Finding structure in logographic writing with library learning II: Grapheme, sound, and meaning systematicity

Guangyuan Jiang[1,⋆], **Matthias Hofer**[1], **Jiayuan Mao**[2], **Lionel Wong**[3], **Joshua B. Tenenbaum**[1,2], and **Roger P. Levy**[1]

[1]MIT BCS [2]MIT CSAIL [3]Stanford University

⋆correspondence to jianggy@mit.edu

Abstract

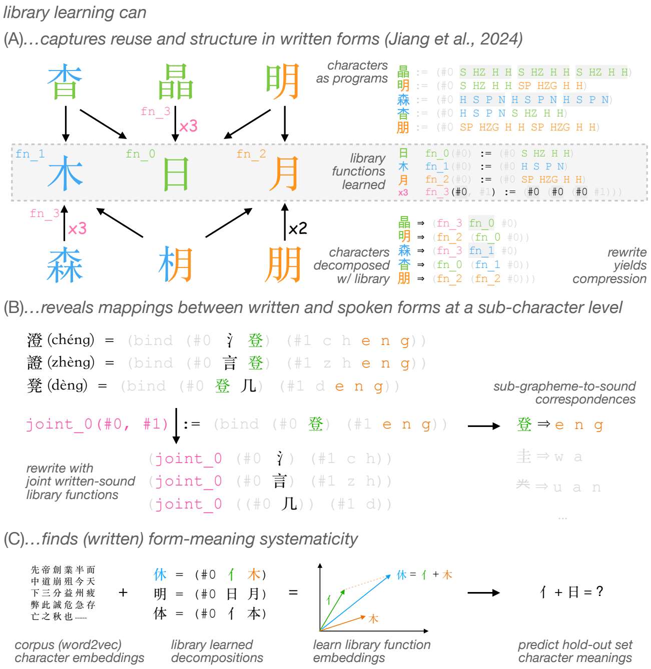
Writing systems are structured to depict the various facets of human language, from sounds to meanings. Chinese writing, as a logographic system, offers a distinctive opportunity to study the structural relationships between written forms and their sounds and meanings all at once. In this companion paper to Jiang et al. (2024), we explore a computational model based on library learning that can capture the compositional structure of Chinese characters and their relationship to sound and meaning. We extend the written-only library learning framework from Jiang et al. (2024) by incorporating written-sound joint compression and distributional semantic representations. The joint compression component allows the model to uncover structural relationships between a character's graphical components and its pronunciation, mirroring the function of phonetic and semantic radicals in Chinese orthography. With distributional semantics, the model also learns systematic links between the graphical structure and the meaning of characters, enabling it to predict the meanings of unseen characters based on their constituent parts. Moreover, our model allows us to explore historical shifts in how written Chinese has represented spoken language. We anticipate that our library learning model to be a unified computational account of writing's interaction with multi-level structures of human language.

Keywords: language learning; evolution; phonology; Bayesian modeling

Introduction

Human language is structured across different levels of representation, and writing is one key, permanent form of this structure. As cultural inventions, writing systems are generally thought to have been created to record spoken language, with their orthographies designed to capture both the sounds and meanings of the language (Chafe & Tannen, 1987; Sproat, 2000; Frost, 2012). But to what degree do writing systems represent the structure of both sound and meaning in language? Moreover, what computational principles govern the structural relations between writings and other aspects of language?

In this paper, we present a computational framework studying the systematic structural relations between writing, sound, and meaning in a language (Frost, 2012; Sproat, 2010). We focus on one of the oldest and most widely used writing systems—Chinese, alongside its most widely spoken Sinitic variety, Standard Mandarin. The Chinese writing system is traditionally recognized as logographic, and so, in contrast to phonetic writing, each graphical symbol (character) is associated with a semantic component, such as a word or morpheme (Gelb, 1963; Coulmas, 2003; Sproat & Gutkin, 2021). This logographic nature offers a unique opportunity to examine the relation between written form, sound, and meaning within a single language.


library learning can

(A)…captures reuse and structure in written forms (Jiang et al., 2024)

杳
晶
明
fn_3
x3
fn_1
木
fn_0
日
fn_2
月
fn_3
x3
x2
森
枂
朋
characters as programs
晶 := (#0 S HZ H H S HZ H H S HZ H H)
明 := (#0 S HZ H H SP HZG H H)
森 := (#0 H S P N H S P N H S P N)
杳 := (#0 H S P N S HZ H H)
朋 := (#0 SP HZG H H SP HZG H H)
library functions learned
日 fn_0(#0) := (#0 S HZ H H)
木 fn_1(#0) := (#0 H S P N)
月 fn_2(#0) := (#0 SP HZG H H)
x3 fn_3(#0, #1) := (#0 (#0 (#0 #1)))
characters decomposed w/ library
晶 ⇒ (fn_3 fn_0 #0)
明 ⇒ (fn_2 (fn_0 #0))
森 ⇒ (fn_3 fn_1 #0)
杳 ⇒ (fn_0 (fn_1 #0))
朋 ⇒ (fn_2 (fn_2 #0)))
rewrite yields compression

