Hanzi Grid

Toward a Knowledge Infrastructure for Chinese Character-based Cultures

Ya-Min Chou[1], Shu-Kai Hsieh[2], and Chu-Ren Huang[3]

[1] Jin-Wen Institute of Technology, Taiwan
[2] National I-Lan University, Taiwan
[3] Institute of Linguistics, Academia Sinica, Taiwan

Abstract. The long-term historical development and broad geographical variation of Chinese character (Hanzi/Kanji) has made it a cross-cultural information sharing platform in East Asia. However, due to the lack of proper research framework, the integration of heterogeneous knowledge grounded in Hanzi and its variants has been a thorny problem. In this paper, we propose a theoretical framework for the knowledge representation of Hanzi in the cross-cultural context. Our proposal is mainly based on two resources: Hantology and Generative Lexicon Theory. Hantology is a comprehensive Chinese character-based knowledge resource created to provide a solid foundation both for philological surveys and language processing tasks, while Generative lexicon theory is extended to catch the abundant knowledge information of Chinese characters within its proposed qualia structure. We believe that the proposed theoretical framework will have great influence on the current research paradigm of Hanzi studies, and help to shape an emergent model of intercultural collaboration.

1 Introduction

1.1 Motivation

Chinese character (Hanzi) has existed for over thousands of years. The long-term historical development and broad geographical variation of Hanzi has made it a valuable resource for multi-linguistic and cross-cultural mediation in Asia, especially among the *Sinosphere* (a new-coined term also known as 漢字文化圈 Chinese character cultural sphere), which denotes a grouping of regions and countries where Chinese characters were adopted and integrated to their languages, or historically under Chinese cultural influence (Bradley et al (2003)).

However, due to the lack of proper research framework in information science, the integration of heterogeneous knowledge grounded in Hanzi and its variants has been a thorny problem. To achieve that ultimate goal, the first attempt is to provide a linguistic proper and computational interoperable framework, which can facilitate collecting and integrating the usage and knowledge of Chinese characters and their variants, in different spatial and temporal dimensions. In this paper, we propose a Knowledge-driven Hanzi-centered framework for diachronic and cross-cultural knowledge representation. We hope that the construction of this proposed framework will facilitate the intercultural collaborative research on Chinese characters.

1.2 Interfacing Ideographic Script and Conceptual Knowledge

In the history of western linguistics, writing has long been viewed as a surrogate or substitute for speech, the latter being the primary vehicle for human communication. Such "surrogational model" (Harris (2000)) which neglects the systematicity of writing in its own right has also occupied the predominant views in current computational linguistic studies. This paper is set to provide a quite different perspective along with the Eastern philological tradition of the study of scripts, especially the ideographic one i.e., Chinese characters (Hanzi).

For phonological writing systems, a character as a writing unit usually represents a phoneme or a syllable. Since the number of phonemes is finite, only a small set of phonetic symbols is needed to represent sounds of words for such languages. A character in the Chinese writing system, however, is a writing unit that represents a compact package of concept and pronunciation. Through over 3000 years of use, the complete Chinese writing system consists of at least 40,000 characters, over 100,000 variants are counted. In theory, it is still possible to invent and add new characters to the inventory today, although in practice, this is a very rare event. Each Chinese character represents one or more different concepts. Like alphabetic or syllabic characters, the Chinese characters serve as a basis of lexical classification. Unlike phonological writing systems, however, the classification is largely conceptual or semantic. In other words, the linguistic ontology of Chinese characters is explicitly marked with logographic features.

From the view of writing system and cognition, human conceptual information has been regarded as being *wired* in ideographic scripts. We believe that the conceptual knowledge information which has been *grounded* on Chinese characters could be used as a cognitively sound and computationally effective ontological lexical resource in performing some NLP tasks, and it will have contribution to the cross-cultural collaboration in the *Sinosphere* within the context of *Semantic Web* as well.

1.3 Hanzi and Conventionalized Conceptualization

Since Chinese characters act as information bearers for over thousands of years, some researchers proposed that the whole set of Chinese characters can be viewed as an *encyclopedia* in essence. In terms of knowledge representation, we prefer to refer to it as a kind of ontological knowledge. But, can an ontology be psychologically real and be evidenced by shared human experience? This is one of the critical issues that linguistic ontologies, such as WordNet (Fellbaum (1998)), tries to answer. The successful applications of WordNet seem to give a positive reply to this question. However, all the conceptual relations (or lexical semantic relations) of WordNet are annotated by experts, not conventionalized. Hence there is no direct evidence of the psychological reality.

