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

Ansel Zhang, Kellogg College, University of Oxford

anselzhang2021@gmail.com

I: Challenges of Historical Chinese Orthography

- While alphabetic scripts have a relatively small inventory of letters and relatively limited numbers of allographs (e.g. ligatures; 'ɑ' and 'a'), Chinese characters constitute an open-ended historical repertoire. The base character dì 帝 has over 1,000 allograph in the Western Zhou bronze scripts.
- Lexicographers have tended to select one form as 'standard,' recording other forms as 'variants.' In the digital age, this has often meant that only the 'standard' gets encoded into Unicode.
    - This is fine for everyday purposes, but philological and historical studies require these variants to be faithfully preserved and transcribed.

II: 3rd Edition of the *Hanyu da zidian* (Great Dictionary of the Chinese Language)

- Considered the most scholarly dictionary of the Chinese language.
- Main Fault of the first (1990) and second (2010) editions of this dictionary → They do not signpost the provenances of the variants recorded, leading users to mistake different textual transmissions as editorial errors.
- The third edition of the *Hanyu da zidian,* scheduled for publication in 2031, should identify a base text and mark variant quotations as documented textual variation, not unexplained inconsistency.[1]

III: Chinese Characters Repertoire Project

- Variants of a character interrelate in different ways (loans, derivatives, traditional and simplified) – Unless these variants are connected to an information database about its form, pronunciation, meaning, and sources.
- The project's small, experimental font *Zhonghua shuju songti* 中華書局宋體 works on the principle that the Unicode points are not sufficient for historical Chinese orthography.
    - Instead, it aims to create a co-ordinated system linking character repertoires, fonts, databases, input methods, source evidence, as well as human philological analysis.[2]

---

[1] 王祝英、楊麗 Wang Zhuying, and Yang Li. '《漢語大字典》修訂雜談' [Discussions on the Revision of Hanyu Da Zidian.] 中國語言學研究 *Journal of Studies on Languages in China*, no. 4 (2024): 229-243

[2] 朱翠萍 Zhu Cuiping in '系列筆談之六：古籍數字化與漢字編碼、字符集' [Digitisation of Ancient Texts and Chinese Character Encoding and Character Sets.] 數字人文 Digital Humanities, no. 2 (2023): 128-133

IV: The Persistence of *Ad hoc* Glyphs

- Despite the *Zhonghua shuju songti,* many publishers outsource typesetting to external companies. Instead of real characters, these companies often create visually acceptable substitutes that, often into form of an inserted image.
- Ancient-text typesetting is labour-intensive, poorly standardised, and commercially marginal. Since each book may require its own set of rare glyphs, it is often cheaper and faster to create throwaway glyphs than to incorporate every form into a durable repertoire with full retrieval support. The implication is that the crisis of missing characters is not only a Unicode problem, nor only a font-design problem. It is also a workflow problem.
- A glyph created for one book or one layout may not be reusable elsewhere.
    - It may not be searchable, survive copying and pasting, or be associated with any stable code point or variant relationship

V: Componential Analysis

- Component-based search offers one possible solution: instead of relying on pronunciation or stroke count, it allows the user to identify characters through their internal structure, which is especially valuable for rare or partially unreadable graphs.
- FSung already demonstrates how powerful this can be, since it combines a very large repertoire with component-based retrieval, but its coverage of premodern variants is still incomplete and awaiting further manual analysis.
- In practice, this means breaking rare forms down into their constituent parts, checking whether the form already exists in the system, and recording missing forms so that the repertoire can continue to expand.
- The comparison between *Zhonghua Shuju Songti* and FSung shows that two font projects can address the same problem from opposite ends: Zhonghua Shuju Songti is repertoire-first and provenance-heavy, while FSung is workflow-first and retrieval-heavy.
    - The two fonts are not interchangeable, and because they do not assign the same glyphs to the same PUA code points, the same file may display differently depending on which font is installed.
    - Users needing large numbers of variant ideographs will gravitate toward the ecosystem that combines the broadest repertoire with the most usable input and retrieval tools.

VI: Conclusion

- Implication for the third edition of the *Hanyu da zidian*: Its future depends on whether the wider repertoire project actually produces infrastructure that can be used in practice.
- If the revisers in Sichuan are able to collaborate effectively with the repertoire teams, and if the encoding and font work is completed in time, then the third edition could become a genuinely digitised dictionary in which users locate variant quotations by typing a head character, a related variant, or even a structural component.
- If that coordination fails, however, the likely result is another print-first dictionary, or a digital edition that remains trapped inside proprietary search software and therefore cannot fully realise the philological ambitions of the revision.