(B)…reveals mappings between written and spoken forms at a sub-character level

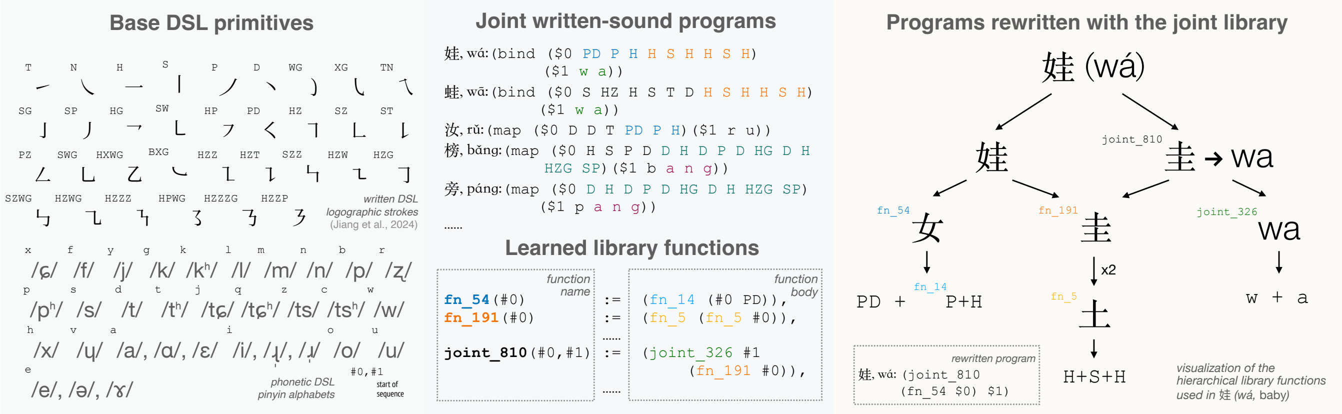
澄 (chéng) = (bind (#0 氵 登) (#1 c h e n g))
證 (zhèng) = (bind (#0 言 登) (#1 z h e n g))
凳 (dèng) = (bind (#0 登 几) (#1 d e n g))
sub-grapheme-to-sound correspondences
joint_0(#0, #1) := (bind (#0 登) (#1 e n g))
登 ⇒ e n g
圭 ⇒ w a
关 ⇒ u a n
…
rewrite with joint written-sound library functions
(joint_0 (#0 氵) (#1 c h))
(joint_0 (#0 言) (#1 z h))
(joint_0 ((#0 几)) (#1 d))

(C)…finds (written) form-meaning systematicity

先帝創業半而
中道崩殂今天
下三分益州疲
弊此誠危急存
亡之秋也……
+
休 = (#0 亻 木)
明 = (#0 日 月)
体 = (#0 亻 本)
=
亻
休 = 亻 + 木
木
亻 + 日 = ?
corpus (word2vec) character embeddings
library learned decompositions
learn library function embeddings
predict hold-out set character meanings

Figure 1: **(A)** Library learning represents logographic forms as programs and discovers recurring structural patterns in characters by compressing and rewriting programs. **(B)** Library learning over written and sound forms discovers rules for mapping graphical parts to sounds. **(C)** Our model reveals systematicity between forms and meanings by predicting character's meaning vector from its library function embeddings.

Inspired by recent work on leveraging library learning as an efficiency-based (Gibson et al., 2019) computational model to study the structure and evolution of written forms in Chinese (Jiang et al., 2024), we develop our computational framework based on a library learning approach. The library learning line of work is best described in inductive program synthesis contexts (Ellis et al., 2021; Bowers et al., 2023; Lake, Salakhutdinov, & Tenenbaum, 2015); it compresses program representations by iteratively growing libraries of program abstractions from the program corpora. Grounded in Chinese characters, we represent characters in program-like representations based on stroke sequences and phonetic alphabets. The


Base DSL primitives
T N H S P D WG XG TN
SG SP HG SW HP PD HZ SZ ST
PZ SWG HXWG BXG HZZ HZT SZZ HZW HZG
SZWG HZWG HZZZ HPWG HZZZG HZZP
written DSL logographic strokes (Jiang et al., 2024)
x /ɕ/ f /f/ y /j/ g /k/ k /kʰ/ l /l/ m /m/ n /n/ b /p/ r /ʐ/
p /pʰ/ s /s/ d /t/ t /tʰ/ j /tɕ/ q /tɕʰ/ z /ts/ c /tsʰ/ w /w/
h /x/ v /ɥ/ a /a/, /ɑ/, /ɛ/ i /i/, /ɿ/, /ʅ/ o /o/ u /u/
e /e/, /ə/, /ɤ/
phonetic DSL pinyin alphabets
#0, #1 start of sequence
Joint written-sound programs
Learned library functions
function name
function body
Programs rewritten with the joint library
娃 (wá)
娃
joint_810
圭 → wa
fn_54
女
fn_191
圭
joint_326
wa
fn_14
PD + P+H
x2
fn_5
土
H+S+H
w + a
rewritten program
visualization of the hierarchical library functions used in 娃 (wá, baby)

```
娃, wá: (bind ($0 PD P H H S H H S H)
              ($1 w a))
蛙, wā: (bind ($0 S HZ H S T D H S H H S H)
              ($1 w a))
汝, rǔ: (map ($0 D D T PD P H)($1 r u))
榜, bǎng: (map ($0 H S P D D H D P D HG D H
              HZG SP)($1 b a n g))
旁, páng: (map ($0 D H D P D HG D H HZG SP)
              ($1 p a n g))
......
```

```
fn_54(#0)          :=  (fn_14 (#0 PD)),
fn_191(#0)         :=  (fn_5 (fn_5 #0)),
                   ......
joint_810(#0,#1)   :=  (joint_326 #1
                             (fn_191 #0)),
                   ......
```

```
娃, wá: (joint_810
        (fn_54 $0) $1)
```

Figure 2: **Overview of our joint written-sound form joint compression model. (Left)**: Base DSL (Domain Specific Language) primitives containing logographic stroke primitives and phonetic alphabet (pinyin) symbols. **(Middle)**: Characters' written-sound correspondences are modeled by `bind` functions that take in written and sound sequences. Library learning on written-sound programs yield high-level abstractions describing sub-grapheme to sound mappings, resembling phonemic rules in Chinese writing systems. **(Right)**: Visualization of a character's written form and sound form's joint hierarchical decompositions.

library learning model identifies recurring graphical patterns and written form-sound bindings and stores them in a library of abstractions. Using these abstractions, characters can be rewritten in a compressed, hierarchical form, reflecting an efficient organization of linguistic forms and aligning with empirical theories on the hierarchical organization of Chinese character orthography (Jiang et al., 2024).