Based on our observation, we find that Chinese writing system can be treated as a linguistic ontology since it represents and classifies lexical units according to semantic classes. Having been used continuously for over 3000 years, it has conventionalized a system of semantic classification. The system is richly structured and robust, and adopted by other languages belonging to different language families. For example, Chinese characters have been incorporated into the writing systems of Japanese (called *Kanji*), Korean (called *Hanja*), and Vietnamese (called *Chunom*), which belong to Japonic, Altaic (debated) and Austro-Asiatic language family, respectively. [4]

2 Hanzi as a Multi-Levels Knowledge Resource : Review of Current Works

In general, a Chinese character is an *ideogram* composed of mostly straight lines or "poly-line" strokes. A number of characters contain relatively independent substructures, called components (or glyphs), and some common components (traditionally called radicals) are shared by different characters. Thus, the structure of Chinese characters can be seen to consist of a 3-layer affiliation network: *character, component (glyph)* and *stroke.*

[4] Besides Japanese, Korean, and Vietnamese, a number of Asian languages have historically been written with Chinese characters, or with characters modified from Han characters. They include: Khitan language, Miao language, Nakhi (Naxi) language (Geba script), Tangut language, Zhuang language and so on.

Linguistically, a Hanzi is regarded as an ideographic symbol representing *syllable and meaning* of a “morpheme” in spoken Chinese, or, in the case of polysyllabic word, one syllable of its sound. Namely, character, morpheme and syllable are *co-extensive*.

The fact that characters can be investigated from different angles, resulting in the various approaches to the knowledge mining from them. For example, there are several studies on the creation of Chinese characters database. One important study is Chinese glyph expression database which consists of 59000 glyph structures (Juang and Hsieh, 2005). The glyphs of Chinese characters are decomposed into 4766 basic components. Each Chinese character can be expressed by the basic components. Chinese glyphs database also contains oracle bone, bronze, greater seal and lesser seal scripts. The largest Chinese characters database is Mojikyo font database which contains more than 110000 characters (Ishikawa, 1999). Both Chinese glyph expression database and Mojikyo font database contain only glyph knowledge. Yung created an ancient pronunciations database for Chinese characters (Yung, 2003). Hsieh and Huang (2006) proposed an ontological lexical resource based on Chinese characters called HanziNet, in which Chinese characters are located within the context of upper level ontology.

These previous studies unveil many single dimensions of Chinese characters. However, each Chinese character consists of glyphs, scripts, pronunciations, senses, and variants dimensions. To meet the need of computer applications as well as the Chinese philological studies, Chou and Huang (2005) propose a language resource called Hantology (Hanzi Ontology). The construction of Hantology focuses on the comprehensive and robust description of linguistic and conceptual knowledge encoded in Chinese writing system through the decomposition and composition of Chinese characters.

The linguistic knowledge described in Hantology includes various information concerning glyph, script, pronunciation, sense, and variants of Chinese characters. In addition, Hantology has been mapped to SUMO (Suggested Upper Merged Ontology), and it is fully encoded in OWL for shareability and for future Semantic Web applications as well.

3 Radical-centered General Theoretical Framework

In this and following sections, we will propose a radicals-centered general framework for the knowledge representation of Hanzi.

Generally speaking, each Chinese character is composed of two parts: a radical representing semantic classification, and a phonetic indicating phonological association. This generalization applies to the majority of Chinese characters, though not all. A minority estimated at less than 20% of all Chinese characters show other forms of composition. However, it is still true that these characters contain at least one semantically significant component. A small set of examples based on the radical 馬 (*ma3*, 'horse') are given below to show the range of assigned meanings. In these examples, 馬 is both a character and a radical denoting 'horse':

驩:	a kind of horse
驫:	many horses
騎:	to ride a horse
驍:	a good horse
驚:	to be scared (referring to a horse)

These Chinese characters shown above suggest that radicals are indeed concept-based. However, it also been shown the *conceptual clustering is more complex than a simple taxonomy.* In this paper, we focus mainly on the knowledge structure of radicals of Chinese characters, whose significance in representing semantic taxonomy was first observed by Shen Xu (Xu, 121). The following will elaborate on these observations.

3.1 Bootstrapping Conceptual Representation with Chinese Radicals

Any formal account of a conceptual system faces the dilemma of choosing a representational framework. Since a representational framework is itself build upon certain conceptualization, any choice is potentially an *a priori* distortion of the account. A possible solution to this dilemma is a shared upper ontology that is conceptually complete and yet general and robust enough to cover different conceptual systems under consideration. Take the Hantology for example, the Suggested Upper Merged Ontology (Niles and Pease, 2001) was adopted in this resource. All concepts expressed in Chinese characters are mapped to SUMO representation in the hope that the mapping can be transformed to a specialized ontology later.

One of the first implications of adopting SUMO representation is the fact that we are now able to formally represent knowledge inference based on the linguistic knowledge provided by radicals. Recall that the radical part of a Chinese character encodes semantic classification. For instance, all characters containing the 魚 ('fish') radical can be assigned to this SUMO's concept with the same knowledge.

