The Extended Architecture of Hantology for Kanji

Ya-Min Chou Ming Chuan University Taipei, Taiwan milesymchou@yahoo.com.tw

Chu-Ren Huang Academia Sinica Taipei, Taiwan churen@sinica.edu.tw

Jia-Fei Hong National Taiwan University Taipei, Taiwan jiafei@gate.sinica.edu.tw

Abstract

Chinese writing system is not only used by Chinese but also used by Japanese. The motivation of this paper is to extend the architecture of Hantology which describes the features of Chinese writing system to integrate Japan Kanji into the same ontology. The problem is Chinese characters adopted by Japan have been changed, thus, the modification of the original architecture of Hantology is needed. A extended architecture consists orthographic , pronunciation ,sense and derived lexicon dimensions. is proposed in this paper. The contribution of this study is that the extension architecture of Hantology provides a platform to analyze the variation of Chinese characters used in Japan. The analytic results of variation for a specific Kanji can be integrated into Hantology, so it is easier to study the variation of Chinese characters systematically.

1. Motivation

Hantology has been created to provide the linguistic resources for Chinese processing (Chou, 2005; Chou & Huang, 2006; Chou & Huang, 2007). The current version of Hantology provides 2100 high- frequency used Chinese characters. But Hantology only take in consideration of Chinese characters used in China and Taiwan. Actually, Japan has been using Chinese characters for more than one thousand years. In Japan, these Chinese characters are named Kanji and still be used wildly in modern Japanese writing system. If Japan Kanji can be integrated in Hantology, it will be an important resource for studying the distribution and variation of Chinese characters between China and Japan.

2. The Introduction of Hantology

Chinese language uses a different writing system with others. Chinese characters are ideographic writing system and have been used for more than 3000 years. Chinese writing system is more complicated than phonetic systems. Lots of useful knowledge provided y Chinese characters are not properly represented in computer systems for further studying. In recent years, Hantology has been developed to represent knowledge of Chinese characters for researchers and Chinese information processing. Hantology is able to represent the orthographic forms(glyphs) , the evolution of script, , pronunciations, senses, variants, lexicalization for different Chinese characters.

To demonstrate the contents of the Hantology, Chinese character '家' is taken as an example. The figure 1 illustrate partial content of Hantology for Chinese character '家'. It shows the composition, the principle of formation, glyph expression, evolution of glyph, variants and pronunciations. The content of Hantology indicates:(1) The composition of '家' have a semantic symbol and phonetic symbol. The composition of Hantology is to decompose the structure of characters into several symbols. The symbols used in Chinese characters can be divided into semantic and phonetic symbols. (2) The principle of formation is '形聲'(semantic & phonetic). The principle of formation is to describe the method of creating characters. (3) The glyph evolution illustrates that the lesser seal script of '家' is '豭' which was used in China about two thousands years ago. (4) The glyph expression of '家' is '宀㒳豕' .The glyph expression is to describe the structure of Chinese characters. The glyph expression used in Hantology is developed by Jung and Hsieh(2005) (5)Hantology describes the variant relation between different characters. Variants are different characters with the same pronunciation and sense. Figure 1 indicates '家' and '傢' are variants with the sense '移去' (move). This is a loan sense of '家' . The reason is the pronunciations of '家' and '傢' are the same.

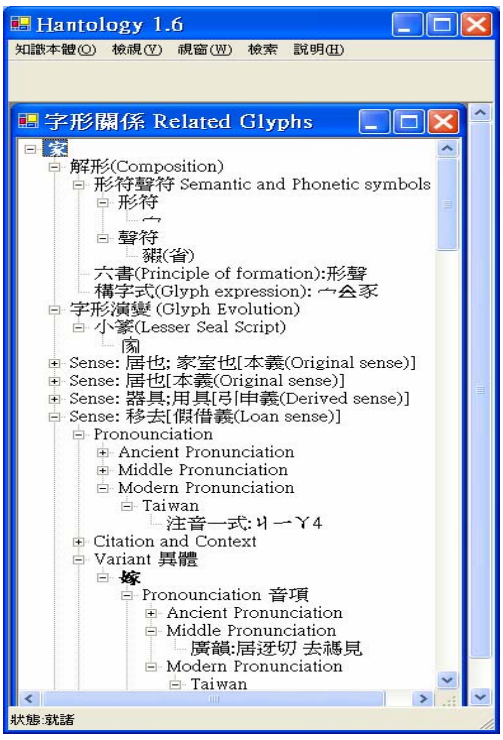

Figure 1:The orthographic forms , evolution of glyph, pronunciation and variants of Chinese character '家'