We augment the written-only library learning model to jointly model the meaning and sound, allowing us to examine and discover the structural nature between Chinese characters and their sounds and meanings. Our investigation is organized into two parts. In Part I, we develop a (written) form-sound joint compression model based on the library learning framework. By identifying sub-graphemic elements predictive of pronunciation—analogous to phonetic radicals in Chinese orthography, the model uncovers systematic phonetic cues embedded in logographic characters. We validate these learned form-sound relationships by demonstrating that the model can successfully predict phonetic properties of character parts in phono-semantic compounds.

In Part II, we extend our analysis to (written) form-meaning systematicity at the sub-character level. We represent character meanings as distributional semantic vectors derived from Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Chen, Corrado, & Dean, 2013) and model meaning compositionality using additive transformations of learned library embeddings. Analyzing over ten thousand Chinese characters, our results reveal a systematic relationship between graphical forms and meanings, capable of predicting held-out character's meanings from its library learned written form decomposition. In addition, our meaning model enables us to investigate diachronic changes in form-meaning systematicity, comparing the transition from Classical Chinese (Peyraube, 2016) to modern Vernacular Chinese. Our results suggest that Chinese characters exhibit a greater form-meaning systematicity when encoding Classical Chinese compared to when representing modern spoken Mandarin.

Overall, our findings demonstrate how a unified computational framework based on library learning can capture the compositional nature and structural relationship between written form, meaning, and sounds in the Chinese language. More broadly, leveraging Chinese writing as a unique testbed, this framework may offer new insights into how efficiency shapes a writing system's role in representing the structure of human language and help us reverse-engineer how writing systems reflect human intuition about language structure, driven by efficiency principles.

Part I: Sound

While Chinese is traditionally classified as a logographic system, phonetic cues remain deeply embedded within its characters, particularly in phono-semantic compounds. However, the extent to which written forms encode sound at a sub-character level remains debated. In this section, we develop a computational framework that jointly compresses written and sound representations, revealing underlying phonetic regularities. By leveraging this library learning approach, we identify systematic mappings from graphical components to sounds that align with empirical theories on phono-semantic compounds.

Methods

Preliminary: Library learning on logographic forms Our library learning model is built on top of Jiang et al. (2024)'s model on Chinese logographic forms. In a few words, the library learning model represents individual logographic characters in a writing system $\mathcal{W}$ as program sequences p_{char} of stroke primitives $\{$`T, N, ..., HZZP`$\} \subset \mathcal{L}_{\text{base}}$. The library learning model discovers recurring patterns `fn` in character programs, adds recurring pattern programs to a library $\mathcal{L} = \mathcal{L}_{\text{base}} \cup \{$`fn`$\}$, and rewrites character programs in the most efficient and compressive way (efficiency measured in minimum description length $C(\mathcal{W})$). We define the objectives used to discover the recurring patterns in eq. (1).

$$
\begin{aligned}
\mathrm{DL}_{\mathcal{L}}(\mathcal{W}) &= \overbrace{\sum_{p \in P_{\mathcal{L}_{\text{base}}}(\mathcal{W})} \mathrm{DL}\left(\textsc{Rewrite}(p, \mathcal{L})\right)}^{\substack{\text{description length} \\ \text{of the rewritten characters}}} + \overbrace{\mathrm{DL}\left(\mathcal{L}\right)}^{\substack{\text{description length} \\ \text{of the library}}} \\
\mathrm{DL}\left(\mathcal{L}\right) &= \sum_{\texttt{fn} \in \mathcal{L}} \mathrm{DL}\left(\textsc{Body}(\texttt{fn})\right) \\
C(\mathcal{W}) &= \min_{\mathcal{L}} \mathrm{DL}_{\mathcal{L}}(\mathcal{W})
\end{aligned} \tag{1}
$$

We refer the reader to Part I of Jiang et al. (2024) for further details of the library learning model. To help build intuition about the computational framework, we will use an example of three characters {侗, 洞, 汹} throughout our paper. The initial programs under the base DSL $\mathcal{L}_{\text{base}}$ are defined as follows:

$$
\begin{aligned}
p_{\text{侗}} &:= \texttt{(\#0 P S S HZG H S HZ H)}, \\
p_{\text{洞}} &:= \texttt{(\#0 D D T S HZG H S HZ H)}, \\
p_{\text{汹}} &:= \texttt{(\#0 D D T P N SZ S)}.
\end{aligned} \tag{2}
$$

We represent characters as sequences of strokes, encoded in DSL (domain specific language) primitive symbols shown in fig. 2 left. From the base stroke programs (2), the model identifies recurring stroke patterns 同 $\texttt{fn\_0(\#0)} := \texttt{(\#0 S HZG H S HZ H)}$ and 氵 $\texttt{fn\_1(\#0)} := \texttt{(\#0 D D T)}$. By adding these pattern programs iteratively to the library $\mathcal{L} = \mathcal{L}_{\text{base}} \cup \{\texttt{fn\_0}, \texttt{fn\_1}\}$, the algorithm rewrites the base programs with reference to the patterns learned in $\mathcal{L}$:

$$
\begin{aligned}
\textsc{Rewrite}(p_{\text{侗}}, \mathcal{L}) &= \texttt{(fn\_0 (\#0 P S))}, \\
\textsc{Rewrite}(p_{\text{洞}}, \mathcal{L}) &= \texttt{(fn\_0 (fn\_1 \#0))}, \\
\textsc{Rewrite}(p_{\text{汹}}, \mathcal{L}) &= \texttt{(fn\_1 \#0 P N SZ S)}.
\end{aligned} \tag{3}
$$