Applying the linking between SUMO and WordNet (Niels and Pease, 2003) the following inference is possible whenever an English word is classified as a hyponym of fish in WordNet. However, no labor intensive manually classification is needed for Chinese. The same inference can be achieved automatically with logographic information, by assigning the default inference rule to all characters with the 魚 ('fish') radical.

```
(subclass Fish ColdBloodedVertebrate) (disjointDecomposition
ColdBloodedVertebrate Amphibian Fish Reptile) (=>
   (instance ?FISH Fish)
   (exists
      (?WATER)
      (and
         (inhabits ?FISH ?WATER)
         (instance ?WATER Water))))
```

3.2 The Ontology of a Semantic Radical: A Generative Lexicon Approach

One illuminating discovery that we make while trying to map radicals to ontology nodes is that each radical actually represents a cluster of concepts that can be associated to the core meaning by a set of rules. We take the 艸 (cao3, 'grass') radical for instance. It is generally accepted that 艸 represents the concept 'plant'.

Of the 444 characters containing the semantic symbol 艸, there is no doubt that they are all related to the concept 'plant'. But what is interesting is that the conceptual clustering is not simply of taxonomic classification. As seem in figure 1, there are four productive relations described by the radical: being a kind of plant (e.g., 蘭 ('orchid')), being a part of a plant (e.g., 葉 ('leaves')), being a description of a plant (e.g.,落 ('fallen')), and being the usage of a plant (e.g. 藥 ('medicine')). The concepts of most radicals that represent concrete objects can be classified into name, part, description and usage. For example, the concepts represented by radical 馬 ('horse'), 牛 ('cow'), and 木 ('wood') also could be divided into the same four classes.

We observe that this is similar with theory of generative lexicon, where *formal, constitutive, telic* and *agentive* are the four parts of the qualia structure of a word which describe motivated semantic changes and coercions (Pustejovsky, 1995). It is interesting to note that all except the Agentive aspect were attested with the conceptual clustering of Chinese characters derived from the grass radical. Since Pustejovsky's Agentive aspect is strongly associated with artifacts and

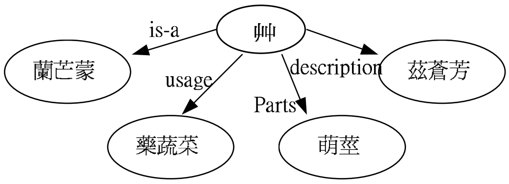


Fig. 1. Conceptual Classes Represented by Semantic Radical 艸 (grass), with derived characters 蘭 (orchid), 葉 (leaves), 落 (fallen), 藥 (medicine) and so on.

other human creations, it is not unreasonable that the radicals based on natural objects lack any obvious semantic extension on how it was created. In addition, the descriptive attributes can be subsumed by the formal aspect of the Qualia structure.

Indeed, the Agentive aspect is attested by a different radical that is conceptually associated with man-made objects. The radical that we take as example is 金 (jin1,'metal'). Since metals are not useful to human in its natural form, they are shaped by human to become different tools. In the conceptual clusters classified according to the semantic radical 金 , there is a substantial sub-set defined by how a metal object was made, as in figure 2.

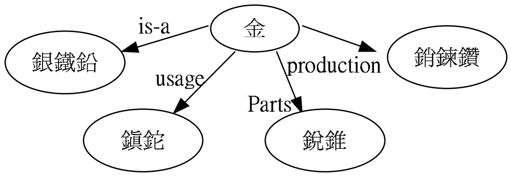


Fig. 2. Conceptual Classes Represented by Semantic Radical 金 (gold), with derived characters 銀 (silver), 鉛 (lead), 鑽 (diamond), 銳 (sharp) and so on.

It is also interesting to observe that there is no instantiation of the Constitutive aspect for the semantic radical 金. This can be easily explained since metal in its natural form is a mass and does not have any components. Hence we show that the seeming idiosyncrasies in the conceptual clustering under each radical are actually dependent on real world knowledge. Hence we find the conceptual structure of encoded by semantic radicals in the Chinese writing system supports Pustejovesky's theory of Generative Lexicon and Qualia structure. These are the same principles used for deriving Chinese characters 3000 years ago suggests that there is cognitive validity.

4 A Sketch on Hanzi-based Intercultural Collaborative Projects

This section depicts the proposed implementation framework and underlying reasons.