The figure 2 is another partial content of Hantology for Chinese character '家'. It describes the original sense , modern senses and generated words of characters. (1)The original sense of '家' is '住所'(means home) according to '說文' (Shuo Wen). ShuoWen is an ancient dictionary of Chinese characters. Hantology based on ShuoWen to represent the original sense of Chinese characters (2)One of the senses of '家' is family. The senses of Chinese characters also are mapped to SUMO(Suggested Upper Merged Ontology)(Niles & Pease, 2003;Huang, 2004). (3)The generated words of '家' are '家父' and '家母' when '家' means family. (4) According to the position at the generated words, Hantology indicates prefix, suffix and infix for '家'. There are some information not illustrated in figure 1 and figure 2. First, there are 540 semantic symbols used by Chinese characters. The sense of each semantic symbol has been created in Hantology and mapped into SUMO. Second, Hantology has a formal representation by using OWL.

Figure 2: The original sense, modern senses and generated words of Chinese character '家'

3. The Extension of Architecture

The original architecture t of Hantology is designed for Chinese characters, so the content of Hantology does not any description for Kanji. To integrate Kanji into Hantology, the extended architecture is proposed in this paper and can be classified into four dimensions as follows:

(i) Orthographic extension

The Orthographic forms of Japan Kanji and Chinese characters are not all the same. The Hantology are needed to include different he Orthographic form of Japan Kanji. These different Orthographic forms of Japan Kanji are caused by many reasons. The first reason is that the Orthographic forms of Kanji have many variants in different locations of Japan. Some Kanji variants are adopted from Chinese ancient characters. These Kanji variants increase writing and reading complexity. The second reason is Japan not only use Chinese characters but also invent hundreds Kanji which even be used in Chinese. The extended architecture of Hantology has been designed to represent the orthographic forms of Japan Kanji and relationships with Chinese characters.

(ii) Pronunciations extension

The pronunciations of Kanji are different with Chinese characters and can be classified into Ondoku (おんどく) and Kunyomi (くんよみ). The Ondoku is similar to Chinese pronunciation. Kunyomi is pronounced by Japanese language. For example, Kunyomi reading of '海' is umi(うみ). Ondoku reading of '海' is kai(かい). To integrate Kanji in Hantology, Ondoku and Kunyomi must both be part of Hantology framework. A Kanji may have many Ondoku and Kunyomi. These features are similar with Chinese characters.

(iii) Senses extension

The meaning of Kanji and Chinese characters are similar but not always the same. The orginal architecture of Hantology can also classify the meaning of Chinese into original, derived and loaned. The original sense is added in Hantology only for Kanji invented by Japan.

(iv) Derived lexicons extension

Kanji can derive many lexicons. Some derived lexicons are also used in Chinese. For instance, Kanji '家' derived '家出', '家見', '家後', '家集', '家職'. These lexicons are not used in Chinese language. But '家父', '家人', '家計', '家譜' are also used in Chinese language. In addition, even the same lexicon may have different meaning. For instance, '勉強' are used in Chinese and Japanese , but the meaning are different. The extension architecture is illustrated in figure 1. There are two matrixes for Chinese characters and Kanji writing system. In this figure, 'C' stands for Chinese characters. 'S' stands for the sense. 'P' stands for the pronunciation. 'K' stands for: Japan Kanji. 'R' stands for Ondoku. 'Q' stands for Kunyomi. The senses of Chinese characters and Kanji are mapped to SUMO(Suggested Upper Merged Ontology). Generally, every Kanji have Ondoku and Kunyomi. But, Kanji invented by Japan do not have Ondoku pronunciation.
There are links between Chinese characters and Kanji. These links can let users compare the usage and variety

between China and Japan for a specific character. The links can be divided into two categories. The type I links are created when Chinese characters and Kanji have the same orthographic form. The type II links are created when Kanji have been modified from Chinese characters

The extension architecture of Hantology is able to describe the variants for Chinese characters and Kanji.The variants are the important features of Chinese characters and Kanji.

