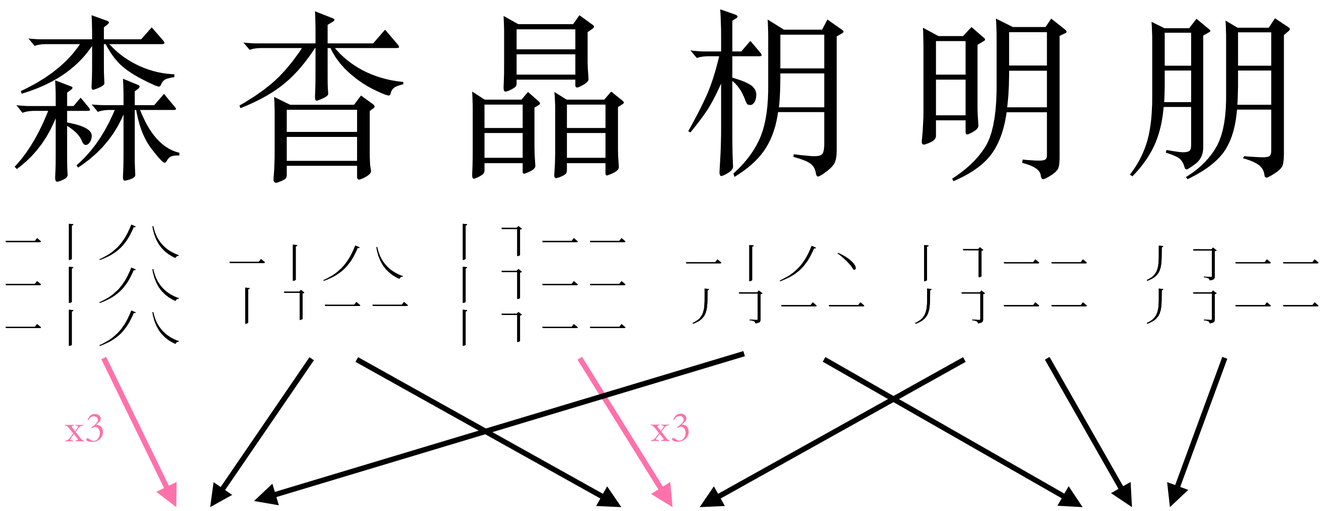
## Library learning as finding MDL (minimum description length)



$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

# Structure discovery with library learning

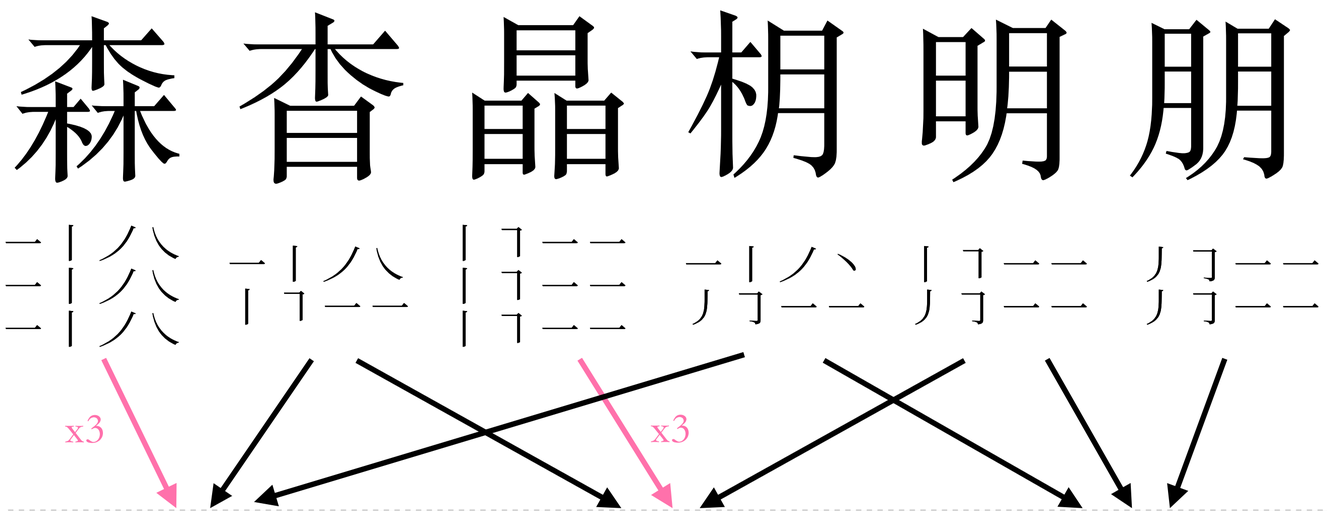
## Library learning as finding MDL (minimum description length)




$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

# Structure discovery with library learning

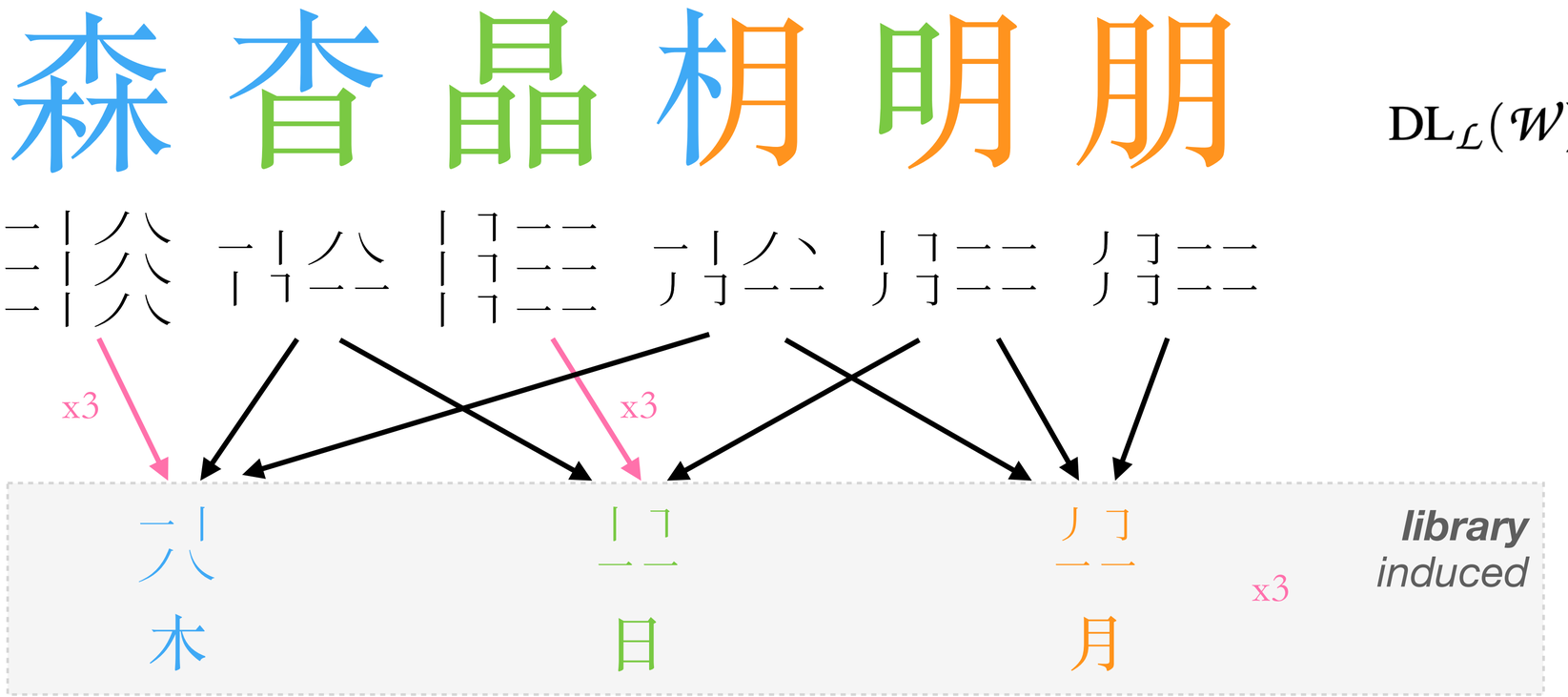
## Library learning as finding MDL (minimum description length)




$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

# Structure discovery with library learning

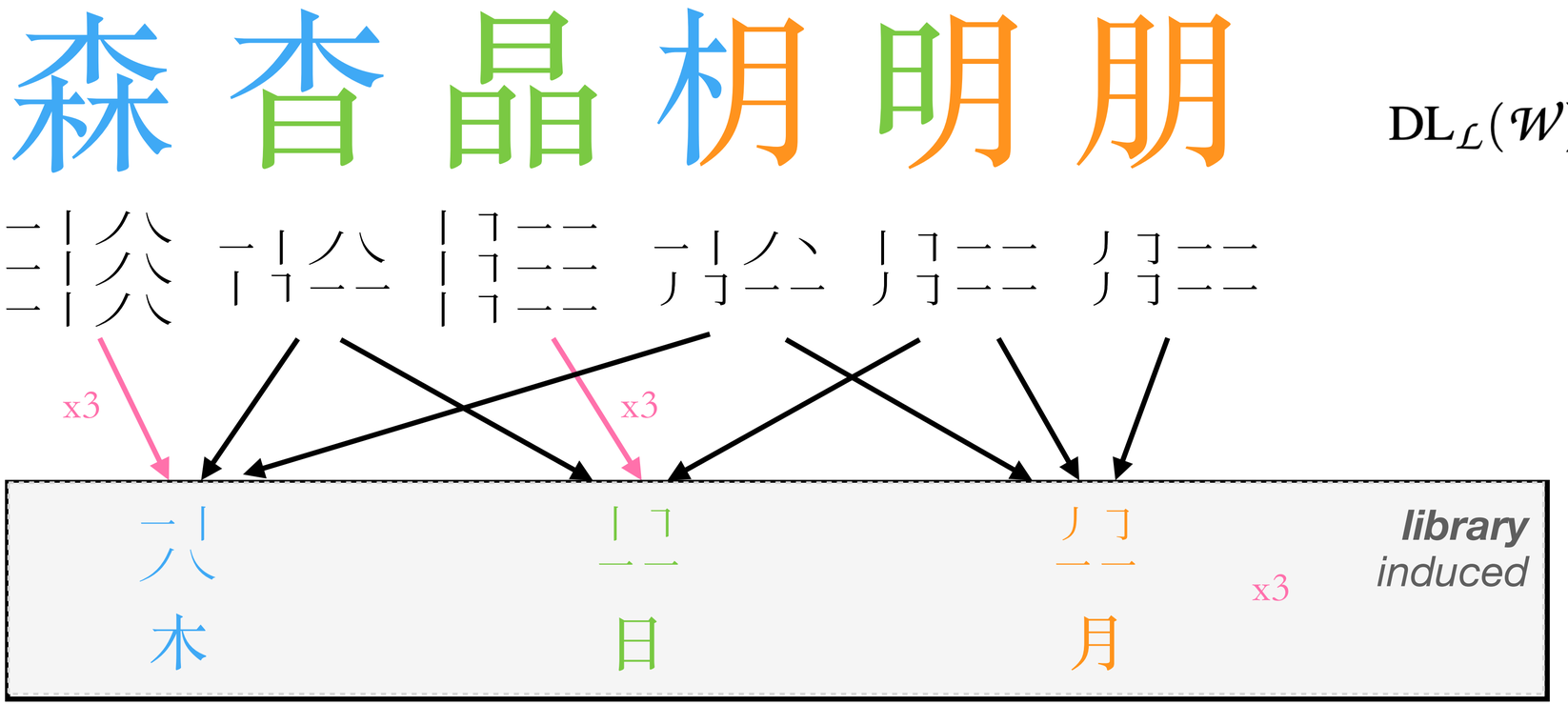
## Library learning as finding MDL (minimum description length)