4.1 Hantology as a Prototypical Cross-cultural Knowledge Platform

In this paper, we propose to elaborate the overall design of Hantology (Chou, 2005) as a prototypical Cross-cultural Knowledge Platform. The methodological and technical considerations are as follows:

First, Hantology is the first comprehensive linguistic ontology of ideographic writing systems. This approach significantly augments knowledge available to the glyph-based Chinese encoding systems. It also allows this systemic knowledge to be applied to facilitate natural language processing. Most important of all, within this framework, the diachronic changes and synchronic variations can be handled by the same mechanism. In other word, cross-lingual variations of Hanzi can be handled with the same mechanism dealing with the variations of Tang usages. (see Figure 3 and Figure 4).

In addition, the glyphs of Chinese characters have undergone historical changes and regional variations, the glyph of each character is different on different period. These relationships are described in Hantology. The descriptions of glyphs include kaishu, lesser-seal, bronze and oraclebone scripts. If two glyphs have evolution relationships, then, hasAncientGlyph and isAncientGlyphOf predicates are used. hasAncientGlyph and isAncientGlyphOf predicates both have inversed and transitive features that are able to infer evolution relationships. The statements of hasAncientGlyph is shown as follows (also see Figure 5):

```
if hasAncientGlyph(G$_i$, G$_j$ )
  then  isAncientGlyphOf(G$_j$,G$_i$)
```

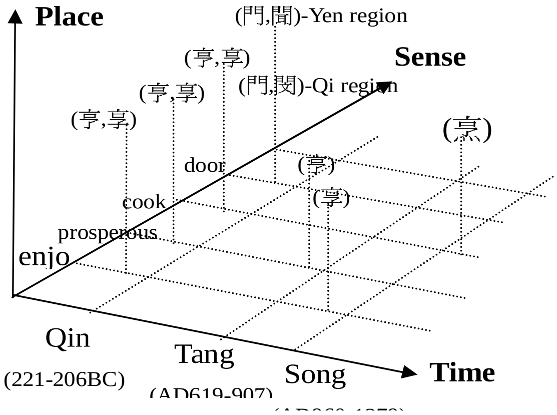


Fig. 3. Character Variants: Temporal and Locational Dependencies

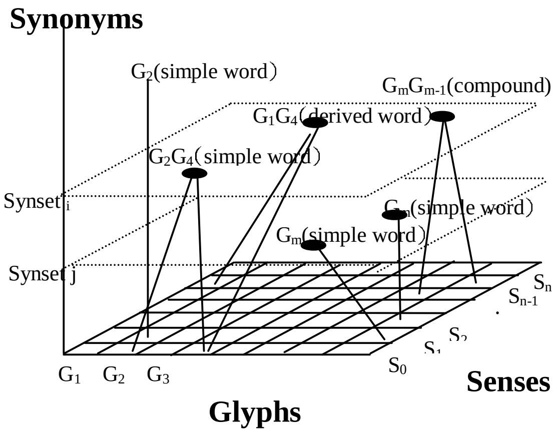


Fig. 4. Words Generated from Chinese Characters

```
if hasAncientGlyph(G$_i$, G$_j$ ) and hasAncientGlyph(G$_j$,G$_k$)
  then hasAncientGlyph(G$_i$, G$_k$ )
```

Second, a linguistic context for describing the relation of character variants is proposed in Hantology. Chinese character variants are an important characteristic of Chinese texts. Unfortunately, so far, the relations of variants have not been properly represented. For this, we proposed a linguistic context for describing the relation of variants. Evaluation results show that this linguistic context provide significant improvement over previous counterpart schemes.

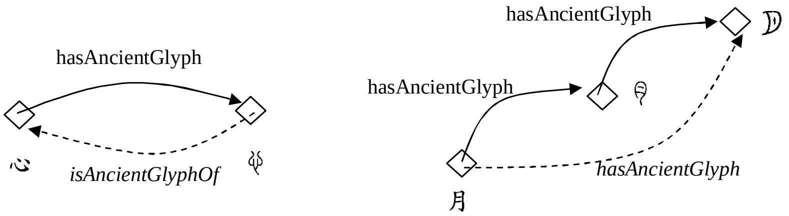


Fig. 5. Ancient Glyphs and Inferred Glyphs

Language always changes over time. Any linguistic ontology should not ignore the variation of language. Hantology is the first linguistic ontology describing the variation of languages. The aspects of variation described by Hantology include orthographic form, pronunciation, sense, lexicalization and variants relation. This approach can systematically illustrate the development of Chinese writing system.

Third, Hantology is knowledge-based and equipped with technologies suitable for web services. Figure 6 and Figure 7 shows the OWL Semantic Model of Glyph in Hantology and Knowledge Structure of Animal-related Radicals interplayed with SUMO ontology.