In practice, when scaling our tiny example to thousands of Chinese characters, the library learning framework would also discover radicals 亻 ($\texttt{fn\_2}$) and 凶 ($\texttt{fn\_3}$), yielding new rewrites $\textsc{Rewrite}(p_{\text{侗}}, \mathcal{L}_{\text{written}}) = \texttt{(fn\_0 (fn\_2 \#0))}$ and $\textsc{Rewrite}(p_{\text{汹}}, \mathcal{L}_{\text{written}}) = \texttt{(fn\_3 (fn\_1 \#0))}$. By applying library learning to a large set of Chinese characters, the learned library functions and the rewritten character programs are able to capture the hierarchical written form structure in Chinese characters (Jiang et al., 2024; Myers, 2019).

Joint compression of written and sound forms Jiang et al. (2024) demonstrated that a library learning model can faithfully capture the hierarchical structure within Chinese characters. Here, we extend the written-form-only model to jointly model written and sound forms. For sound form representations, we utilize Pinyin (also known as the Chinese phonetic alphabet, see L. L. Chen (2016) for an introduction) to encode sounds. In our {侗, 洞, 汹} example, pronunciations are 侗 (dòng), 洞 (dòng), 汹 (xiōng). Similar to how we encode stroke sequences, we use the same Lisp-like grammar to encode phonetic alphabet primitive sequences (e.g., $p_{\text{dòng}} := \texttt{(\#0 t o n g)}$, note that we omit tones for simplicity, see fig. 2 for more examples). To characterize written and sound form mappings, we use an additional DSL function $\texttt{(bind \#written\_form \#sound\_form)}$ that intuitively links written and sound representations in a single joint program. We represent the joint programs for {侗, 洞, 汹} as follows (note we already rewrote part of the joint program with library functions learned on written forms for efficient compression):

$$
\begin{aligned}
p_{\text{侗, dòng}} &:= \texttt{(bind (fn\_0 (fn\_2 \#0)) (\#1 d o n g))}, \\
p_{\text{洞, dòng}} &:= \texttt{(bind (fn\_0 (fn\_1 \#0)) (\#1 d o n g))}, \\
p_{\text{汹, xiōng}} &:= \texttt{(bind (fn\_3 (fn\_1 \#0)) (\#1 x i o n g))}.
\end{aligned} \tag{4}
$$

We follow the compression objectives defined in eq. (1), but now instead of compressing only logographic forms (e.g., programs (2)), we compress joint programs (e.g., programs (4)) to find recurring patterns between logographic written forms and sounds. From this example, we can get library functions $\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{written}} \cup \{\texttt{joint\_0}, \texttt{joint\_1}\}$ and rewritten programs (5). New library functions learned from compressing joint programs resemble phonetic abstractions and rules for mapping graphical parts to phonetic abstractions. For example, $\texttt{joint\_0(\#0,\#1)} := \texttt{(bind \#1 (\#0 o n g))}$ resembles the Pinyin coda (final) *ong* /ʊŋ/, $\texttt{joint\_1(\#0)} := \texttt{(joint\_0 d (fn\_0 \#0))}$ represents the rule of mapping radical 同 → *dong*.

$$
\begin{aligned}
\textsc{Rewrite}(p_{\text{侗, dòng}}, \mathcal{L}_{\text{joint}}) &= \texttt{(joint\_1 (fn\_2 \#0))} \\
\textsc{Rewrite}(p_{\text{洞, dòng}}, \mathcal{L}_{\text{joint}}) &= \texttt{(joint\_1 (fn\_1 \#0))} \\
\textsc{Rewrite}(p_{\text{汹, xiōng}}, \mathcal{L}_{\text{joint}}) &= \texttt{(joint\_0 (x i) (fn\_3 (fn\_1 \#0)))}
\end{aligned} \tag{5}
$$

Another challenge for modeling the written-sound relationship is the presence of polyphonic characters in Chinese. In other words, some Chinese characters have more than one pronunciation. During our library learning process, we optimize the most compressive written-sound mappings. For example, for the Chinese character 都, which has two pronunciations *dū* and *dōu*, we model the description length of the rewritten character as $\min_{\text{sound} \in \{\text{dū}, \text{dōu}\}} \mathrm{DL}\left(\textsc{Rewrite}(p_{\text{都},\text{sound}}, \mathcal{L})\right)$.

Results

Joint compression reveals mappings between written and sound forms at a sub-character level We applied our joint compression model on 12,085 Chinese characters together with their pronunciations (16,316 joint programs considering polyphones). The 12,805 characters cover most of the set of characters used in describing the Chinese Wikipedia (12,805 out of 14,710, Wikipedia Contributors (2025)), excluded ones are due to the lack of ground-truth raw stroke programs.

Our library learning model discovered 1,825 library functions representing higher-order abstractions in written forms; these written-form abstractions resemble radicals discovered by Jiang et al. (2024). Additionally, our joint compression of written and sound forms yielded 1,923 sound-related abstractions.

By detailed examination of the 1,923 library functions learned, we identified highly interpretable semantics from the learned abstractions. In particular, 376 library functions are coda-like abstractions from phonetic alphabets

Table 1: **Examples of learned library functions from joint compression**, ordered according to their frequency of usage (and their percentile rank). Beta normal forms (in lambda calculus expressions, see Alama and Korbmacher (2024) for a brief introduction) can be interpreted as a flattened version of hierarchical library functions.