$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)



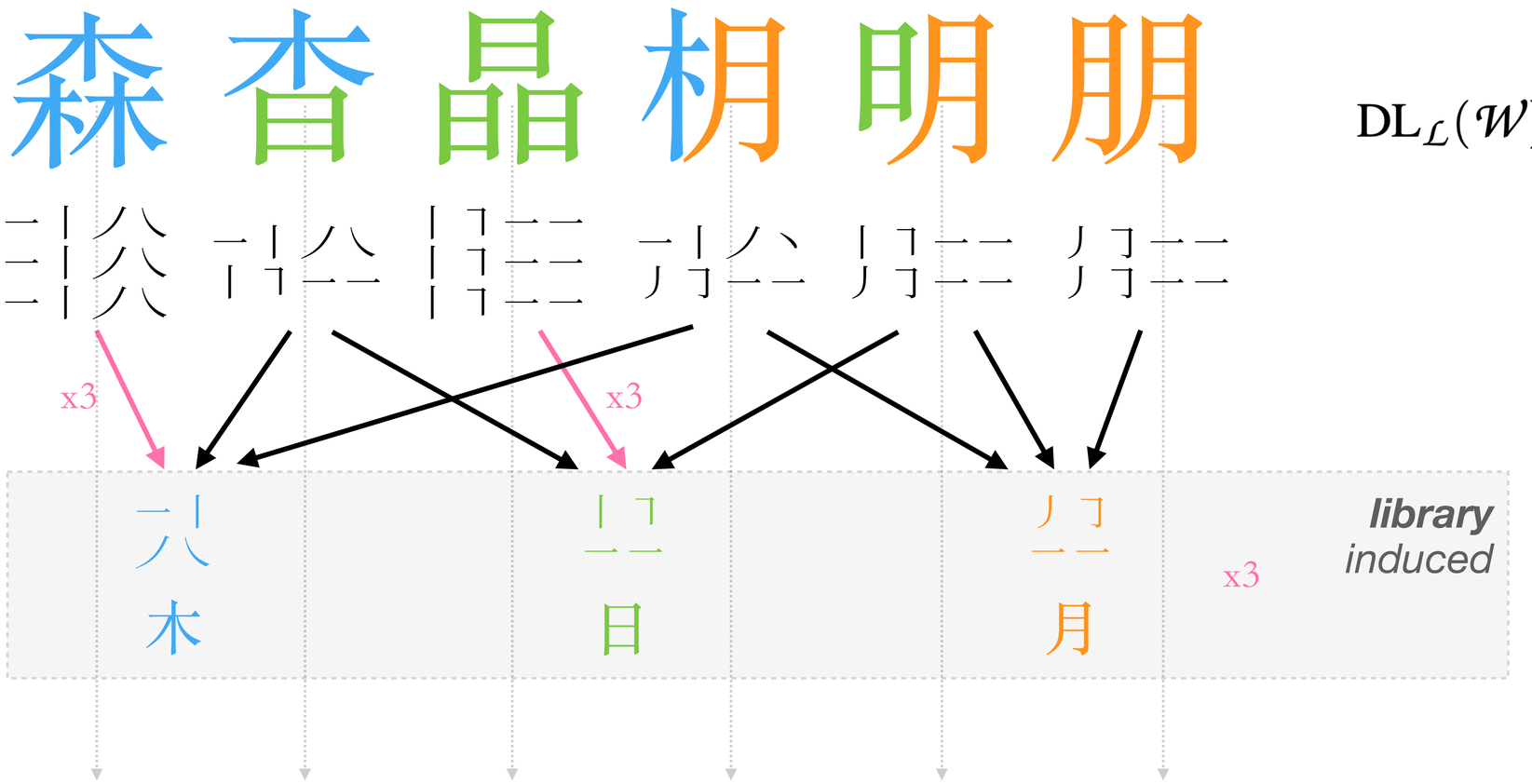

$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}}$$

A library of recurring patterns discovered.

How do we know it is a good library of patterns?

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)



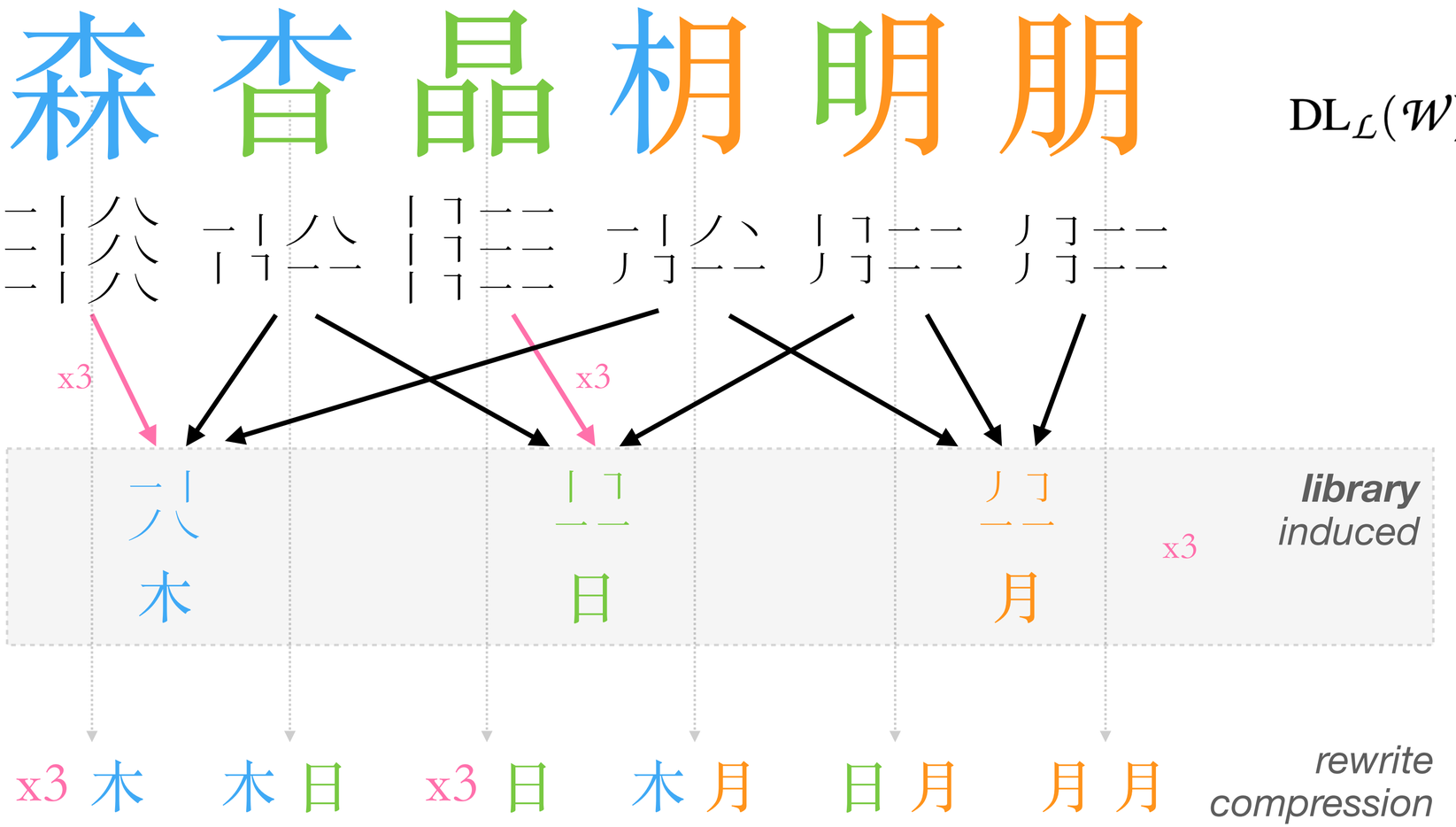

$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

$$+ \quad \overbrace{\mathrm{DL}\left(\mathcal{L}\right)}^{\text{description length of the library}}$$

$$\mathrm{DL}\left(\mathcal{L}\right) = \sum_{\texttt{fn} \in \mathcal{L}} \mathrm{DL}\left(\textsc{Body}(\texttt{fn})\right)$$

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)



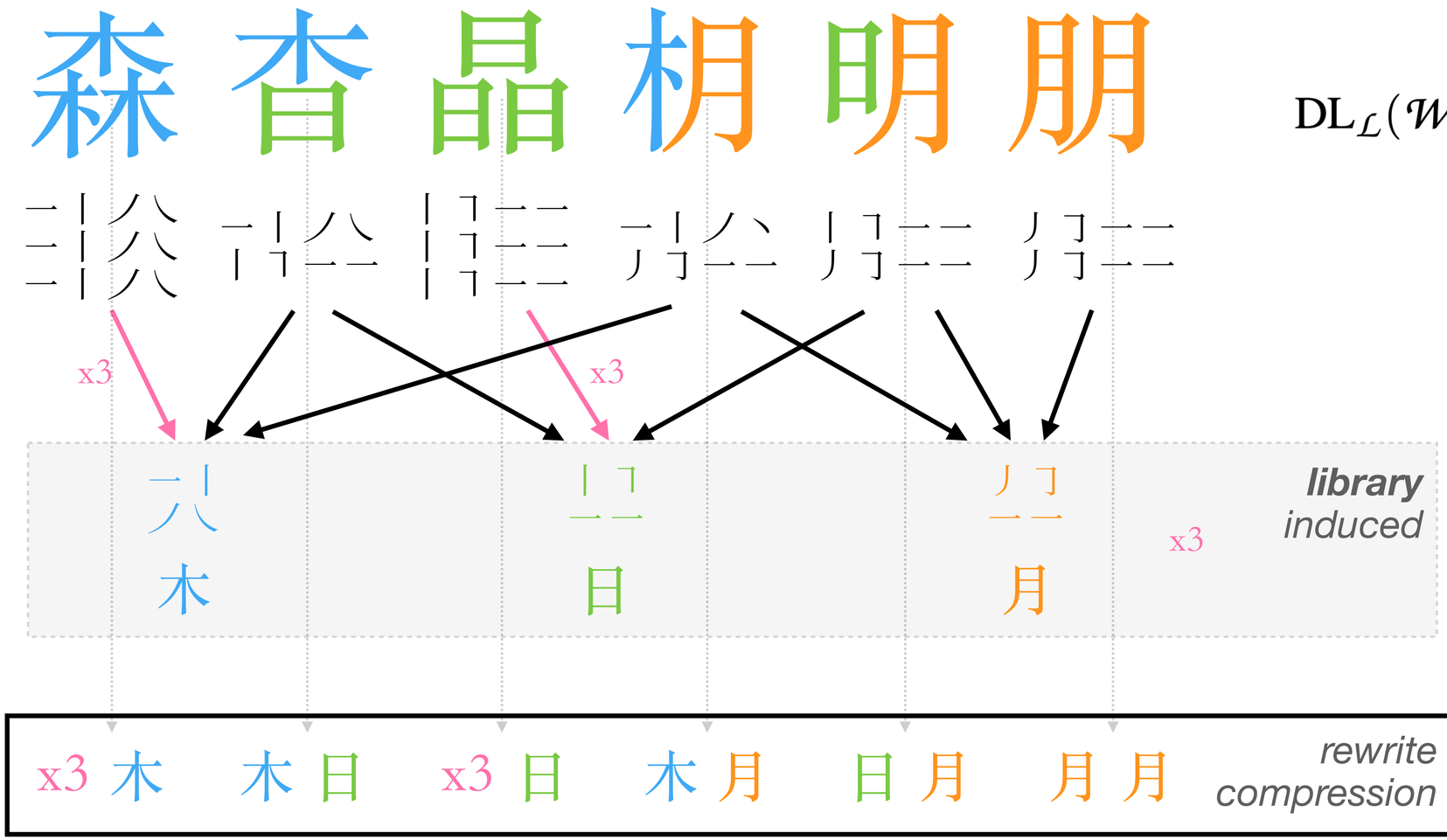

$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

$$+ \quad \overbrace{\mathrm{DL}\left(\mathcal{L}\right)}^{\text{description length of the library}}$$

$$\mathrm{DL}\left(\mathcal{L}\right) = \sum_{\texttt{fn} \in \mathcal{L}} \mathrm{DL}\left(\textsc{Body}(\texttt{fn})\right)$$