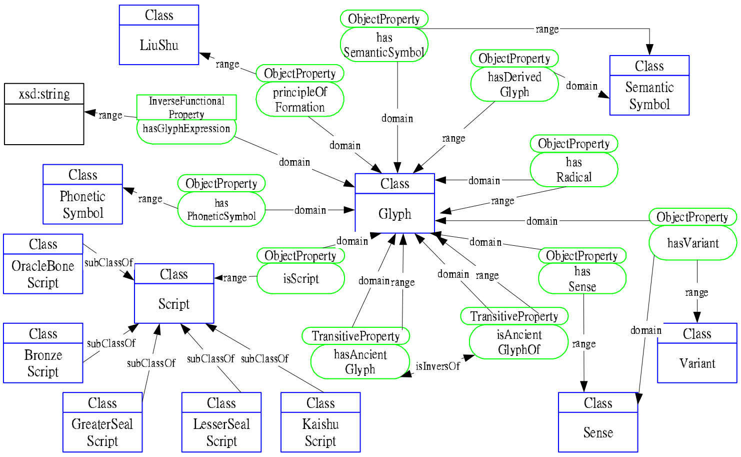
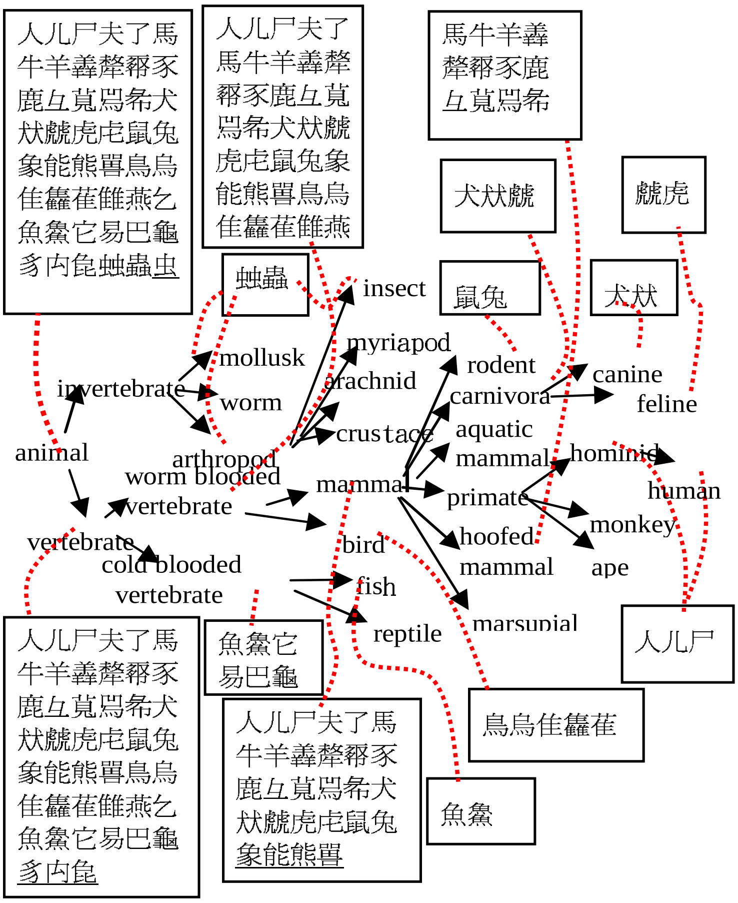


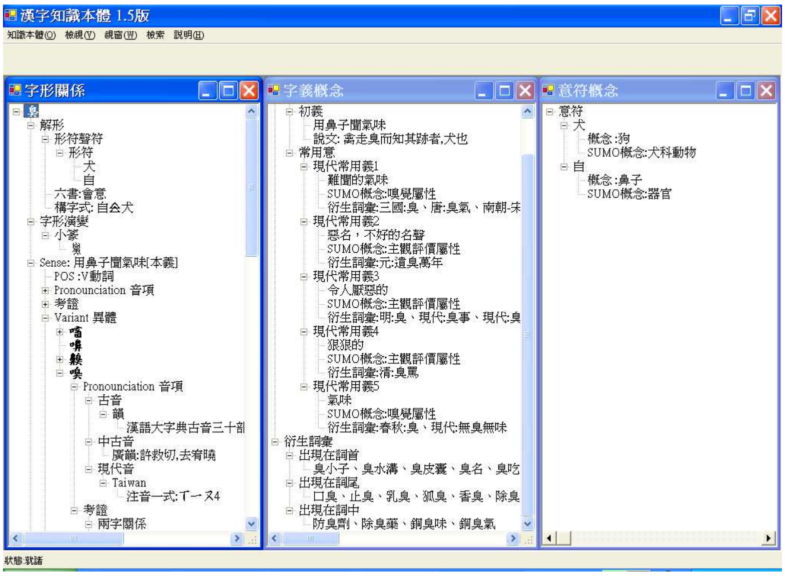
Fig. 6. OWL Semantic Model of Glyph in Hantology



