Analyzing of Chinese Character Form Database (CCFD) and the Study of CCFD

Jian-Yu LIU

School of Chinese Literature, Bohai University, 19 Keji Road, Songshan District, Jinzhou, Liaoning, P.R.C.

prcmap@163.com

Keywords: Chinese Character Form Database (CCFD), the Study of CCFD, Present Situation.

Abstract: After the combination of the study of Chinese character and the computer information science, Chinese Character Form Database (CCFD) has been a new thing, which can meet the objective requirement of the current collecting and research work for Chinese character. CCFD is an important foundation of modernization and informatization of Chinese character. The Study of CCFD is an interdisciplinary subject, based on theoretical and practical researches, and the study of CCFD can provide a powerful theoretical weapon for the collecting and the research of Chinese character in the information era. At present, the hypostatic construction of CCFD and the systematic research of the study of CCFD are still in the primary stage, but there will be a promising and significant development in the future.

The 21st century is the information age and Big Data era. Language and script are the two most important information carriers. Language informatization is the basis of the informatization of the whole society. By the means of information, the systematic, comprehensive and in-depth study of Chinese and Chinese character is necessary for China to be in an invincible position in the 21st century.

Innovation is the life of academic scholarship, and academic development is inseparable from the use of new methods and new techniques. In the past 20 years, with the gradual popularization of computers, computer technology has been applied to study Chinese character and gotten a lot of theoretical and practical results; Chinese information processing has become a core area in the current study of Chinese and Chinese character. However, comparing with the mature Chinese corpus and Chinese corpus linguistics, it is still relatively weak to collect and study Chinese character by computer technology and fall far behind the objective requirements of character collection normative work. To change this situation, we must use Chinese Character Form Database (CCFD), and establish a new discipline—the study of CCFD as soon as possible to guide practical work of Chinese character collecting and research.

CCFD

The Chinese Character form (CCF) data and CCFD derive from the conception of linguistic data and Corpus relatively. The CCF data refers to Chinese form material collected and organized from the actual text for Chinese character ontology research. CCFD is not a simple pile of the CCF. In reference to the relevant names in corpus, a character which has not been annotated can be regarded as raw CCF and a character has been annotated can be considered as mature CCF. The value of mature CCF is much higher than the raw one. Only the CCF is annotated, then they will provide various kinds of Chinese character information and be a useful resource which conforms to the research requirements of the systematically collation and research on Chinese characters. [1] CCFD means the large-scale electronic character resource library with the objective of collecting and researching characters, on the principle of linguistics and philology, by collecting the written form of characters which stands for specific character or character variety in actual usage, through computer technology. CCFD is a kind of ordered aggregation of real character written form based on the
The Study of CCFD

The Study of CCFD is an interdisciplinary subject, which is based on character materials in real texts in order to define and describe the phenomenon as the starting point of Chinese characters through the collection, storage, annotation, retrieval and statistical analysis to put forward new theory or verifying and amending existing theories of Chinese characters, and learning how CCFD applies to all levels of Chinese teaching and other related subjects.

The research includes four fields such as CCFD-driven research, CCFD-based research, Applied CCFD research and research on other subjects. The Study of CCFD not only has the dual attributes of social science and natural science, but also has the dual properties of theoretical and applied science. The study contents include the CCFD ontology and construction entity, research on character attributes and regularization based on CCFD, CCFD application research etc. The CCFD-driven research belongs to normative, using the bottom-up method of induction. The CCFD-based research belongs to an empirical research repository, using the top-down method of interpretation.

The Study of CCFD has not only opened up a whole new field of research, but also provides a new research approach. Just like corpus-based linguistic research has become a mainstream language, The Study of CCFD will also undoubtedly become mainstream of the Chinese philology in Big-Data era. Being different form “intuitive data” and “data-induced” in the past research, CCF data in CCFD are comprehensive and empirical. By the qualitative and quantitative description and summary of the actual use of CCFD materials provided by the Chinese character facts, there will be a fundamental guarantee for the accuracy and reliability of the study conclusions. The Study of CCFD is an interdisciplinary subject, based on theoretical and practical researches, and the study of CCFD can provide a powerful theoretical weapon for the collecting and the research of Chinese character in the information era.

Present Situation Analysis

At present, the hypostatic construction of CCFD and the systematic research of the study of CCFD are still in the primary stage, there are still many problems to be solved.

First, there are many problems in the hypostatic construction of CCFD. To do a good job, one must first sharpen one’s tools. Only by establishing some CCFD with the highest standards and the deepest processing, researchers can search, statistics, compare and analyze a variety of valuable material conveniently and efficiently. However, there has not been a CCFD widely recognized with the above function and in public usage. Even though CCFD or “Resemblance CCFD” have been developed, such as the CCFD of Beijing Normal University (BNUZLK, see Fig.1), the CCFD of Dictionary of Bohai University(CCFDDt. , see Fig.2), the Ancient and Modern Texts and Ancient Dictionary Repository of East China Normal University (http://www.wenzi.cn/web/content.aspx?moduleid=22&parentid=20), and the Structure of Chinese
Characters Database of “Academia Sinica” in Taiwan (http://xiaoxue.iis.sinica.edu.tw), Stone Rubbings Database of the University of Tokyo (http://coe21.zinbun.kyoto-u.ac.jp/djvuchar?query=), etc.

Fig. 1. the CCFD of Beijing Normal University.

Fig. 2. the CCFD of Dictionary of Bohai University.

These databases can just only meet the requirements of the lower levels of research. A further debugging and optimization of the research is need in the framework design, characters’ acquisition, the degree of fidelity, data integration, attribute tagging, quality inspection, Chinese characters relationship organization, system functional design and the overall efficiency, etc. This problem should be focus on for some time in the present and the future to solve.

Secondly, the theory ontology research in CCFD and the study of CCFD is very weak. There are still many gaps in theoretical system and many problems need to be solved, such as the comparison of similarities and differences between CCFD and Corpus, the definition of CCFD, founding background and conditions, theoretical basis, subject status, relations with related disciplines, research of theoretical and practical significance, main content and methods, key points, bottlenecks encountered, the concept of system, internal classification, the application of data and a series of macro and micro theoretical issues related to CCFD entity construct. CCFD and the ontology theory research are still quite heavy and arduous.

Finally, the researchers of CCFD and the study are still needed further improvement. In addition to a good open scientific thinking and innovation, the researchers should also have a solid system of
multi-disciplinary knowledge base, master the scientific research methods and techniques, which are the literacy for scholars engaging in interdisciplinary research. Nowadays, the number of scholars who specialize in the study become more and more, but there is a small part of researchers often take the computer as research tools to conduct study and even a smaller part to combine the computer database application systems development and knowledge of Chinese characters to form expertise in two areas of study. Therefore, training researchers who engage in CCFD and the study is a very urgent task.

Prospect Outlook

The 21st century is the information technology and network era, the era of knowledge economy prosperity and the time based on knowledge and innovation. It is believable that in the near future, as more and more scholars attach importance to and actively involved in the study, along with the appearance of a group of a certain size CCFD, the study based on large-scale Chinese character database will be carried out gradually. The study of CCFD, as a new interdisciplinary, will subsequently grow and keep pace with the corpus linguistics, which will have a significant and far-reaching influence on the development and improvement of Chinese graphology and information science in the new century.

Acknowledgement

This research was financially supported by the Youth Program of the National Social Science Fund of China (Grant No.14CYY060) and the Youth Program of the Humanities and Social Science Fund of Ministry of Education of China (Grant No. 13YJCZH117)

References