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)




$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

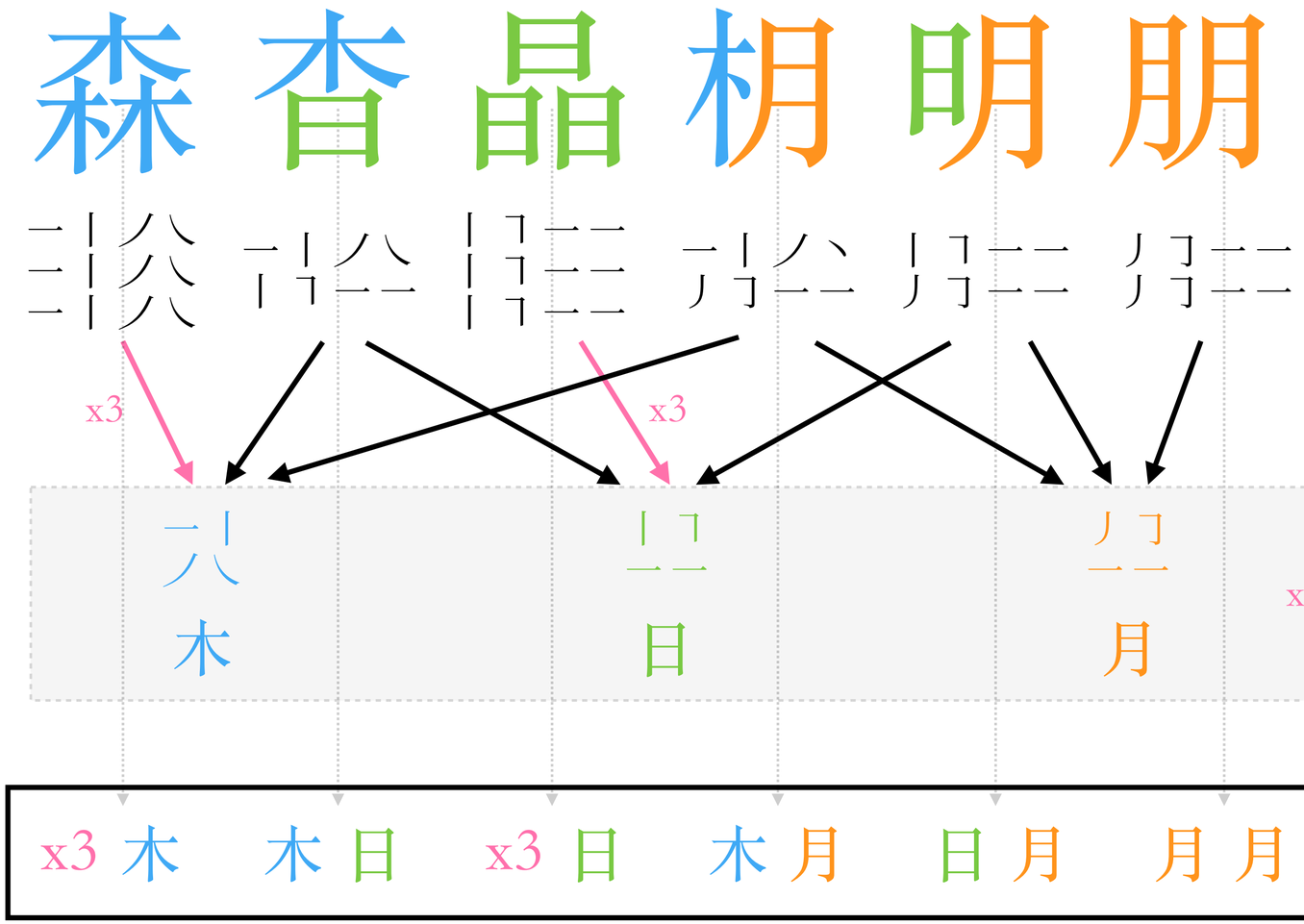
**compressed length** = 12
(#primitives used)

$$+ \quad \overbrace{\mathrm{DL}\left(\mathcal{L}\right)}^{\text{description length of the library}}$$

$$\mathrm{DL}\left(\mathcal{L}\right) = \sum_{\texttt{fn} \in \mathcal{L}} \mathrm{DL}\left(\textsc{Body}(\texttt{fn})\right)$$

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)




$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

**compressed length** = 12
(#primitives used)

$$+ \quad \overbrace{\mathrm{DL}\left(\mathcal{L}\right)}^{\text{description length of the library}}$$

$$\mathrm{DL}\left(\mathcal{L}\right) = \sum_{\texttt{fn} \in \mathcal{L}} \mathrm{DL}\left(\textsc{Body}(\texttt{fn})\right)$$

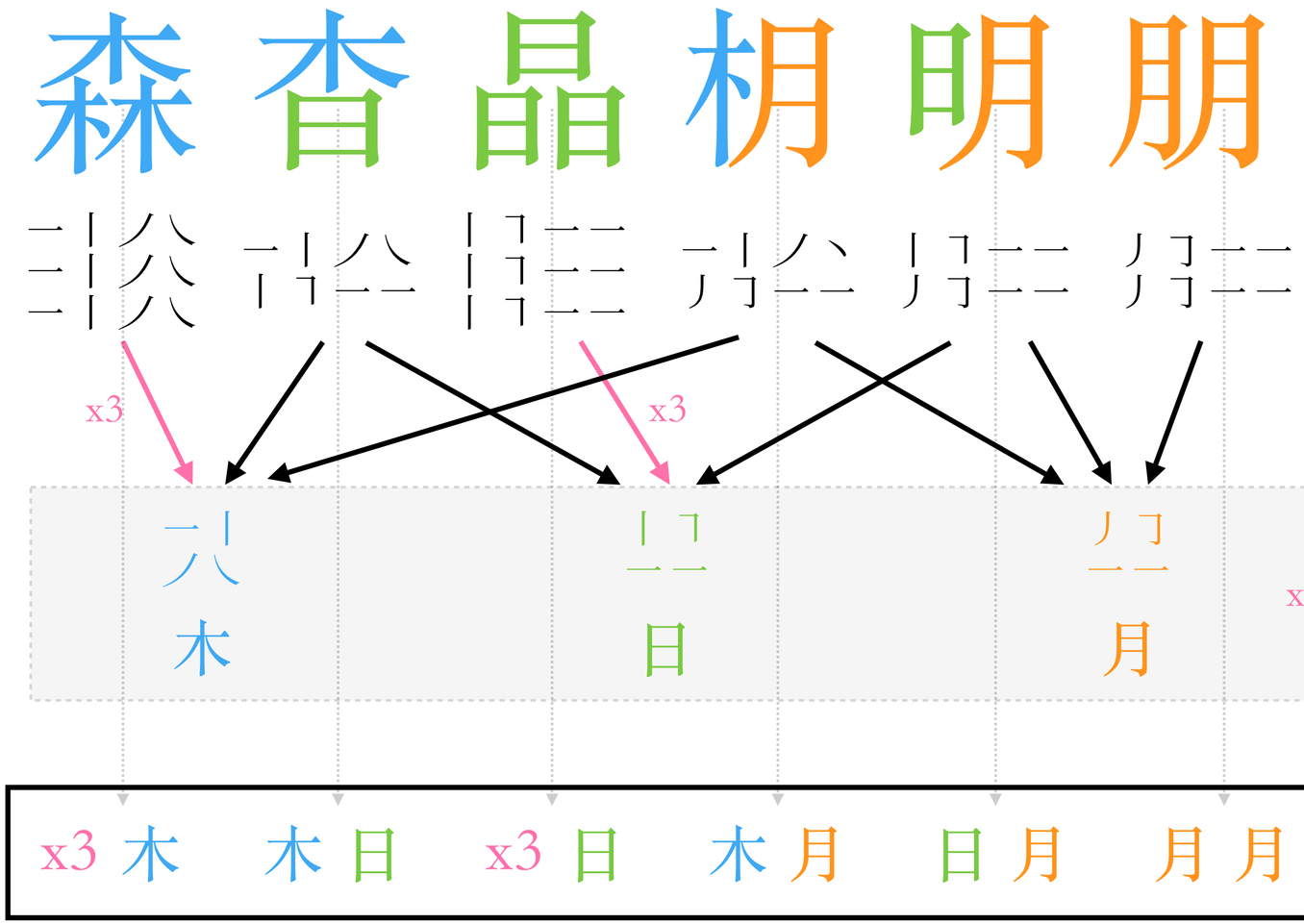
**library overhead** = 13
(sum of #primitives used in every abstraction)

# Structure discovery with library learning

## Library learning as finding MDL (minimum description length)

**Compressed DL = 25 (12 + 13)**

**Raw DL = 56**




$$\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) = \overbrace{\sum_{p \in P_{\mathcal{L}_{base}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\text{description length of the rewritten characters}}$$

**compressed length** = 12
(#primitives used)

$$+ \quad \overbrace{\mathrm{DL}(\mathcal{L})}^{\text{description length of the library}}$$

$$\mathrm{DL}(\mathcal{L}) = \sum_{\texttt{fn} \in \mathcal{L}} \mathrm{DL}\left(\textsc{Body}(\texttt{fn})\right)$$