人儿尸夫了馬牛羊羴犛豩豕鹿彑萈舄㣇犬犾虤虎虍鼠兔象能熊羆鳥烏隹雥萑雔燕乙魚鱻它易巴龜豸内㲋蚰蟲虫
人儿尸夫了馬牛羊羴犛豩豕鹿彑萈舄㣇犬犾虤虎虍鼠兔象能熊羆鳥烏隹雥萑雔燕
馬牛羊羴犛豩豕鹿彑萈舄㣇
犬犾虤
虤虎
蚰蟲
insect
鼠兔
犬犾
myriapod
mollusk
rodent
canine
invertebrate
arachnid
carnivora
feline
worm
aquatic
crustace
animal
mammal
hominid
arthropod
worm blooded
mammal
primate
human
vertebrate
monkey
vertebrate
hoofed
bird
cold blooded
mammal
ape
fish
vertebrate
marsupial
reptile
人儿尸夫了馬牛羊羴犛豩豕鹿彑萈舄㣇犬犾虤虎虍鼠兔象能熊羆鳥烏隹雥萑雔燕乙魚鱻它易巴龜豸内㲋
魚鱻它易巴龜
人儿尸
人儿尸夫了馬牛羊羴犛豩豕鹿彑萈舄㣇犬犾虤虎虍鼠兔象能熊羆
鳥烏隹雥萑
魚鱻

Fig. 7. Knowledge Structure of Animal-related Radicals

To illustrate the major contents of Hantology, we use the character 臭 (‘scent’) as an example. The figure 8 shows the glyphs, pronunciations and variants for 臭. The principle of formation is 會意 (‘ideographic compound’), one of the six criteria for character formation (六書 (liu4shu1,‘six methods’). Glyph evolution shows the derivational history of the script. 臭 originated as a verb and referred to the act of smelling by nose. There are four variants for the sense of smell. 後作 in the figure means 臭 is replaced by 嗅 to express the sense of smelling later. The first citation appears in period of 唐 (‘Tang dynasty’, 619AD-907AD). For the aim of intercultural colloboration, an example based on Japanese Kanji is shown in Figure 9.

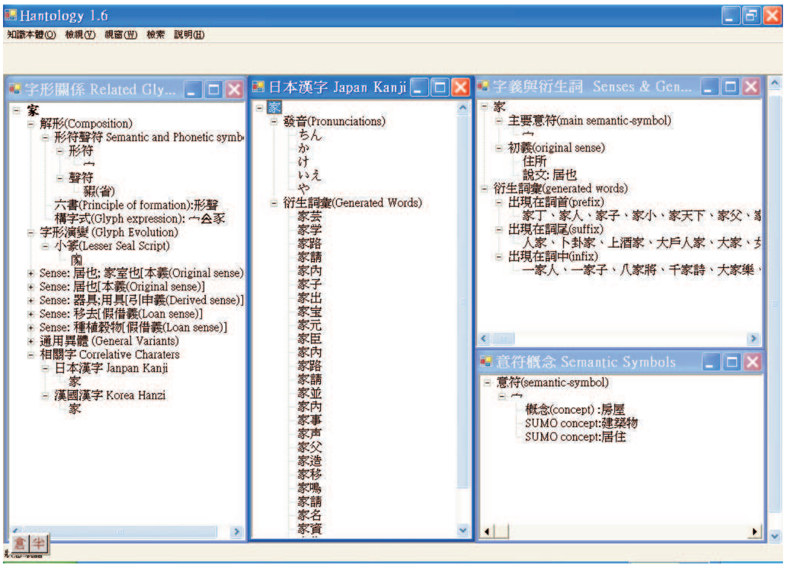


```
⊟ 臭
  ⊟ 解形
    ⊟ 形符聲符
      ⊟ 形符
          犬
          自
      六書:會意
      構字式: 自会犬
  ⊟ 字形演變
    ⊟ 小篆
        臭
  ⊟ Sense: 用鼻子聞氣味[本義]
      POS :V動詞
    ⊞ Pronounciation 音項
    ⊞ 考證
    ⊟ Variant 異體
      ⊞ 嗅
        㚖
      ⊞ 齅
      ⊟ 嗅
        ⊟ Pronounciation 音項
          ⊟ 古音
            ⊟ 韻
                漢語大字典古音三十部
          ⊟ 中古音
              廣韻:許救切,去宥曉
          ⊟ 現代音
            ⊟ Taiwan
                注音一式:ㄒㄧㄡ4
        ⊟ 考證
          ⊟ 兩字關係
```

```
  ⊟ 初義
      用鼻子聞氣味
      說文: 禽走臭而知其跡者,犬也
  ⊟ 常用意
    ⊟ 現代常用義1
        難聞的氣味
        SUMO概念:嗅覺屬性
        衍生詞彙:三國:臭、唐:臭氣、南朝-宋
    ⊟ 現代常用義2
        惡名，不好的名聲
        SUMO概念:主觀評價屬性
        衍生詞彙:元:遺臭萬年
    ⊟ 現代常用義3
        令人厭惡的
        SUMO概念:主觀評價屬性
        衍生詞彙:明:臭、現代:臭事、現代:臭
    ⊟ 現代常用義4
        狠狠的
        SUMO概念:主觀評價屬性
        衍生詞彙:清:臭罵
    ⊟ 現代常用義5
        氣味
        SUMO概念:嗅覺屬性
        衍生詞彙:春秋:臭、現代:無臭無味
⊟ 衍生詞彙
  ⊟ 出現在詞首
      臭小子、臭水溝、臭皮囊、臭名、臭吃
  ⊟ 出現在詞尾
      口臭、止臭、乳臭、狐臭、香臭、除臭
  ⊟ 出現在詞中
      防臭劑、除臭藥、銅臭味、銅臭氣
```

```
⊟ 意符
  ⊟ 犬
      概念 :狗
      SUMO概念:犬科動物
  ⊟ 自
      概念 :鼻子
      SUMO概念:器官
```