Library function discovered (beta normal form)	#Uses	(percentile)	Semantic	Example usage

`joint_0(#0,#1) := (bind #1 (#0 n g))`
`(λ (bind $0 ($1 n g)))`
	2585	99.90%	ng (coda)	唔 (ńg) 嗯 (ńg) 哼 (hēng) 哽 (gěng)

`joint_13(#0) := (joint_0 (#0 i))`
`(λ (bind $0 ($1 i n g)))`
	511	99.10%	ing (coda)	`<intermediate abstraction>`

`joint_22(#0) := (joint_5 (#0 u))`
`(λ (bind $0 ($1 u n)))`
	460	98.80%	un (coda)	轮 (lún) 遁 (dùn) 蹲 (dūn) 囷 (qūn)

`joint_63(#0) := (joint_13 (#0 l))`
`(λ (bind $0 ($1 l i n g)))`
	73	95.60%	ling (syllable)	另 (lìng) 磷 (lín) 釘 (dīng) 翎 (líng)

`joint_64(#0) := (joint_22 (#0 y))`
`(λ (bind $0 ($1 y u n)))`
	79	96.40%	yun (syllable)	運 (yùn) 运 (yùn) 尉 (yùn) 盾 (yǔn)

`joint_188(#0,#1) := (joint_63 #1 (fn_213 #0))`
`(λ (bind (fn_213 $1) ($0 l i n g)))`
	29	87.80%	令→ling	令 (lìng) 冷 (lǐng) 零 (líng) 岭 (lǐng)

`joint_245(#0,#1) := (joint_28 #1 (fn_136 #0))`
`(λ (bind (fn_136 $1) ($0 y u)))`
	20	85.10%	俞→yu	瑜 (yú) 喻 (yù) 俞 (yú) 愉 (yú)

`joint_810(#0,#1) := (joint_326 #1 (fn_191 #0))`
`(λ (bind (fn_191 $1) ($0 w a)))`
	9	70.80%	圭→wa	娃 (wá) 洼 (wā) 蛙 (wā) 哇 (wā)

`joint_546(#0,#1) := (joint_104 #1 (fn_718 #0))`
`(λ (bind (fn_718 $1) ($0 l u o)))`
	8	68.50%	累→luo	螺 (luó) 漯 (luò) 骡 (luó) 騾 (luó)

`joint_892(#0,#1) := (joint_412 #1 (fn_600 #0))`
`(λ (bind (fn_600 $1) ($0 s h u o)))`
	5	49.60%	朔→shuo	溯 (shuò) 朔 (shuò) 搠 (shuò) 鎙 (shuò)

`joint_1494(#0,#1) := (joint_235 #1 (fn_1233 #0))`
`(λ (bind (fn_1233 $1) ($0 s h a)))`
	4	30.40%	妾→sha	接 (shà) 霎 (shà) 翣 (shà) 唼 (shà)

`joint_1447(#0,#1) := (joint_64 #1 (fn_579 #0))`
`(λ (bind (fn_579 $1) ($0 y u n)))`
	3	23.10%	軍→yun	暈 (yūn) 惲 (yùn) 韗 (yùn)

`joint_1828(#0,#1) := (joint_187 #1 (fn_21 (fn_780 #0)))`
`(λ (bind (fn_21 (fn_780 $1)) ($0 z h e n g)))`
	2	2.10%	灬+丞→zheng	蒸 (zhēng) 烝 (zhēng)

or syllables in Chinese sounds. We selected and interpreted a subset of sound-related library functions and visualized them in Table 1 for reference. In the example shown, `joint_0(#0, #1)` := `(bind #1 (#0 n g))` and `joint_13(#0)` := `(joint_0 (#0 i))` are two of the most frequently used library functions discovered, resembling the Chinese coda *ng* /ŋ/ and *ing* /iŋ/. Moreover, phonetic library functions are also hierarchically structured, as syllables are composed of simpler initials and codas (e.g., `joint_63` *ling* is composed of `joint_13` *ing* + `l` *l*).

The remaining 1,547 library functions depict sub-grapheme to library functions depicting sub-grapheme to phonemes mappings at multiple levels. In Table 1, we notice that library functions describe interpretable logographic parts to sound mappings. The logographic parts involved in the library functions intuitively correspond to phonetic radicals in Chinese characters (DeFrancis, 1986). For example, in 娃 (*wá*, baby), 洼 (*wā*, swamp), 蛙 (*wā*, frog), and 哇 (*wā*, wow), they all have the same 圭 part that can be attributed to the source of their sounds – *wa*. The 圭 part is thus identified as a phonetic radical that contributes to the pronunciation in characters. Of note, a phonetic radical does not always contribute to characters' sounds in Chinese: 圭 can appear in 佳 (*jiā*, good), 卦 (*guà*, divinatory diagram) and itself as a standalone character 圭 (*guī*, sceptre), they all differ from the sound *wa* as expected.

Library learning discovers phono-semantic compounds In the previous section, we intuitively found sub-grapheme to sound patterns in our library-learned functions. Inspired by the widely developed theory of phono-semantic compounds in Chinese characters (Hsiao & Shillcock, 2006; Myers, 2019), we further quantitatively evaluate the validity of our library learning model in identifying phonetic parts in phono-semantic compounds. As we demonstrated in the 圭 example, phono-semantic compounds are highly irregular and inconsistent (Zhou, 1978), it remains challenging to attribute phonetic parts.

We collected phonetic regularity data of Chinese characters from the Hanzipy library (Synkied, 2023). We retained characters that can be decomposed into two graphical parts both by the Hanzipy library and our model, resulting in 3,598 phono-semantic compounds as our ground truth to compare.

To identify the phonetic part in graphical forms, we leverage our joint compression model by rewriting written-sound programs with learned library functions. Based on the rewritten outcome, we further analyzed the library functions used in rewriting the joint programs: if a written-form abstraction appears in one of the used sound-related library functions, then the graphical part corresponding to the written-form abstraction is likely contributing to the character's pronunciation, or in other words, the part is a phonetic radical. To help understand this, we take the 娃 (*wá*, baby) character and its sound as an example.