**library overhead** = 13
(sum of #primitives used in every abstraction)

# At scale analysis of simplified Chinese script

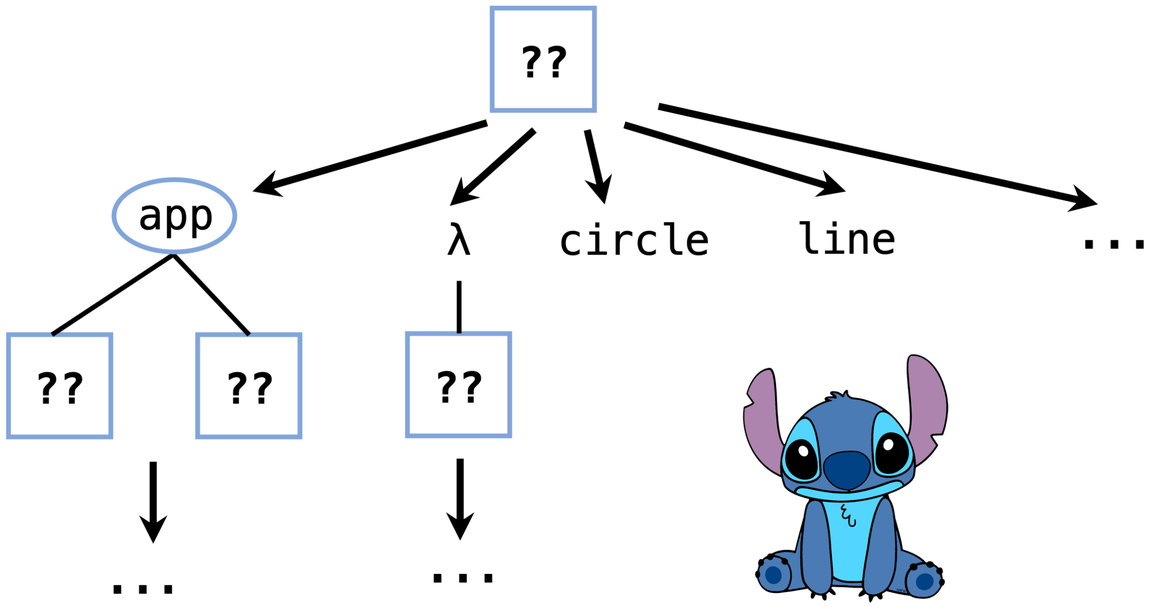
| | oracle | seal | traditional | simplified |
|---|---|---|---|---|
| *sink* |  |  | 瀋 | 沈 |
| *float* |  |  | 漂 | 漂 |
| *color* |  | 彩 | 綵 | 彩 |
| *insect* |  | 昆 | 崑 | 昆 |
| *orange* |  | 桔 | 桔 | 桔 |
| *peace* |  |  | 寧 | 宁 |

1500 BC 1050 BC 200 AD 1950 AD

# At scale analysis of simplified Chinese script

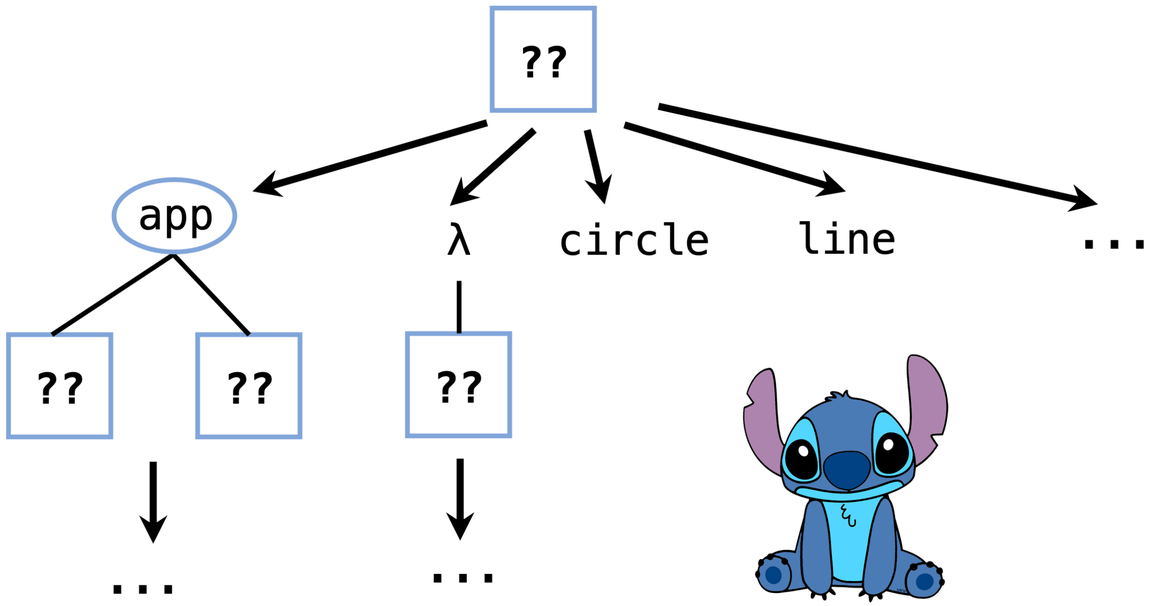
# Scaling up to the simplified Chinese script

- Leverage the **Stitch** (Bowers et al., 2023; also see DreamCoder) for efficiently discovering library functions.



# Scaling up to the simplified Chinese script

- Leverage the **Stitch** (Bowers et al., 2023; also see DreamCoder) for efficiently discovering library functions.

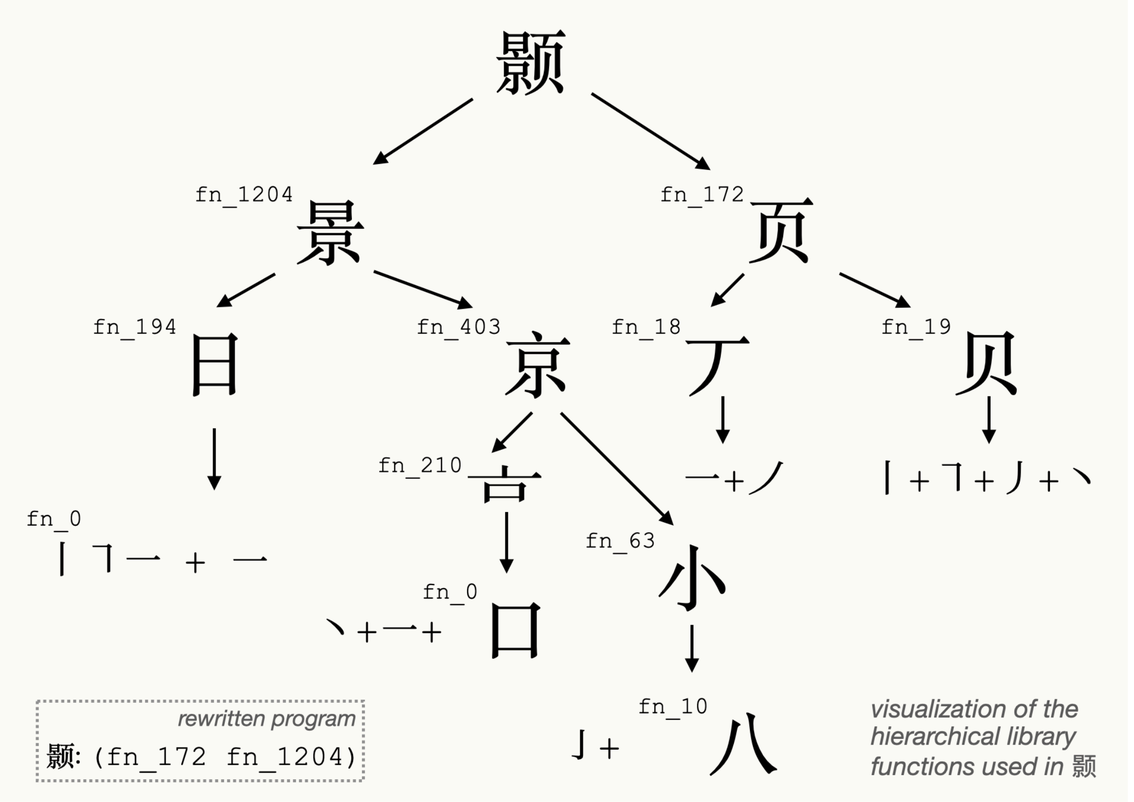



- 6,596 simplified Chinese characters. Represented as **programs**.

Our model rediscovers widely recognized theories of combintorial structure in the Chinese orthography.

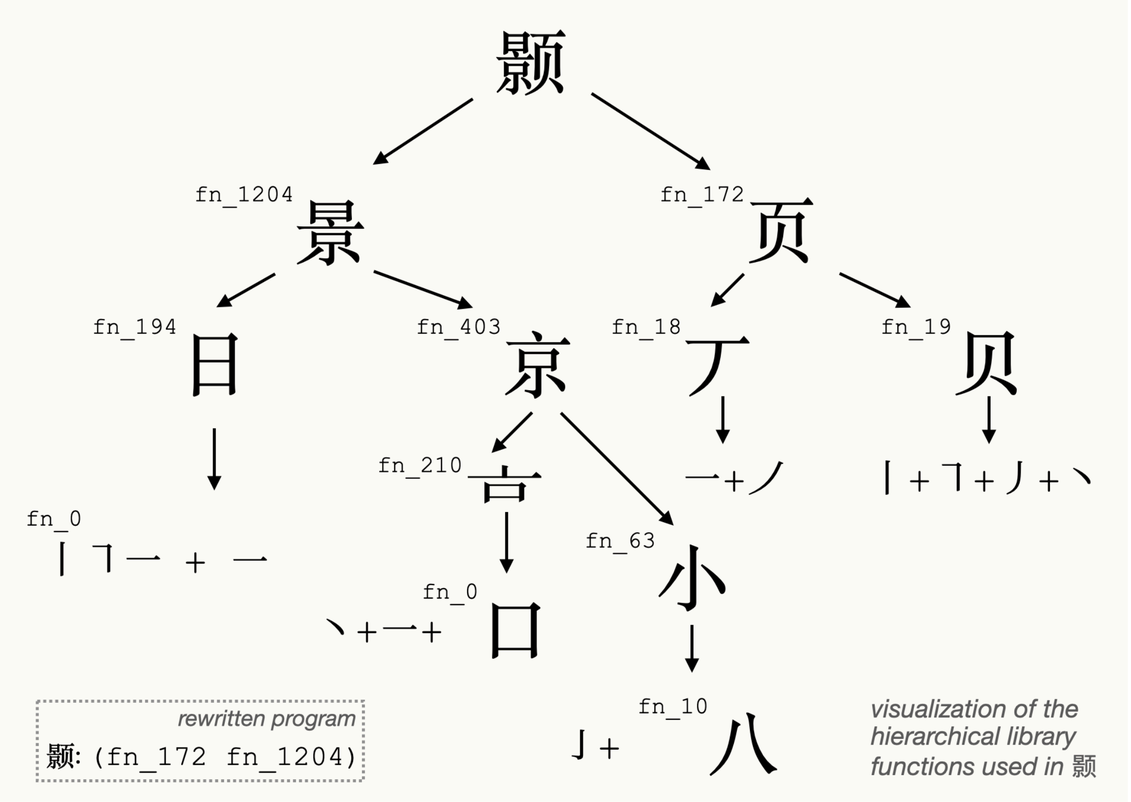
# What are the library functions learned?