Fig. 8. The Glyphs and Variants Knowledge for 臭


Hantology 1.6
知識本體(O) 檢視(V) 視窗(W) 檢索 說明(H)
字形關係 Related Gly...
日本漢字 Japan Kanji
字義與衍生詞 Senses & Gen...
意符概念 Semantic Symbols
倉
半

```
⊟ 家
  ⊟ 解形(Composition)
    ⊟ 形符聲符 Semantic and Phonetic symb
      ⊟ 形符
          宀
      ⊟ 聲符
          豭(省)
      六書(Principle of formation):形聲
      構字式(Glyph expression): 宀会豕
  ⊟ 字形演變 (Glyph Evolution)
    ⊟ 小篆(Lesser Seal Script)
        家
  ⊞ Sense: 居也; 家室也[本義(Original sense)
  ⊞ Sense: 居也[本義(Original sense)]
  ⊞ Sense: 器具;用具[引申義(Derived sense)]
  ⊞ Sense: 移去[假借義(Loan sense)]
  ⊞ Sense: 種植穀物[假借義(Loan sense)]
  ⊞ 通用異體 (General Variants)
  ⊟ 相關字 Correlative Charaters
    ⊟ 日本漢字 Janpan Kanji
        家
    ⊟ 漢國漢字 Korea Hanzi
        家
```

```
⊟ 家
  ⊟ 發音(Pronunciations)
      ちん
      か
      け
      いえ
      や
  ⊟ 衍生詞彙(Generated Words)
      家芸
      家学
      家路
      家請
      家内
      家子
      家出
      家宝
      家元
      家臣
      家内
      家路
      家請
      家並
      家内
      家事
      家声
      家父
      家造
      家移
      家鳴
      家請
      家名
      家資
```

```
⊟ 家
  ⊟ 主要意符(main semantic-symbol)
      宀
  ⊟ 初義(original sense)
      住所
      說文: 居也
⊟ 衍生詞彙(generated words)
  ⊟ 出現在詞首(prefix)
      家丁、家人、家子、家小、家天下、家父、家
  ⊟ 出現在詞尾(suffix)
      人家、卜卦家、上酒家、大戶人家、大家、女
  ⊟ 出現在詞中(infix)
      一家人、一家子、八家將、千家詩、大家樂、
```

```
⊟ 意符(semantic-symbol)
  ⊟ 宀
      概念(concept) :房屋
      SUMO concept:建築物
      SUMO concept:居住
```

Fig. 9. The Kanji Example for 家

Lastly, the missing characters and variants retrieval problems are solved. It is an essential requirement to properly represent characters and symbols for any information processing and philological studies as well. However, current Chinese computer systems fail to meet this requirement for decades. Consequently, users always have to face the missing characters and variants retrieval problem. We propose to change the representation of Hanzi to increase the knowledge owned by computers. By integrating missing characters with Hantology, the missing characters and variants problem are solved successfully.

4.2 Meaning-bearing Radicals as Inter-Lingual-Index

The *Inter-Lingual-Index* (ILI) was proposed by the project of EuroWordNet (Vossen, (1998)), which is used to connect synsets across all the languages. We propose to adopt *Meaning-bearing Radicals* (MBRs, 意符) to provide an efficient mapping across the CJKV languages on the one side, and other European languages on the other sides.