Our library learning model encodes 娃 (*wá*) as $p_{\text{娃, wá}}$:= `(bind (#0 PD P H H S H H S H) (#1 w a))`. We write $p_{\text{娃, wá}}$ with $\mathcal{L}_{\text{joint}}$ yielding $\textsc{Rewrite}(p_{\text{娃, wá}}, \mathcal{L}_{\text{joint}}) =$ `(joint_810 (fn_54 #0) #1)`, where `fn_54` is the left part of 娃–女, `fn_191` is the right part 圭, `joint_326` is the syllable abstraction *wa*.

We identify the key library function used for the mapping from written to sound is `joint_810(#0,#1) := (joint_326 #1 (fn_191 #0))`. It binds form 圭 and sound *wa*. The other part of 娃–女 is not involved in the mapping function. Hence, our model acknowledges 圭 as the phonetic part and 女 as the semantic part in the phono-semantic compound 娃.

By scaling up our analysis on 3,598 phono-semantic compounds, our library learning model correctly predicted phonetic and semantic attributions with a well-above-chance accuracy of 72.9% (2,624/3,598). These findings suggest our model can faithfully capture the highly irregular and inconsistent patterns between logographic parts and sound forms.

Part II: Meaning

Our results from Part I demonstrate that a library learning model can effectively capture both the combinatorial structure of written forms and the systematic mappings between written and spoken forms. In this section, we aim to further advance the model's capabilities for structure discovery by examining the relationships between the learned library functions and their associated meanings. We leverage distributional semantic meaning representations and additive compositionality of vector meanings to demonstrate that characters' meanings can be effectively predicted from their constituent library functions. By testing the predictability of characters, our model assesses the form-meaning systematicity in Chinese characters and reveals historical changes in the relationship between written and spoken language.

Methods

Modeling meanings of characters and library functions Following common practices in modeling meaning systematicity (Pimentel, McCarthy, Blasi, Roark, & Cotterell, 2019; Gutiérrez, Levy, & Bergen, 2016; Piantadosi et al., 2024), we represent and extract meanings by taking vector representations of Chinese characters with a distributional semantic model, Word2Vec's CBOW (Continuous Bag-of-Words) (Mikolov, Chen, et al., 2013). For the implementation of CBOW, we use fastText (Joulin, Grave, Bojanowski, & Mikolov, 2017) to train our character embedding models.

We define a Chinese character c 's distributional semantic vector representation as $\boldsymbol{v}^{(c)} \in \mathbb{R}^d$. In practice, we trained a CBOW model with $d = 300$ on the traditional Chinese Wikipedia corpus (Wikipedia Contributors, 2025) tokenized at the character level. The Wikipedia corpus contains 863M Chinese character tokens (14,710 unique characters) for training.

For modeling library function's meaning, we draw inspiration from Mikolov, Sutskever, et al. (2013); Bonandrini et al. (2023); Marelli and Baroni (2015) on distributional semantic representation's additive compositionality. Our library function vector representations are learned from additive supervisions. We learn library function embeddings $\boldsymbol{u}^{(i)}$ by predicting a character c's embedding $\hat{\boldsymbol{v}}^{(c)}$ from its rewritten program $\textsc{Rewrite}(p_c, \mathcal{L})$'s constituent library functions (eq. (6)).

$$\hat{\boldsymbol{v}}^{(c)} = \frac{\sum_{i \in \textsc{Rewrite}(p_c, \mathcal{L})} \boldsymbol{u}^{(i)}}{\|\sum_{i \in \textsc{Rewrite}(p_c, \mathcal{L})} \boldsymbol{u}^{(i)}\|_2} \tag{6}$$

We train library embedding vectors $\boldsymbol{u}^{(i)}$ by minimizing the MSE loss between predicted character vectors $\hat{\boldsymbol{v}}^{(c)}$ and CBOW-learned character vectors $\boldsymbol{v}^{(c)}$: $L_{\text{MSE}} = \frac{1}{n} \sum_{c \in \mathcal{W}_{\text{train}}} (\boldsymbol{v}^{(c)} - \hat{\boldsymbol{v}}^{(c)})^2$ using gradient descent. We randomly select 80% of all available character programs $\mathcal{W}$ (as specified in Part I) as the training set $\mathcal{W}_{\text{train}}$ and 20% as the test set $\mathcal{W}_{\text{test}}$.

Measuring form-meaning systematicity We measure systematicity as the degree to which we can predict the meaning of a Chinese character from the decompositions of its written form. Based on our distributional semantic modeling method, the systematicity of the form-meaning corresponds to the performance of predicting the meaning of the character from the meanings of its constituent library functions in its form.

We report the predictability of character meaning $\boldsymbol{v}^{(c)}$ from its form decompositions in ranking terms. Specifically, we calculate the cosine similarity between the predicted meaning vector and the ground-truth: $\cos(\hat{\boldsymbol{v}}^{(c)}, \boldsymbol{v}^{(c)}) = \frac{\hat{\boldsymbol{v}}^{(c)} \cdot \boldsymbol{v}^{(c)}}{\|\hat{\boldsymbol{v}}^{(c)}\| \|\boldsymbol{v}^{(c)}\|}$. We report the average cosine similarity in the test set $\frac{1}{|\mathcal{W}_{\text{test}}|} \sum_{c \in \mathcal{W}_{\text{test}}} \cos(\hat{\boldsymbol{v}}^{(c)}, \boldsymbol{v}^{(c)})$, and the relative hit rate $P@1$ (exact match), $P@10$, $P@100$, we define $P@k$ as follows:

$$P@k = \frac{1}{|\mathcal{W}_{\text{test}}|} \sum_{c \in \mathcal{W}_{\text{test}}} \mathbb{I}\left(\text{Rank}\left(\cos(\hat{\boldsymbol{v}}^{(c)}, \boldsymbol{v}^{(c)}), \left\{ \cos(\hat{\boldsymbol{v}}^{(c)}, \boldsymbol{v}^{(i)}) \middle| \, i \in \mathcal{W}_{\text{test}} \right\} \right) \leq k \right),$$

where $\mathbb{I}$ is the indicator function, $\text{Rank}(\cdot, \cdot)$ calculates a element's ranking index in a list (in descending order).