hierarchically defined **graphical components**



# What are the library functions learned?

hierarchically defined **graphical components**




combinatorial **patterns** and **templates**

`fn_1136(#0) :=(#0 (#0 (#0 list)))` repeat `#0` 3 times

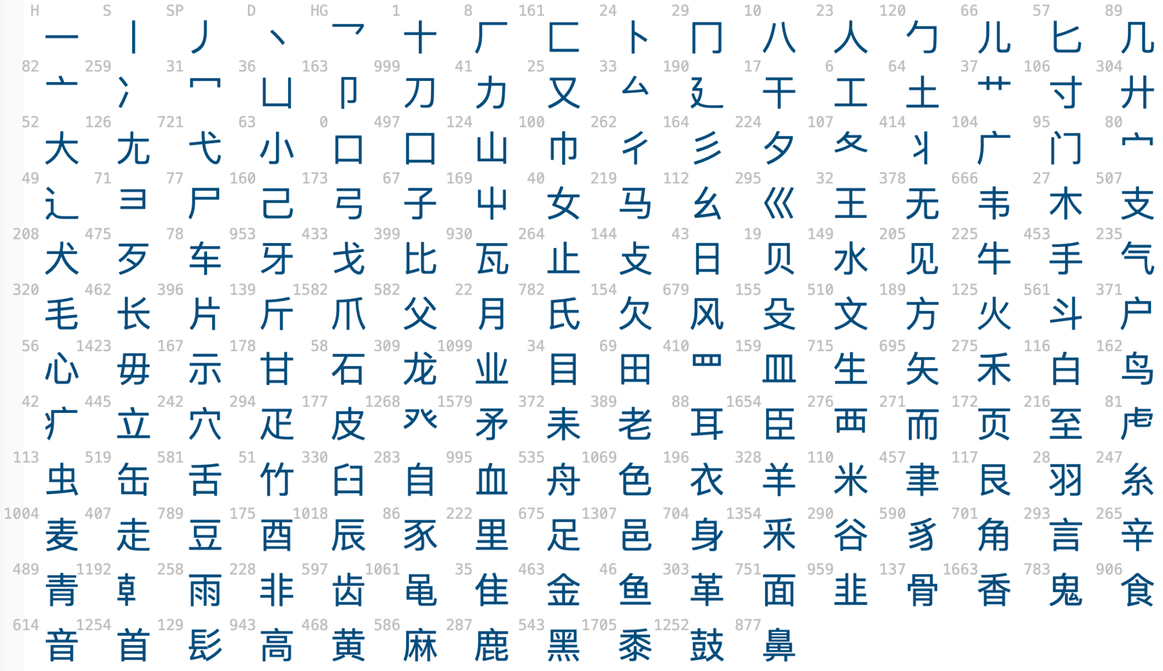
`fn_577(#0,#1):=(lambda (#0 (#0 (#0 #1))))` repeat `#0` 3 times + one radical

`fn_1135(#0,#1):=(lambda ($0 ($1 丶 冖) 一 丨 丿 乀))` fixed part + two radicals

# Library functions resemble expert-defined radicals

## Can library learning models uncover the structural theories underlying the Chinese language?

### Discovered and aligned radicals (187 / 201)



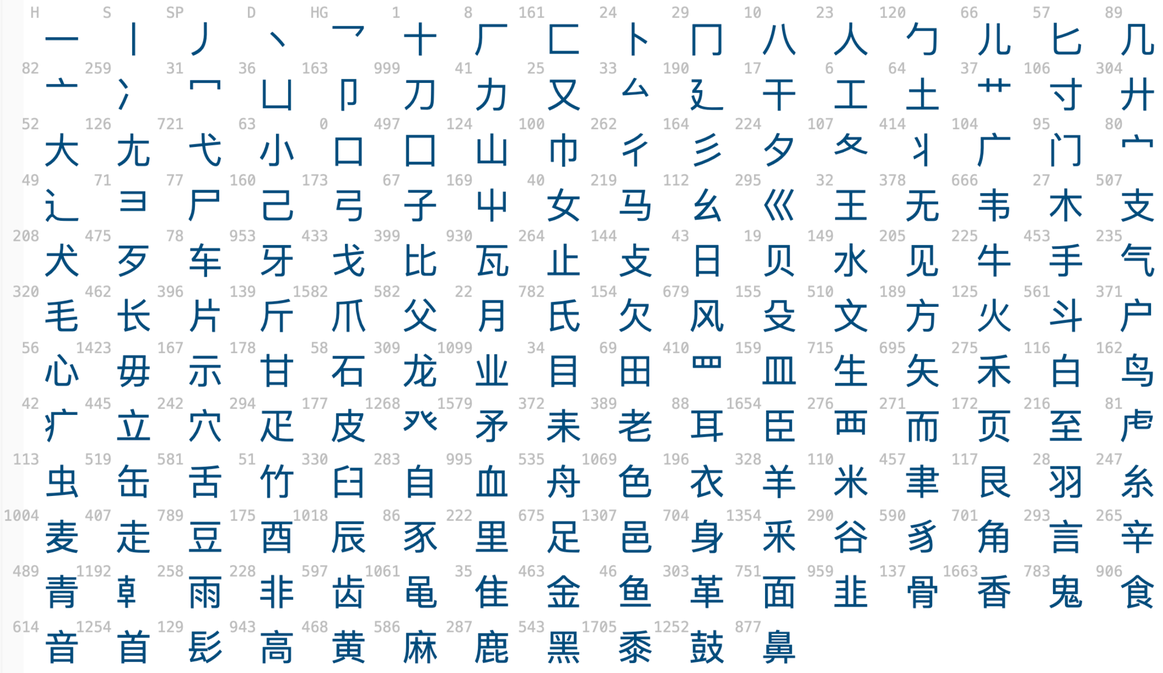

### Radicals failed to discover (14 / 201)

飞 瓜 肉 齐 赤 卤 龟 阜 隶 鬲 鬥 鼎 鼠 龠

# Library functions resemble expert-defined radicals

**Can library learning models uncover the structural theories underlying the Chinese language?**

**Discovered and aligned radicals (187 / 201)**




**Radicals failed to discover (14 / 201)**

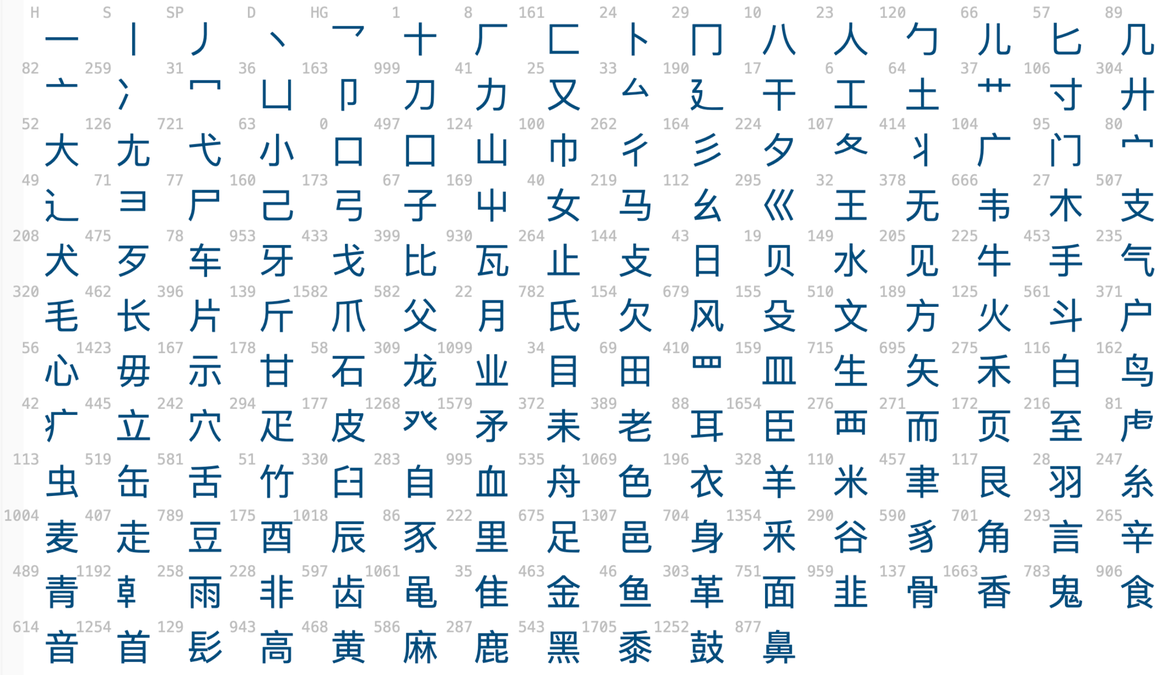
飞 瓜 肉 齐 赤 卤 龟 阜 隶 鬲 鬥 鼎 鼠 龠

- Our model **discovered** 187 **(93.0%) radicals defined by experts**.

# Library functions resemble expert-defined radicals

**Can library learning models uncover the structural theories underlying the Chinese language?**

**Discovered and aligned radicals (187 / 201)**



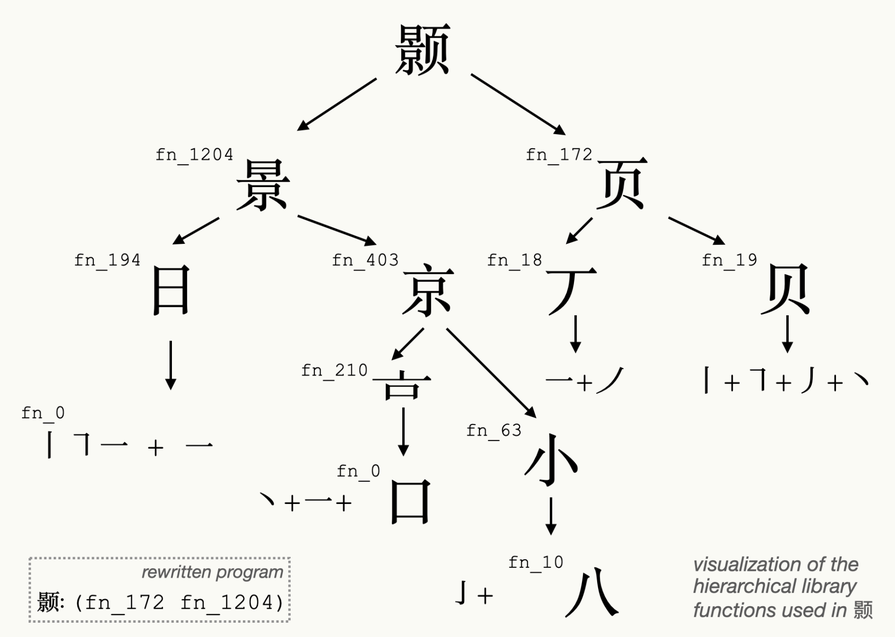
**Radicals failed to discover (14 / 201)**

飞 瓜 肉 齐 赤 卤 龟 阜 隶 鬲 鬥 鼎 鼠 龠

- Our model **discovered** 187 **(93.0%) radicals defined by experts**.

- Recap: radicals are
  - **Graphical components discovered by experts** that **frequently occur** in Chinese characters.

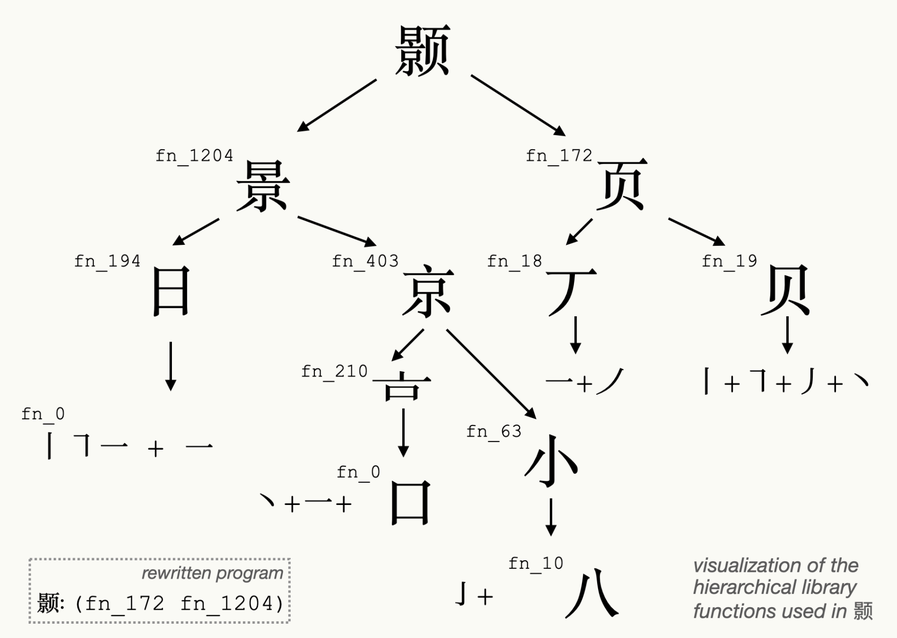
# Learned library captures the hierarchical organization of simplified Chinese characters




| Model | $F_1$ |
|---|---|
| Library learning | 61.6 |
| Baselines | |
| – Balanced binary tree | 34.4 |
| – Random binary tree | 28.5 |
| – Left-branching tree | 30.8 |
| – Right-branching tree | 36.0 |

- **Compared to gold standard.**
- Scores calculated over spans of the parsed trees.

