Recording Usage in an Open Script: Variant Glyphs in Chinese Historical Lexicography

26 June 2026

Ansel Zhang
ansel.zhang@kell.ox.ac.uk

Kellogg College, University of Oxford

The Problem: Recording Usage in an Open Script

?*c*1225 (?*a*1200) Aduersite & prosperte.
Ancrene Riwle (MS Cleopatra C.vi) (1972) 145
[Composed ?*a*1200]

OED Entry of an older spelling of ‘adversity’

ɑ a

Allographs of the Latin ‘a’

Alphabetic Old spelling	Chinese variant glyphs
Finite spelling variants	Open-ended graphic variation
Usually typeable	Often not encoded/typeable
Citation problem	Citation + display + input problem

Small sample of the attested variants of 帝 ‘emperor’

Same Passage, Different Transmitted Graphs

10th Century (Song Dynasty)
manuscript witness
拑 qián

13th Century (Yuan Dynasty)
manuscript witness
箝 qián

For the dictionary, this should appear as documented textual variation, not an unexplained inconsistency.

Interglyph Relationships

Loaned Characters 假借字

An existing glyph borrowed to write another word, usually because the two are similar in pronunciation.

我

Wǒ - Originally referred to a saw-like tool or weapon, but was borrowed to write the first-person pronoun “I/me.

Derivative Characters 分化字

New graph created to distinguish meanings

然 → 燃

rán

然 originally had a meaning connected with burning. Later it was heavily used for grammatical or abstract meanings such as “so, thus.” To clarify the burning meaning, 燃 was created with the fire radical 火.

Interglyph Relationships

Old and New Print Forms 新舊字型

Typographic standardisation; Not necessarily a new graph

眞 → 真

Zhēn ‘Truth’

Simplified and Traditional Forms 簡體/簡化

A graph that is written in a simpler form than another related graph.

體 → 体

Tǐ ‘Body’

These have pre-modern customary usages, but also result from twentieth-century standardisation processes.

Chu (2023): Chinese Characters Repertoire Project

Character Repertoires	**Mostly Complete**: Target c. 300,000 characters
Fonts	**In Progress**: *Zhonghua Shuju Songti* is a small experimental output; larger final fonts still pending.
Databases	Future work likely maps variants onto Ideographic Variation Sequences (IVSes)
Input Methods	**Partial**: Offers input support, but not yet enough for full lexicographic use
Source Evidence	**Strong**: Focuses on sourcing glyphs from texts, dictionaries, rime books, and printed editions.
Philological Analysis	**Strong:** Teams analyse form, sound, meaning, usage, and inter-glyph relationships (字際關係).

Tung (2025): *Ad hoc* glyphs

Character Repertoires	**Bypassed**: Large repertoires do not automatically enter production workflows.
Fonts	**Insufficient**: Fonts exist, but difficult books still require *ad hoc* glyphs.
Databases	**Broken link:** Image-glyphs / throwaway glyphs are not searchable or reusable data.
Input Methods	? Encoded or drawn forms may still be impossible to type in ordinary workflows.
Source Evidence	**Detached**: *Ad hoc* glyphs often lose connection to edition-level evidence.
Philological Analysis	**Absent:** Typesetters solve page layout, not variant relations or historical identity.

Visible on the Page, Invisible to Search

先看表示的层面，即咏蝉的层面。首句，“垂  ”二字写蝉的形象，是拟人法。“  ”是什么呢？是古代绅士结在颔下的帽带下垂部分，又叫冠缨。一说：“蝉首有触须，如人之冠缨。“（刘永济）读者多信而不疑。然而端详蝉的标本，便觉其说不妥——蝉的触须在头顶，而且是短短的两根，像角，也像眉，怎样也不像冠缨。一说：“蝉喙长在口下，似冠之  也。”（孔颖达）按，蝉喙细长如带，部位又在颔下，所以说法成立。接着，“饮清露”三字写蝉的习性。古人不知道蝉吸食树汁以存活，以为它餐风饮露。诗非科学，无妨出以想象。次句，始说蝉声“流响出疏桐”。

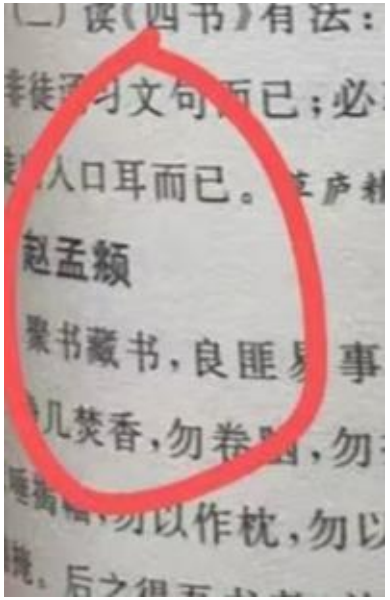
Variant characters are inserted as images (in white) in texts.
These cannot be searched in databases nor contextualised with their source texts

Unicode is Not Enough

Unicode Character "癵" (U+2B5AF)

Unicode 字符 "癵" (U+2B5AF)

癵



The unicode point for a character can exist … but publishers may still need to rely on *ad hoc* glyphs.

Wang (2015): Componential Analysis

Character Repertoires	**Searchable**: Large repertoires become usable when decomposed into components.
Fonts	**Expanding**: FSung has large PUA coverage and keeps growing through discovered variants.
Databases	**Core Mechanism**: Decomposition tables turn glyphs into structured searchable data.
Input Methods	**Completed**: The Componential search algorithm lets users find characters without knowing sound or radical.
Source Evidence	**Needs linking:** Retrieved glyphs must still be checked against attested sources.
Philological Analysis	**Enabled**: Component analysis helps distinguish structure, variants, and misidentified forms.

Componential Analysis Algorithm 部件檢索

A variant of 真 with two internal strokes rather than the standard three, from a Jesuit manuscript c. 1584

It is displayed on the algorithm once its components are entered into the search bar

Search by structure, not by pronunciation

Zhonghua Shuju Sonti vs. *FSung*

	Zhonghua Shuju Songti 中華書局宋體	**FSung 全宋體**
Project type	institutional / repertoire-first	tool-driven / workflow-first
Total glyph coverage	130k+ reported in current font package	near-200k searchable environment
PUA / self-made glyphs	c. 30,000	c. 100,000
Input/retrieval	Guolian input tools	Component Search Algorithm(部件檢索)
Interoperability	PUA not compatible with FSung	PUA not compatible with Zhonghua
Main weakness	not fully integrated as lexicographic workflow	source evidence/provenance must be checked

Conclusion

Best Case	Worst Case
Revisers of the *Hanyu da zidian* and developers of digital repertoires collaborate	Projects remain disconnected
Usable font + database + input tools	Print-first or proprietary system
variants searchable by head character, variant, or component	Variants visible but not reusable
Citable historical orthography	Orthography trapped on the page

A dictionary that records variant glyphs must also make them usable