The variants are two different orthographic forms have the same meaning and pronunciation. For example, in Chinese, variants ‘體’ and ‘体’ have the same meaning and pronunciation, however , the orthographic forms are different. Sometimes, Chinese character and Kanji have the same orthographic form, however, both of them might have different variants. For example, Kanji ‘欠’ and ‘缺’ are the variants because they have the same meaning and pronunciation in Japanese. On the contrary, ‘欠’ and ‘缺’ are not the variants in Chinese,

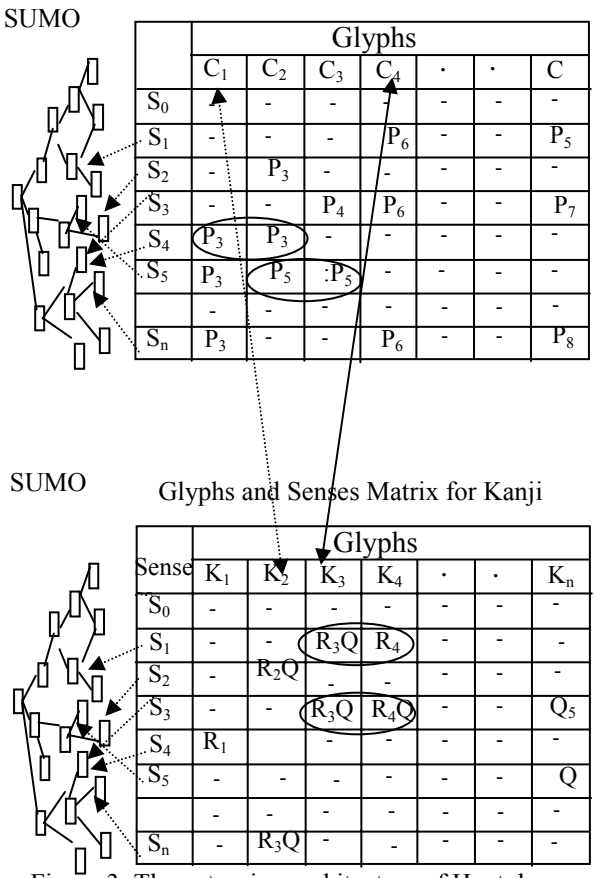

Figure 3: The extension architecture of Hantology for Kanji

To demonstrate the content of Hantology after extension, Chinese character ‘家’ is used as an example. In figure 4, the most right sub-window which shows the contents of Kanji ‘家’ is a new part in Hantology browser. The indications provided by Hantology are as following: (1) ‘家’ is also used in Japan Kanji.(2) Kunyomi reading of ‘家’ is ‘いえ’ and ‘や’. (3)Ondoku reading is ‘ちん’,’か’ and ’け’.(4)Kanji ‘家’does not have any variants. But, Chinese character ‘家’ and ‘傢’ are variants. Furthermore, figure also shows anther partial content of ‘家’ and ‘傢’. It indicate Chinese ‘家’ and ‘Kanj ‘家’ have the same generated word ‘人家’ ,’家臣’,’家事’,’作家’and ‘住家’ But , ‘家路’,’家請’,’家內’,’家出’,’家元’,’家路’,’家並’ are words only used in Japan.

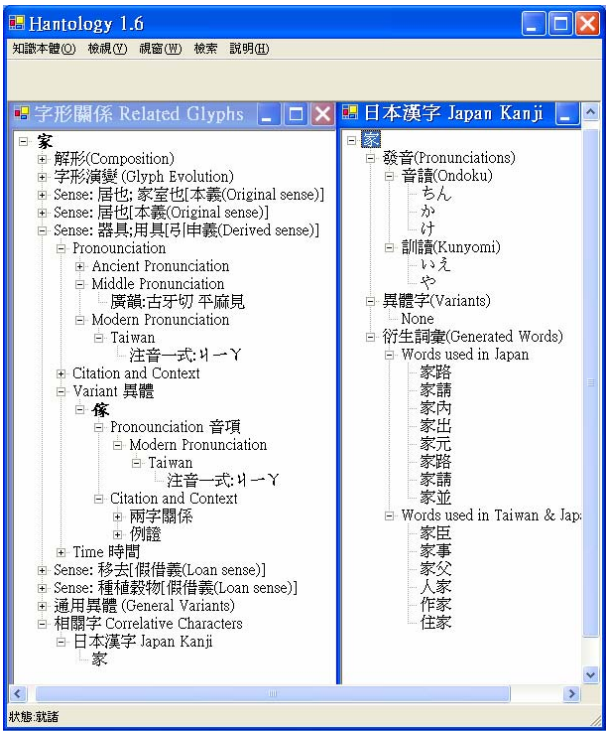

Figure 4: Chinese character ‘家’ and Kanji 家’