# Learned library captures the hierarchical organization of simplified Chinese characters

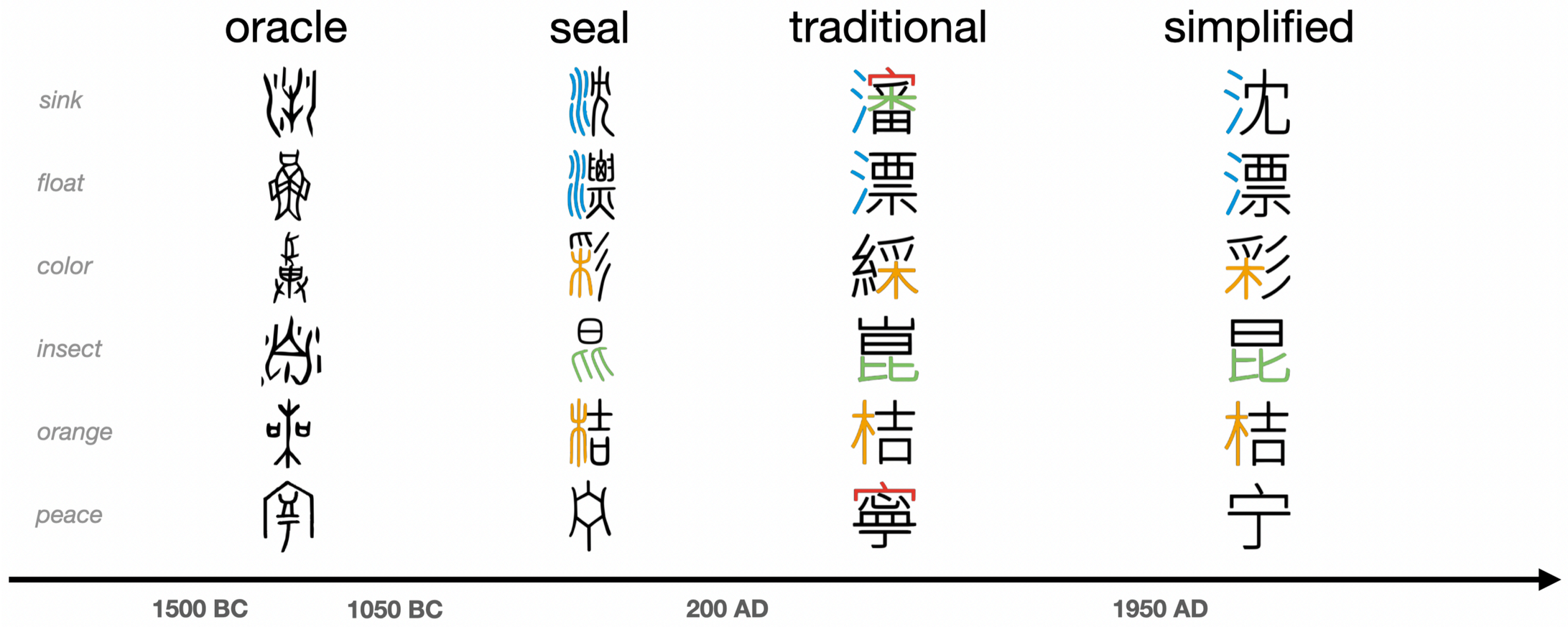



| Model | $F_1$ |
|---|---|
| Library learning | 61.6 |
| Baselines | |
| – Balanced binary tree | 34.4 |
| – Random binary tree | 28.5 |
| – Left-branching tree | 30.8 |
| – Right-branching tree | 36.0 |

- **Compared to gold standard.**
- Scores calculated over spans of the parsed trees.

# Diachronic analysis:

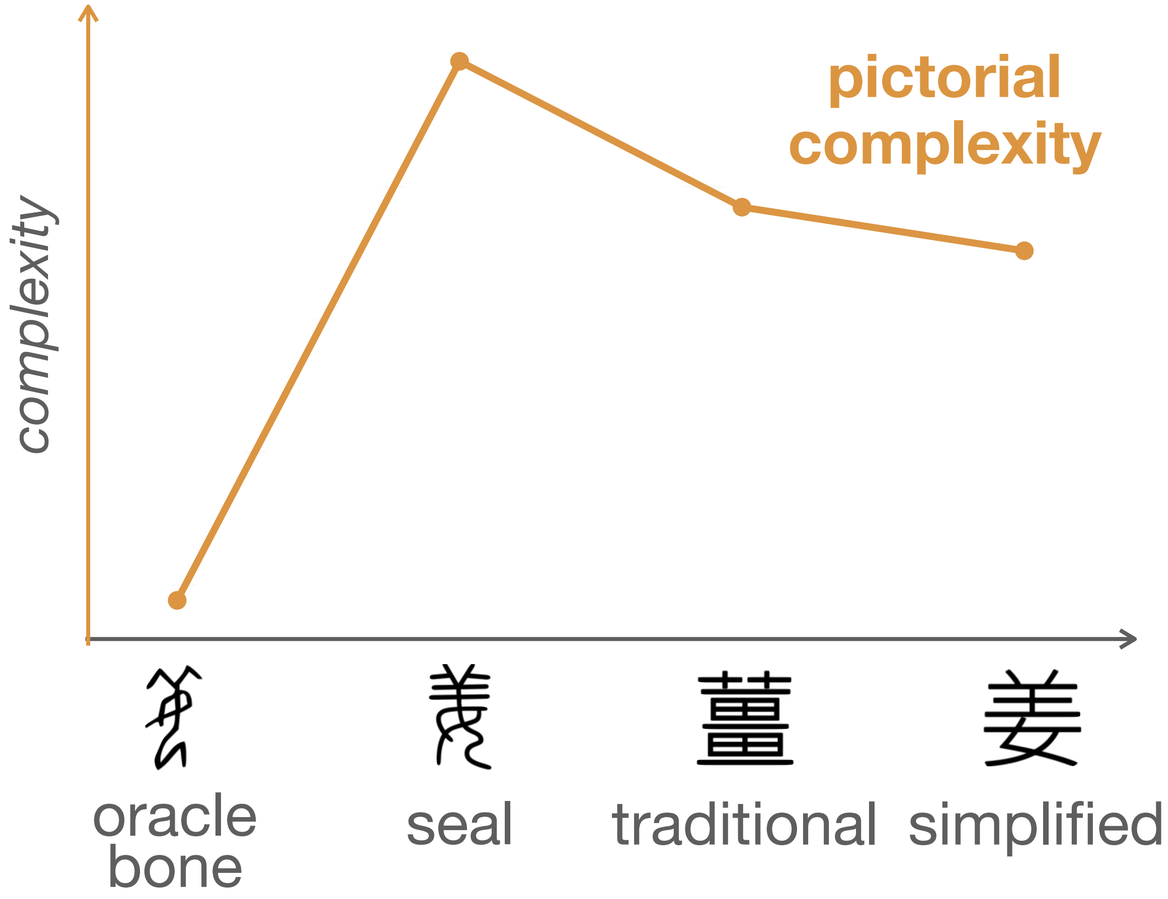
## How has the Chinese writing system evolved?




- A simplication? More efficient? Visually more complex?

# Diachronic analysis:

## How has the Chinese writing system evolved?

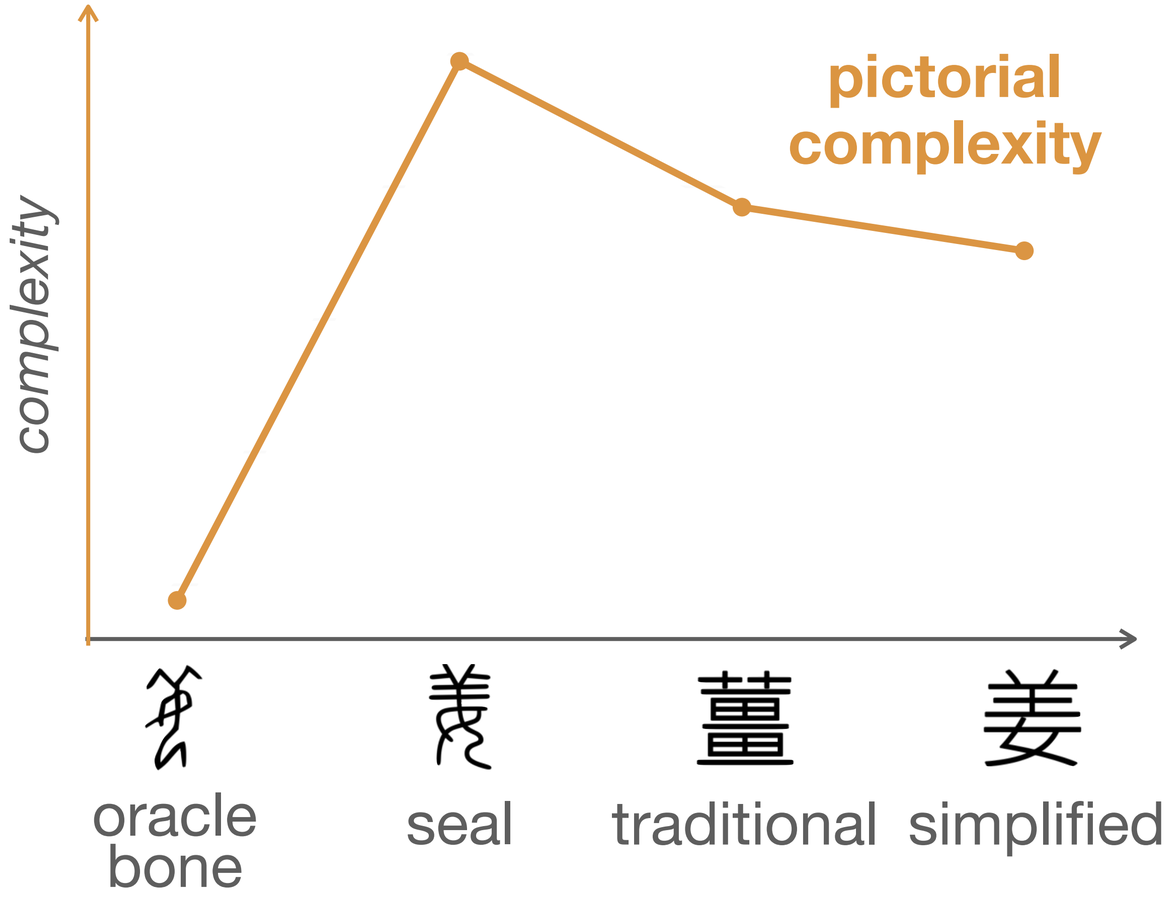



*pictorial (perimetric) complexity:* $C = \frac{P^2}{4\pi A}$

Widely used for drawings and simple shapes!

# Diachronic analysis:

## How has the Chinese writing system evolved?




*pictorial (perimetric) complexity:* $C = \frac{P^2}{4\pi A}$

Widely used for drawings and simple shapes!