Results

Predictable relations between character and sub-character meanings We evaluated the characters' meaning predictability with a Wikipedia-trained Word2Vec model as the source for ground-truth meaning vectors. We report five-fold cross-validation (9,668 characters for training, 2,417 characters for testing) results in Table 2.

Table 2: **Quantitative evaluation of Chinese characters' (written) form-meaning systematicity**, based on the meaning predictability. We report the average cosine similarity between the predicted meaning vector and the ground-truth character's meaning vector. Relative hit rates $P@k$ represent how much percent of the ground truth character falls into the k-nearest neighbor of the predicted vector.

Model	Similarity	$P@1$	$P@10$	$P@100$
Library learning	0.169	3.81%	13.33%	36.31%
Random decomposition	0.022	0.02%	0.45%	4.27%

Our results show that library-learned decompositions of Chinese characters lead to highly predictable meaning compositions compared to a random decomposition baseline. Table 3 demonstrates three successfully predicted characters and three failed cases. We found that highly predictable characters generally have one semantically transparent radical (e.g. 魚 /鱼 fish, 钅 iron, 疒 sickness, 鳥 bird).

Phonetic radicals contribute less to character meanings To quantitatively test how semantic radicals affect meaning systematicity, we performed a simple ablation by removing phonetic radicals. In detail, we identified written-form library functions that associate with sounds (from Part I), and substituted all phonetic-related library functions with a placeholder function `fn_sound`. We re-trained the constituent library function embeddings and evaluated how removing these library functions impacts meaning systematicity.

Table 3: **Examples of predicting characters' meaning from their learned library constituents.** We show three successfully predicted characters and three failure cases, sorted by the cosine similarity between the additive predicted vector and the ground truth character's meaning vector. Predicted character shown is the nearest neighbor of the predicted meaning vector. Identifiable library functions' graphical illustrations and their approximate english meanings are also marked for reference.

Constituents	Predicted	Ground-truth	Similarity
fn_111 (鱼, fish), fn_837 (沙, sand)	鲨, shark	鲨, shark	0.757
fn_728 (矣, done), fn_0 (口, mouth)	唉, sigh	唉, sigh	0.593
fn_519 (也, also), fn_54 (女, female)	她, her	她, her	0.530
fn_13 (旦, dawn) , fn_239 (少, few)	眸, eye	省, save	-0.239
fn_82 (立, stand), fn_20 (扌, hand)	扔, throw	拉, pull	-0.161
fn_5 (土, earth), fn_1443 (刑, penalty)	肇, cause	型, type	-0.145

A total of 225 phonetic-related logographic library functions were removed. For comparison, we also performed a baseline study by randomly removing the same number of library functions (table 4). We found phonetic library functions (phonetic radicals) contribute less to character's meaning systematicity than the average. This result further validates the library learning model's ability in capturing written-sound and form-meaning relationships in a unified view.

Table 4: **Phonetic library function's contribution to meaning.** We compare between removing phonetic library functions and random removal on impairing the predictability of character's meaning. Phonetic library functions has minimal impact on meaning predictions.

Model	Similarity	*P*@1	*P*@10	*P*@100
Full	0.169	3.81%	13.33%	36.31%
Phonetic removal	0.171	3.43%	12.71%	35.81%
Random removal	0.165	3.37%	11.99%	28.73%

Classic Chinese character's meaning is more predictable than Vernacular (Modern) Chinese In the history, there was an increasing gap between written and spoken Chinese during the imperial China eras. In simple words, the spoken Chinese language evolved much faster than the written Chinese. The old written Chinese (also called Literary Chinese) was used and designed to represent classical Chinese instead of the spoken forms that people use everyday (Peyraube, 2016). However, this huge gap was not bridged until the New Culture Movement (or, more specifically, the Written Vernacular Chinese Reform (Chow, 1960)) in the early twentieth century. During the Written Vernacular Chinese Reform, scholars tried to modify the Literary Chinese scripts to make it close to Beijing Mandarin by modifying the sounds and grammar in the writing, while keeping most written character forms of Literary Chinese unchanged (P. Chen, 1999; Wei, 2014).

Given this historical sociolinguistic background, Chinese characters were designed for the Literary Chinese writing system and represent the meaning of the Classic Chinese language. Inspired by Jiang et al. (2024)'s analysis on the deliberate simplification from tradition to simplified Chinese has broken the consistency of the form structures across characters, we similarly hypothesize that the deliberate Written Vernacular Chinese Reform may break the (written) form-meaning systematicity of the Chinese writing system at a character level.

To validate this hypothesis, we compared the form-meaning systematicity of Chinese characters in two different versions of the Chinese language. We employed the 四庫全書 (*Complete Library of the Four Treasuries*, the largest imperial Chinese encyclopedia (Egan, 2001)) for modeling Classic Chinese meanings and Chinese Wikipedia for modern Chinese.

We trained two Word2Vec models on the two different versions of Chinese. The 四庫全書 used is at the same order of magnitude (690M Chinese character tokens) compared to the Chinese Wikipedia corpus (863M). We used characters that appear in both versions of Chinese, resulting in 10,037 character programs encoded, and 1,625 library functions learned by applying the library learning model on the character programs.

