Viewpoints on structure description of Chinese character

Morioka Tomohiko

Center for Informatics in East Asian Studies
Institute for Research in Humanities, Kyoto University

June 18th, 2020

Introduction

Many Chinese characters (漢字) are complex characters composed of multiple components. So we can describe their structures: e.g.

林=⿰木木　　雲=⿱雨云　　広=⿸广厶

But in some cases, their are ambiguity to analyze their structures and components: e.g.

- 旗 = ⿰方⿱其 or ⿸扩其
- 嬴 = ⿱亡⿰月女⿰凡 or ⿸羸女

Who am I?

Works:

- CHISE (CHaracter Information Service Environment) `http://www.chise.org/`
- Bibliography of Oriental Studies on the Web `http://ruimoku.zinbun.kyoto-u.ac.jp/`
- MeCab-Kanbun (Morpheme Analyzer for classical Chinese; Joint research) `https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun/mecab-kanbun`
- etc.

CHISE IDS database

- `https://gitlab.chise.org/CHISE/ids`
  - one of the most comprehensive IDS dataset with a large number of characters that supports almost all CJKV Unified Ideographs coded in UCS.
- CHISE character ontology
  - CHISE IDS database is a part of CHISE character ontology. Each components are defined in the ontology.
- CHISE IDS Find `http://www.chise.org/ids-find`
  - a Web service for searching Chinese characters that contains specified components. It is also an entrance to the CHISE character ontology.

Structural description requirements

There are a lot of Chinese characters, so it is not easy to maintain data quality.

- Versatility: Write once, use anywhere
- Consistency
- Coverage of components: describe all Chinese characters with as few components as possible
- Intelligibility (especially for native users and classical Chinese scholars)

$\rightarrow$ We need models

Description based on apparent structure

Components are a visible objects

- 林 = ⿰木木
- 雲 = ⿱雨云
- Then, if 嬴 = ⿱亡口⿰月女丮,
  is ⿰月女丮 a component?

Description based on functional structure

Component is an interface to associate phonetic and/or semantic values and shapes

→ In this view, 月女孔 is not a component

- If you do not know the target character, you will not know the functional components (maybe it is the goal)

Description based on glyph design variation of component

Component is a unit to describe glyph variations of Chinese characters. cf. unification rules

- 「習」「習」「習」: 「羽」「羽」「羽」

  If an abstract component 〈羽〉 = { 羽 , 羽 , 羽 } *is defined, it is possible to describe abstract character* 〈習〉 = ⿱〈羽〉白

Description based on productivity

Components are objects that combine them to create Chinese characters

- → Components that can produce many Chinese characters have high "componentness".
- → If a component is included in only one Chinese character, it is meaningless to regard it as a component (inappropriate decomposition?)
- Mechanical analysis is possible using the CHISE IDS database