Previous results based on **pictorial complexity did not** show a gradual simplification over time (Han et al., 2022).

# Is pictorial complexity enough?

- It is good for **simple drawings**.
- But is **not** capable of **capturing complexity and reuse** at a **system** level.

# Is pictorial complexity enough?

- It is good for **simple drawings**.
- But is **not** capable of **capturing complexity and reuse** at a **system** level.

# Is pictorial complexity enough?

- It is good for **simple drawings**.
- But is **not** capable of **capturing complexity and reuse** at a **system** level.

# Is pictorial complexity enough?

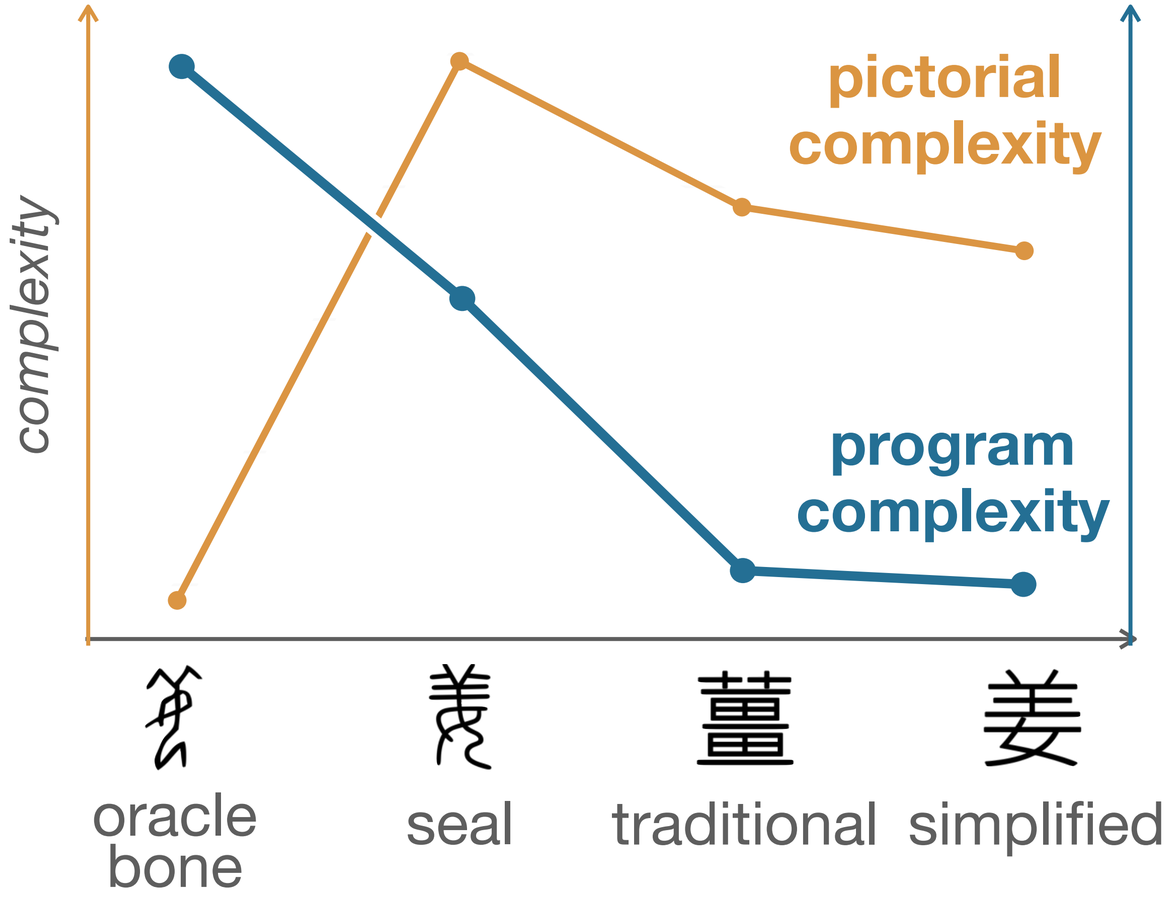
- It is good for **simple drawings**.
- But is **not** capable of **capturing complexity and reuse** at a **system** level.

These two characters have comparable pictorial complexities.
However, considering reuse and patterns, 矗 is much simpler.

as 矗 = 直 x 3

# Diachronic analysis:

## How has the Chinese writing system evolved?

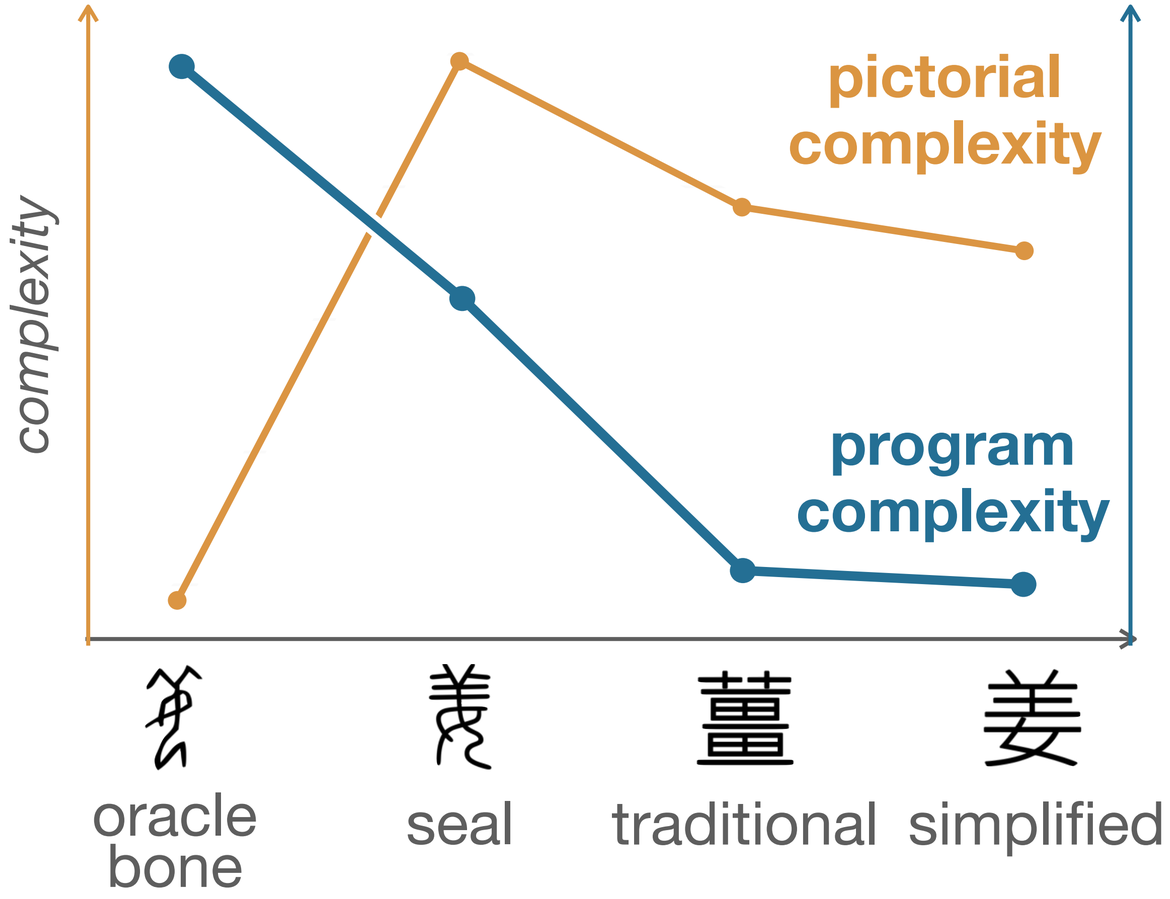



Our prediction:

the **library learning model** should **reveal a gradual simplification** as systems **adapt to these biases** in cultural evolution.

# Diachronic analysis:

## How has the Chinese writing system evolved?




Our prediction:

the **library learning model** should **reveal a gradual simplification** as systems **adapt to these biases** in cultural evolution.

Result:

**Program complexity** $C(\mathscr{W})$ has shown a **monotonic decrease** across time, confirming earlier empirical arguments.

# More on recent changes ~1950s

## Traditional Chinese -> Simplified Chinese

| | oracle | seal | traditional | simplified |
|---|---|---|---|---|
| *sink* |  |  | 瀋 | 沈 |
| *float* |  |  | 漂 | 漂 |
| *color* |  |  | 綵 | 彩 |
| *insect* |  |  | 崑 | 昆 |
| *orange* |  |  | 桔 | 桔 |
| *peace* |  |  | 寧 | 宁 |

1500 BC 1050 BC 200 AD 1950 AD

# More on recent changes ~1950s

## Traditional Chinese -> Simplified Chinese

| | traditional | simplified |
|---|---|---|
| *sink* | 瀋 | 沈 |
| *float* | 漂 | 漂 |
| *color* | 綵 | 彩 |
| *insect* | 崑 | 昆 |
| *orange* | 桔 | 桔 |
| *peace* | 寧 | 宁 |

# More on recent changes ~1950s

## Traditional Chinese -> Simplified Chinese


inconsistent one-to-multiple mapping

inconsistent simplification

- **A real simplification?**
  - This process may have **disrupted established systematicity** and lead to a loss of established semantic-phonetic and graphic patterns (Handel, 2013; Zhao & Baldauf, 2011).

# More on recent changes ~1950s

## Traditional Chinese -> Simplified Chinese

observation
*(traditional ⇒ simplified)*

開 ⇒ 开
鏘 ⇒ 锵

盧 ⇒ 卢
爐 ⇒ 炉

蔣 ⇒ 蒋
獎 ⇒ 奖

*(non-systematic rules observed)*

simplication rules

1 開 ⇒ 开 / 闹

inconsistent one-to-multiple mapping

2 盧 ⇒ 卢 / 户

3 將 ⇒ 将 / 妆

inconsistent simplification

# More on recent changes ~1950s

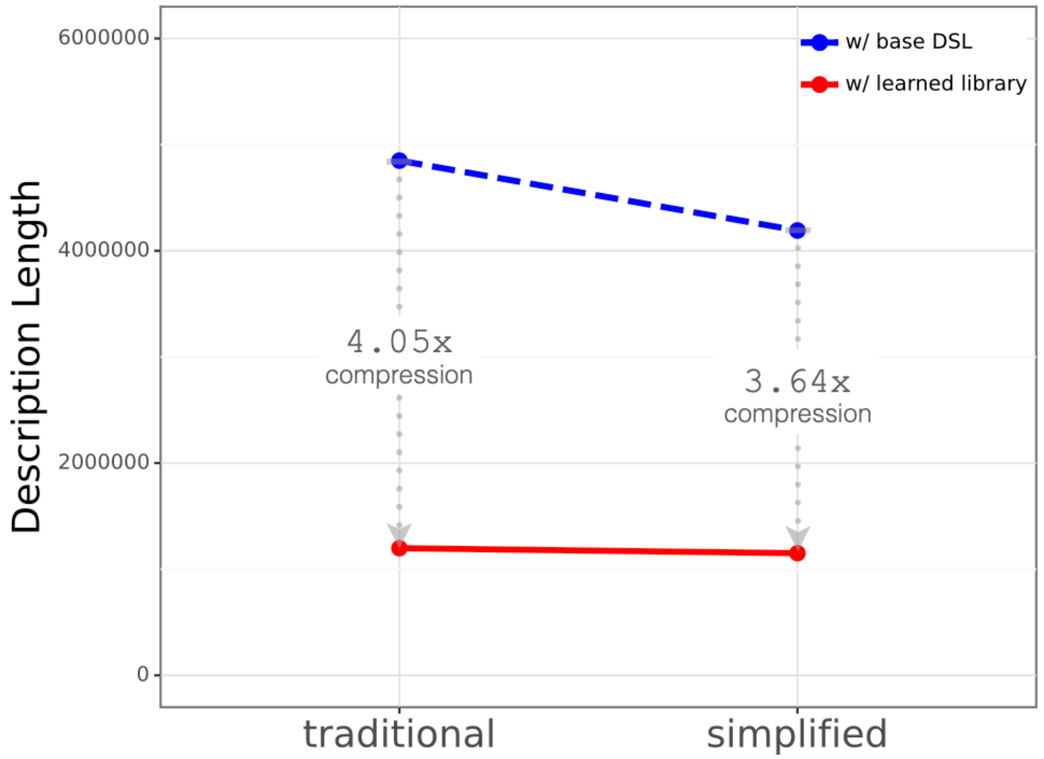
## Traditional Chinese -> Simplified Chinese


inconsistent one-to-multiple mapping

inconsistent simplification

- Can our computational model provide concrete evidence of the loss of systematicity?
- Our prediction:
  - **Systematic scripts should be more compressive.**