Table 5: **Diachronic comparisons of (written) form-meaning systematicity between classic and Modern Chinese**, measured by using the same set of traditional Chinese characters but different meanings.

Script	Similarity	*P*@1	*P*@10	*P*@100
Classic (Siku Quanshu)	0.2259	9.75%	29.97%	62.70%
Modern (Wikipedia)	0.1423	2.30%	10.67%	32.48%

In a character-meaning predictability analysis (table 5), we observed that Chinese characters exhibit a higher degree of (written) form-meaning systematicity when describing Classic Chinese compared to modern Chinese.

Discussion

In this work, we extend an efficiency-based library learning framework to investigate the structural relationships between writing, sound, and meaning. In the first part of our study, we model logographic characters and their pronunciations as structured program-like representations over stroke primitives and phonetic alphabets. Leveraging joint compression, our framework uncovers sub-graphemic mappings between written and phonetic components and the structure of phono-semantic compounds in Chinese writing. In the second part, we examine the relationship between written forms and meanings within Chinese characters. We formulate form-meaning systematicity as additive compositionality within a semantic embedding space, revealing predictable links between a character's logographic structure and its meaning. A diachronic analysis of Classical and Modern Chinese further provides insights into historical changes in how writing encodes meaning.

Together with our prior work (Jiang et al., 2024), this study shows that a library learning-based computational model can capture the inductive biases underlying the emergence and evolution of compositional structures in human language. It emphasizes representational efficiency as a unifying principle that links combinatorial reuse in logographic forms to systematic mappings among form, sound, and meaning. We hope our work can contribute to theories of efficiency as a design feature for writing systems, and how writing efficiently conveys and represents humans' intuitive understanding of sound and meaning structures in human languages.

Acknowledgments

We thank Maddy Bowers and Hongjie Li for many valuable discussions and contributions. MH, JM, LW, and JBT are supported by MIT Quest for Intelligence, AFOSR Grant #FA9550-19-1-0269, the MIT-IBM Watson AI Lab, ONR Science of AI, and Simons Center for the Social Brain.

References

Alama, J., & Korbmacher, J. (2024). The Lambda Calculus. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2024 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2024/entries/lambda-calculus/.

Bonandrini, R., Amenta, S., Sulpizio, S., Tettamanti, M., Mazzucchelli, A., & Marelli, M. (2023). Form to meaning mapping and the impact of explicit morpheme combination in novel word processing. *Cognitive Psychology*, *145*, 101594.

Bowers, M., Olausson, T. X., Wong, L., Grand, G., Tenenbaum, J. B., Ellis, K., & Solar-Lezama, A. (2023). Top-down synthesis for library learning. *Proceedings of the ACM on Programming Languages*, *7*(POPL), 1182–1213.

Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual review of anthropology*, *16*, 383–407.

Chen, L. L. (2016). Hanyu pinyin. In *The routledge encyclopedia of the chinese language* (pp. 522–542). Routledge.

Chen, P. (1999). *Modern chinese: History and sociolinguistics*. Cambridge University Press.

Chow, T.-t. (1960). *The may fourth movement: Intellectual revolution in modern china*. Harvard University Press.

Coulmas, F. (2003). *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press.

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

Egan, R. (2001). Reflections on uses of the electronic siku quanshu. *Chinese Literature: Essays, Articles, Reviews (CLEAR)*, *23*, 103–113.

Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Morales, L., Hewitt, L., ... Tenenbaum, J. B. (2021). Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation* (pp. 835–850).

Frost, R. (2012). Towards a universal model of reading. *Behavioral and brain sciences*, *35*(5), 263–279.

Gelb, I. J. (1963). *A study of writing*. University of Chicago Press.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389–407.

Gutiérrez, E. D., Levy, R., & Bergen, B. (2016). Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2379–2388).

Hsiao, J. H.-w., & Shillcock, R. (2006). Analysis of a chinese phonetic compound database: Implications for orthographic processing. *Journal of psycholinguistic research*, *35*, 405–426.

Jiang, G., Hofer, M., Mao, J., Wong, L., Tenenbaum, J., & Levy, R. (2024). Finding structure in logographic writing with library learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017, April). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (pp. 427–431). Association for Computational Linguistics.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, *122*(3), 485.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Myers, J. (2019). *The grammar of chinese characters: Productive knowledge of formal patterns in an orthographic system*. Routledge.

Peyraube, A. (2016). Ancient chinese. In *The routledge encyclopedia of the chinese language* (pp. 39–55). Routledge.

Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, *28*(9), 844–856.

Pimentel, T., McCarthy, A. D., Blasi, D., Roark, B., & Cotterell, R. (2019). Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1751–1764).

Sproat, R. (2000). *A computational theory of writing systems*. Cambridge University Press.

Sproat, R. (2010). *Language, technology, and society*. Oxford University Press.

Sproat, R., & Gutkin, A. (2021). The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, *47*(3), 477–528.

Synkied. (2023). *hanzipy*. Retrieved from https://github.com/Synkied/hanzipy (Accessed: 2025-01-27)

Wei, S. (2014). 10 writing and speech: Rethinking the issue of vernaculars in early modern china. In *Rethinking east* *asian languages, vernaculars, and literacies, 1000–1919* (pp. 254–301). Brill.

Wikipedia Contributors. (2025). *Wikipedia: Database download.* Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Database_download (Accessed: 2025-01-28)

Zhou, Y. G. (1978). Xiandai hanzihong shengpangde biaoyin gongneng wenti [to what degree are the "phonetics" of present-day chinese characters still phonetic?]. *Zhongguo Yuwen*, *146*, 172–177.