Formally, given a set of Radicals $\mathcal{R}$, it is also a *power set* of $\mathcal{R}$, written $\mathcal{P}(\mathcal{R})$. Hence we can have $\mathcal{P}(\mathcal{R}) = \{\{c, j\}, \{c, k\}, \{c, v\}, \{c, j, k\}, ...\}$, where c, j, k, v denote Chinese, Japanese, Korean and Vietnamese, respectively. These information can thus facilitate more comparative and contrastive studies of characters variants. [5]

4.3 From Hanzi Ontological Lexical Networks to Hanzi Grid

To achieve the goal of collaborative research, there are many software and hardware implementation architecture available, such as Wiki, Semantic Web and Grid. In order to facilitate the process of automatic learning of qualia structure encoded in Meaning-bearing Radicals, we propose to adopt the *LexFlow* Grid computing environment proposed by Soria et al (2006). These are already in preparation.

5 Conclusion

Chinese characters explicitly encode conventionalized conceptualization. It is well-established practice in computational linguistics to manipulate lexical and inter-lexical level knowledge, such as the very active research based on WordNet.

[5] More information such as CJK shared Hanzi and Radicals 中日韓共用漢字表 are available at http://140.111.1.40/fulu/fu5/fu6.htm

However, the knowledge encoded on Chinese characters is intra-lexical and are embedded in the orthography. In this paper, we focused on how to represent the knowledge structure formed by Chinese characters in the cross-lingual and cross-cultural context. We adopt Hantology as a prototypical formal representation of the linguistic ontology conventionalized with the Chinese writing system. We propose that the radicals, the semantic symbols, do form a robust and well-accepted conceptual system, and can be used as ILI. The historical depth of Hantology will allow us to examine how knowledge systems evolve through time. In addition to the foster the knowledge exchange among the Sinosphere, the richly encoded knowledge at the basic writing level will also support multilingual (CJKV) content processing of texts without higher the syntactic processes of segmentation, chunking, or parsing.

References

1. Bradley, David, Randy LaPolla, Boyd Michailovsky and Graham Thurgood (eds). (2003). Language variation: Papers on variation and change in the Sinosphere and in the Indosphere. Canberra: Pacific Linguistics.
2. Chou, Y.M. and Huang, C.R. (2005). Hantology: an Ontology based on Conventionalized Conceptualization. In the *Proceedings of Ontolex Workshop.*
3. Chou, Yia-Min and Chu-Ren Huang. (2005a). 漢字意符知識結構的建立. (The construction of the knowledge structure of meaning components of Hanzi). In: *The 6th Chinese Lexical Semantics Workshop.* Xia-Men: China.
4. Chou, Y.M. (2005). Hantology: The Knowledge Structure of Chinese Writing System and Its Applications. *Ph.D Thesis.* National Taiwan University, Ph.D. thesis.
5. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database, Cambridge: MIT Press.
6. Harris, Roy. (2000). Rethinking Writing. The Athlone Press.
7. Huang, Chu-Ren, Ru-Ying Chang and Shiang-Bin Li. (2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In: *Proceedings of the 4th LREC*, Lisbon:Portugal.
8. Huang, C.R. (2005). Knowledge Representation with Hanzi: The relationship among characters, words, and senses [In Chinese.] *International Conference on Chinese Characters and Globalization.* Taipei.
9. Hsieh, Shu-Kai and Chu-Ren Huang. (2006). When Conset meets Synset: A Preliminary Survey of an Ontological Lexical Resource based on Chinese Characters. *poster paper* at *COLING/ACL 2006*, Sydney.
10. Ishikawa, T. (1999). Mojikyo Font Database, `http://www.mojikyo.org/`
11. Juang, D.M. and Hsieh, C.C. (2005). The Construction and Applications of Chinese Characters Database [In Chinese]. *International Conference on Chinese Characters and Globalization*, Taipei, Taiwan.

12. Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In: *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, Ogunquit, Maine.
13. Pustejovsky, James. (1995). The Generative Lexicon. Cambridge: MIT Press.
14. Soria, Claudia, Maurizio Tesconi, Andrea Marchetti, Francesca Bertagna, Monica Monachini, Chu-Ren Huang and Nicoletta Calzolari. (2006). Towards Agent-based Cross-lingual Interoperability of Distributed Lexical Resources. In: COLING/ACL post-conference Workshop *Multilingual Languageg Resources and Processing*, Sydney.
15. Vossen, Piek. (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.
16. Wong, Shun-Ha. S. and Karel Pala. (2002). Chinese Characters and Top Ontology in EuroWordNet. In *Singh, U.N. (ed).: Proceedings of the First Global WordNet Conference 2002, Indian.*
17. Xu, Shen. (121). 說文解字 (The Explanation of Words and the Parsing of Characters). Cited edition, ZhongHua (2004).
18. Yu, Shiwen, Zhu Xuefeng and Li Feng. (1999). The development and application of modern Chinese morpheme knowledge base. [in Chinese]. In: 世界漢語教學, No.2. pp38-45.