# Simplified Chinese is simpler but less systematic compared to traditional Chinese

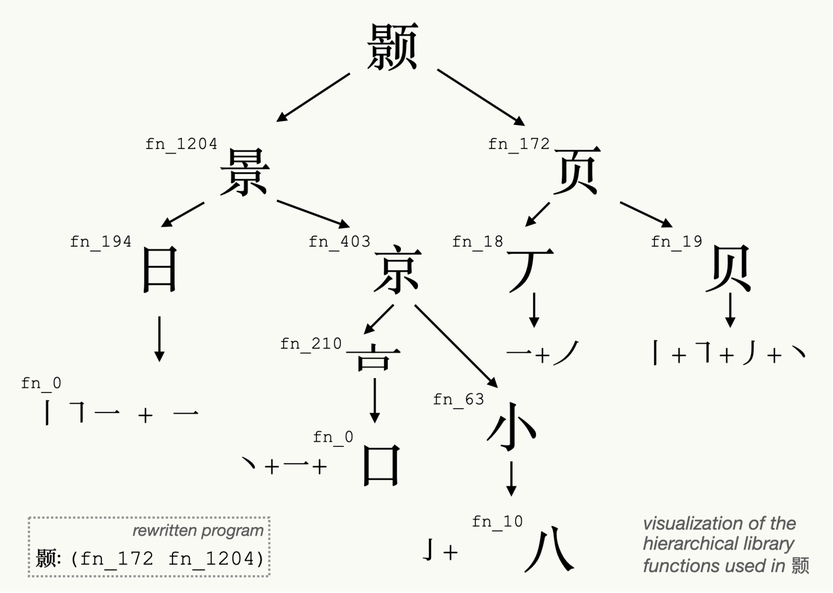



- Compression ratio (Raw DL / Compressed DL):
  - **Traditional > Simplified**
  - Suggesting the simplification process **did break** part of the systematicity.

# Conclusions & Takeaway

- **A library learning-based** computational model can reveal the **inductive biases** behind the **emergence and evolution of combinatorial structures** in human language.
- **Combinatoriality**
  - **Develops** from a **MDL** perspective of representational efficiency
  - **By discovering inventories of reusable parts**
  - **By compressing the language**

# Conclusions & Takeaway

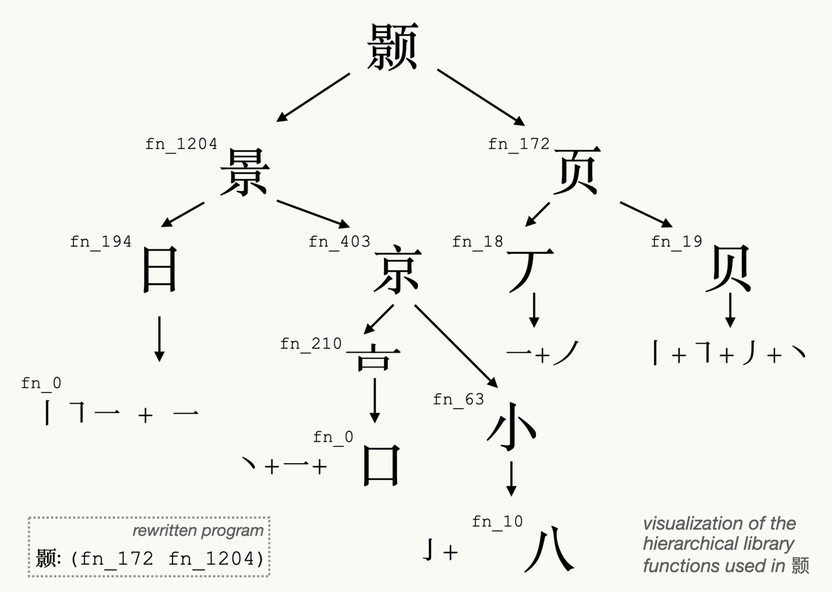



## ⬅️ Synchronic

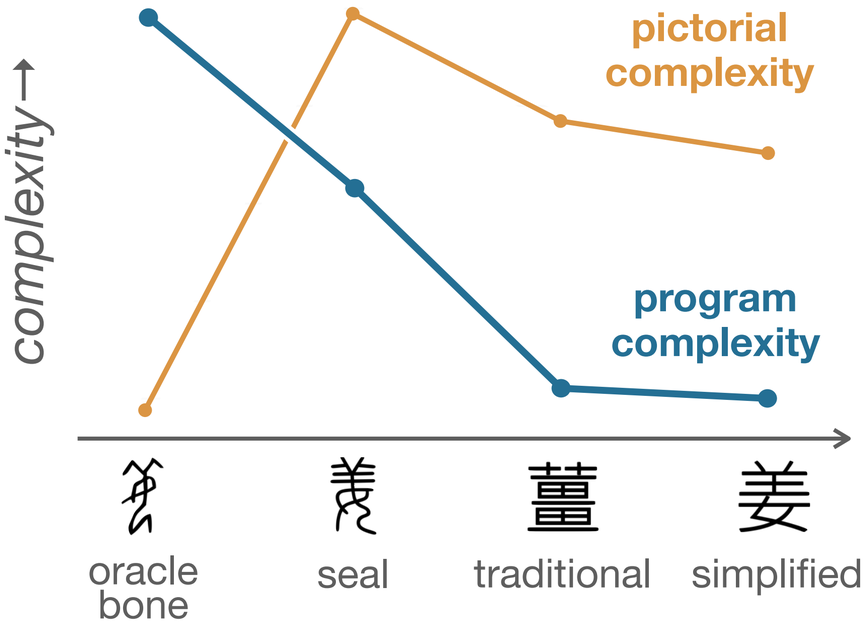
**Discovered and aligned radicals (187 / 201)**

一 丨 丿 丶 ㇇ 十 厂 匚 卜 冂 八 人 勹 儿 匕 几
亠 冫 冖 凵 卩 刀 力 又 厶 廴 干 工 土 艹 寸 廾
大 尢 弋 小 口 口 山 巾 彳 彡 夕 夂 丬 广 门 宀
辶 ⺕ 尸 己 弓 子 屮 女 马 幺 巛 王 无 韦 木 支
犬 歹 车 牙 戈 比 瓦 止 攴 日 贝 水 见 牛 手 气
毛 长 片 斤 爪 父 月 氏 欠 风 殳 文 方 火 斗 户
心 毋 示 甘 石 龙 业 目 田 罒 皿 生 矢 禾 白 鸟
疒 立 穴 疋 皮 癶 矛 耒 老 耳 臣 西 而 页 至 虍
虫 缶 舌 竹 臼 自 血 舟 色 衣 羊 米 聿 艮 羽 糸
麦 走 豆 酉 辰 豕 里 足 邑 身 釆 谷 豸 角 言 辛
青 ⻏ 雨 非 齿 黾 隹 金 鱼 革 面 韭 骨 香 鬼 食
音 首 髟 高 黄 麻 鹿 黑 黍 鼓 鼻

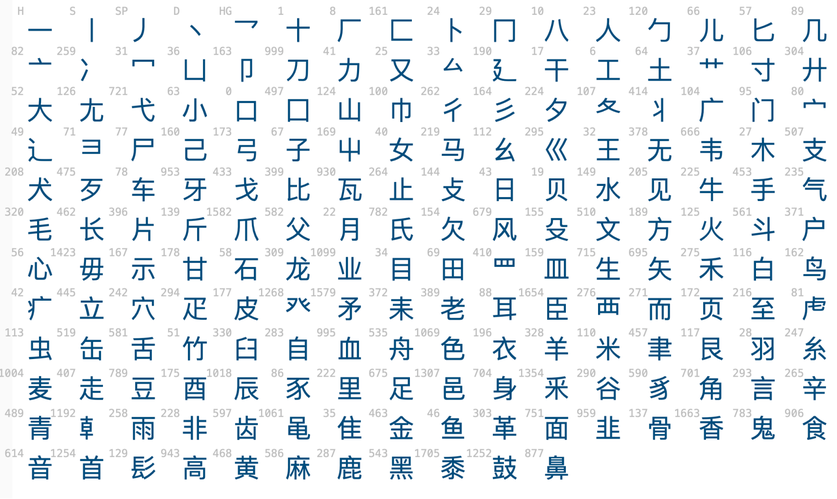
# Conclusions & Takeaway



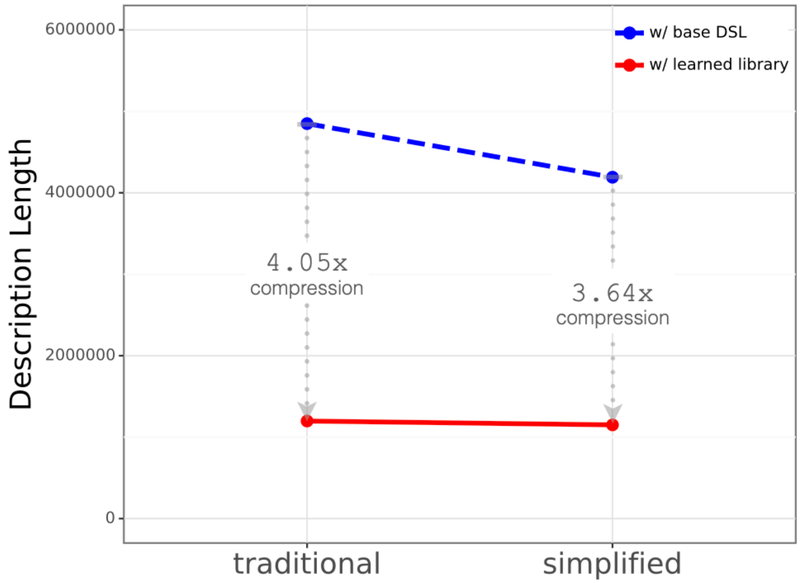

## ⬅️ Synchronic




**Discovered and aligned radicals (187 / 201)**



## Diachronic ➡️



# Future work

- Extend to meaning compositionality:
  - Logographics captures the **multi-level** structure: not only combinatoriality in **forms**, but also compositionality in **meanings**.
- A wider range of logographic languages:
  - Cuneiform, Vai script…
- Consider more factors: frequency, motor cost, iconicity, visual complexity, etc.

# Thanks!

I’m actively seeking PhD positions starting 25fall :)

Guangyuan

Paper