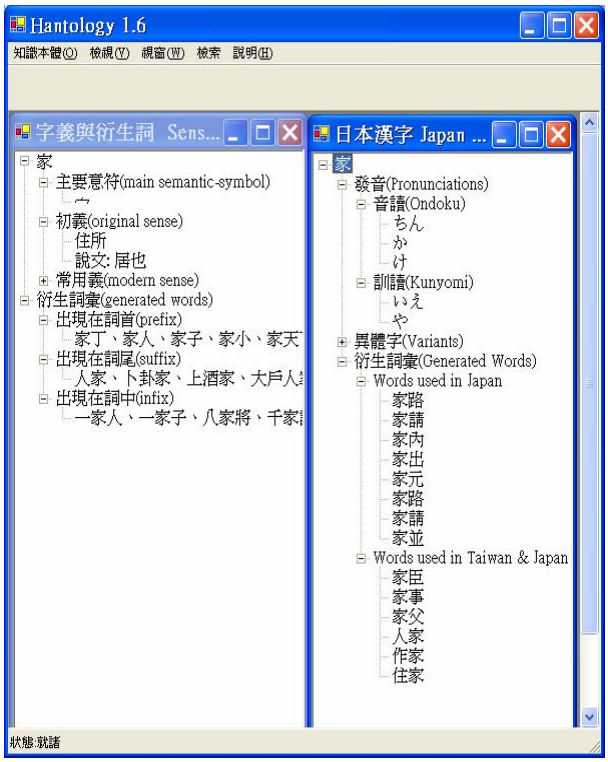

Figure 5: Chinese character’家’and Kanji ‘家’(con’d)

4. Conclusion

The goal of this study is to extend the architecture of Hantology to integrate the Chinese characters and Kanji. To extend the architecture, the original architecture needs to be modified . The basic criteria is to minimize the modification in order to make integration easier. The extension of Hantology consists of orthography, pronunciation and lexicon dimensions. The orthographic forms of Chinese character and Kanji is connected, so that it is easy to understand the different context of Chinese characters used in China and Japan. The new architecture of Hantology is able to represent both Chinese characters and Kanji. It shows the flexibility of original architecture of Hantology.

The contribution of this study is that an integrated ontology of Chinese and Japan Kanji can be created with the extension architecture of Hantology. The extension architecture of Hantology provides a platform to analyze the variation of Chinese characters used in Japan. The analytic results of variation for any specific Kanji can be integrated into Hantology, so it is easier to study the variation of Chinese characters systematically.

5. Acknowledgements

This work was supported in part by the National Science Council under the Grants NSC 95-2411-H-228-001, NSC-96-2411-H-228-002-MY2

6. References

Chou, Y.M. and Huang, C.R., Hantology:an Ontology based on Conventionalized Conceptualization, *Ontologies and Lexical Resources for Natural Language Processing,* Cambridge Press (in press)

Chou, Y.M. and Huang, C.R. (2006), Hantology: Linguistic Resources for Chinese Language Processing and Studying, *Proceedings of Language Resources and Evaluation*, Genoa , Italy, May, 24-26.

Chou, Y.M. (2005). *Hantology-The Knowledge Structure of Chinese Writing System and Its Applications*. Unpublished Dissertation. National Taiwan University

Huang, C.R., Chang, R.Y. and Lee, S.B. (2004) "Sinica BOW (Bilingual Ontological WordNet): Integration of Bilingual WordNet and SUMO." *Proceedings of the 4th International Conference on Language Resources and Evaluation*,. Lisbon. Portugal, pp. 1553-1556.

Jung, D.M. and Hsieh ,C.C.(2005). The Construc-tion and Applications of Chinese Characters Data-base [In Chinese.], *International Conference on Chinese Characters and Globalization*, January 28-30, Taipei, Taiwan.

Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*,. Las Vegas, Nevada, pp. 412-416