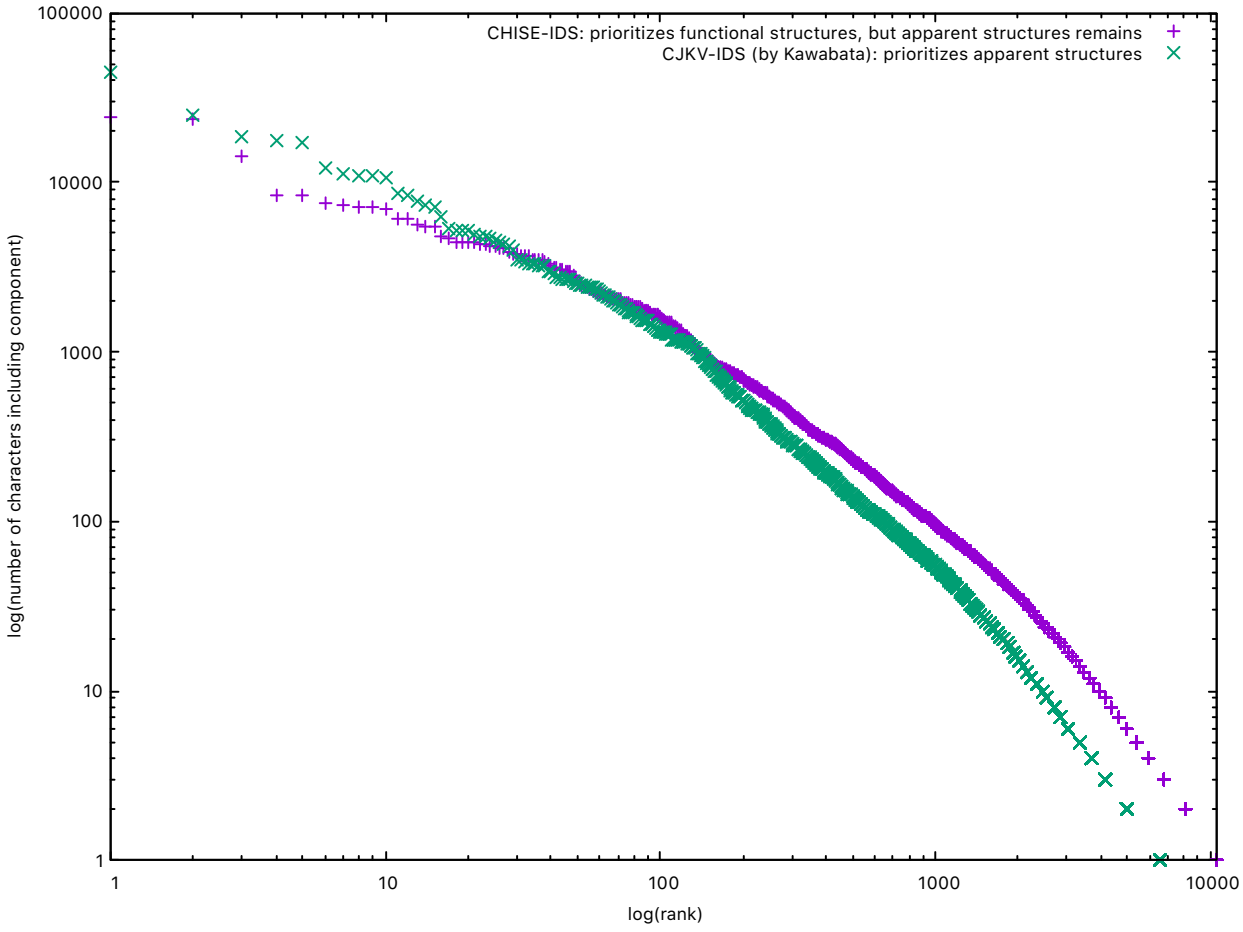
In case 嬴

- 𣎆: 「嬴」「蠃」「贏（贏, 贏, 贏）」「⿳亡口⿲月禾凡」「⿳亡口⿲月木凡」「⿳亡口⿲月歹凡」「⿳亡口⿲月糸凡」「羸」「⿳亡口⿲月叔凡」「⿳亡口⿲月冉凡」「⿳亡口⿲月金凡」「⿳亡口⿲月果凡」「鸁」「⿳亡口⿲月婁凡」「⿳亡口⿲月魚凡」「⿳亡口⿲月鳥凡」「⿳亡口⿲月重凡」「⿳亡口⿲月畐凡」「⿳亡口⿲月喬凡」「羸（羸）」「⿳亡口⿲月虫凡」「⿳亡口⿲月畐凡」「⿳亡口⿲月角凡」 ...
- ⿳亡口月女凡: 「嬴」「⿳亡口⿲月女凡」「瀛（⿳亡口⿲月女凡）」「籯」「籝」

In case 族

- 㫃: 斻, 施, 旆, 斿, 旂, 旸, 㫄, 旉, 旄, 㫕, 旅,
旅, 旆, 㫆, 旇, 㫈, 旒, 旋, 㫉, 旌, 旍, 旎, 族,
㫊, 㫋, 㫌, 㫍, 㫎, 㫏, 旐, 㫐, 㫑, 㫒, 旒, 㫓,
旓, 㫔, 旇, 旖, 㫖, 旗, 㫗, 㫘, 㫙, 遊（旟）,
㫚, 㫛, 簱（旅 ←? 旅）, 旚, 瀋（旛 ←? 旛）,
旘, 旛, 旞, 㫝, 㡖, 旜, 旝, 㫞, 㫟, 旒, 旓, 旞,
㫝, 旞, 旟, 㫟, 𣄠, 㫠, 㫡, 菰（㼉）, 旋, 㫉, ...
- 矢: 族, 樤, 鏃

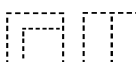
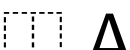
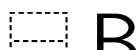
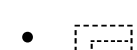
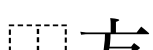
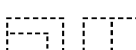
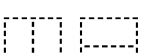
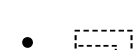
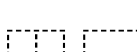
Occurrence of components



This distribution seems to follow the Zipf's law

Equivalence

In many cases, descriptions based on apparent structure and descriptions based on functional structure have equivalent information.
We can write rewriting rules: e.g.

-  ABC →  A  BC
  (旗：  扩 其 →  方箕)
-  ABC →  ABC
  (頦：  须女 →  彡女頁)

Term Rewriting Systems (TRS) can also normalize glyph variants with unification rules.

Ambiguity of apparent structure

虚：□虎业 → □卢⊟七业 → ⊟卜□厂⊟七业

Apparent component is also depended on knowledge.

Conclusion

- Structural description of Chinese character should be based on Chinese character analysis (Chinese character studies), like grammatical analysis of natural language.
- It depends on knowledge, but statistical analysis for CHISE-IDS database helps discover this knowledge.
  - productivity of components
- Grapholinguistic model and algebraic model (such as Term Rewriting System) are the two wheels to describe structure of Chinese characters.