Hantology: An Ontology based on Conventionalized Conceptualization

Ya-Min Chou
Jin Wen Institute of Technology,
Taipei, Taiwan
milesymchou@yahoo.com.tw

Chu-Ren Huang
Institute of Linguistics of Academia Sinica,
Taipei, Taiwan
churen@gate.sinica.edu.tw

Abstract

Hantology is the abbreviated name for Hanzi Ontology, an ontology based on the conventionalized conceptual orthographic system of Chinese characters (or kanji). We treat the Chinese writing system as a linguistic ontology since it represents and classifies lexical units according to semantic classes. This linguistic ontology is robust enough to endure over 3000 years of use by the most populous people, as well as adaptation by neighboring languages. In this paper, this robust and richly encoded ontology is fully and explicitly studied. We map Hantology to SUMO (Suggested Upper Merged Ontology) for systematic and theoretical discussion. In addition, the complete Hantology is fully encoded in OWL for sharability and for semantic web applications.

1. Introduction: Hanzi and Conventionalized Conceptualization

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), try to answer. The successful applications of WordNet in research seem to give a positive reply to this question. However, all the conceptual relations (or lexical semantic relations) of WordNet are annotated by experts, and not conventionalized. Hence there is no direct evidence of the psychological reality. We observe that there is indeed a human language writing system that has conventionalized a system of semantic classification. The system is richly structured and robust, having been used continuously for over 3000 years, and adopted by neighboring languages. This is the writing system of the Chinese characters (*hanzi*, or *kanji* in Japanese). We develop a linguistic ontology to represent the knowledge structure of Chinese characters' radicals, orthographic forms, variants and derived words. In this paper, we focus mainly on the knowledge structure of Chinese characters' radicals.

1.1. The Chinese Writing System

For syllabic and alphabetic writing systems, a character is a writing unit representing a phoneme or a syllable. Because the number of phonemes is finite, we only need finite phonetic symbols to represent sounds of words for a language. For the Chinese writing system, however, a Chinese character is a writing unit that represents a concept. Through over 3000 years of use, the complete Chinese writing system consists of at least 40,000 characters and perhaps has over 100,000 including variants. Each character represents one or more different concepts. Unlike alphabetic or syllabic characters, the Chinese characters not only represent concepts, they are also classified according to a set of semantic symbols. In other words, the linguistic ontology of Chinese characters is explicitly marked with logographic features.

1.2. Logographic Features of Chinese Characters

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 are given below to show the range of assigned meanings. In these examples, 馬 *ma3* is both a character and a radical denoting 'horse':

驪: a kind of horse
驫: many horses

騎: to ride a horse
驍: a good horse
驚: to be scared (referring to a horse)

The Chinese characters shown above suggest that radicals are indeed concept-based. However, it also showed the conceptual clustering is more complex than a simple taxonomy. We continue with the exploration of the system of conceptualization governing the constructions of Chinese characters.

1.3 Bootstrapping Conceptual Representation of 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. We adopt the Suggested Upper Merged Ontology (SUMO, Niles and Pease 2001) in this study. 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 show how knowledge inference can be achieved with the linguistic knowledge provided by radicals. Based on the linking between SUMO and WordNet (Niles and Pease, 2003) the following inference is possible whenever an English word is linked to the 'fish' node in SUMO:

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

What is unique with Chinese characters is that the writing system encodes this conceptual link without any extra cost. Basically, all characters containing the 魚 'fish' radical can be assigned to this SUMO's concept with the same knowledge inference.

2. General Framework (Chou 2005)

In this paper, we elaborate the theoretical motivation as well overall design of Chou's (2005) dissertation. Chou (2005) proposes Hantology, a formal explicit representation of conceptualization for Chinese writing system. In this thesis, he takes Chinese characters are fundamental linguistic units and important resources for natural language processing. Although it is widely accepted that the Chinese writing system is richly encoded with semantic information, a formal account (or ontology) of this knowledge system has never been proposed. Chou (2005) focused on how to represent the knowledge structure of Chinese writing system. Hantology is proposed as a framework that is designed to give a felicitous and robust description of the knowledge structure of Chinese characters and the Chinese writing system.

Hantology describes orthographic forms, phonological forms, senses, variants, variation and lexicalization of Chinese writing system. The orthographic forms of Chinese writing system is ideographic or word-syllable characters. In general, each Chinese character is not only a writing unit but also itself a word or morpheme. The most important feature of Chinese writing system is that orthographic forms and senses are extensions of semantic symbols, so the concepts indicated by semantic symbols become the core of Chinese writing system. In this study, we use 540 radicals of ShuoWenJieZi (Xyu 121) as basic semantic symbols. To enable the conception and relation of semantic symbols to be processed by computer systems, the concepts indicated by each radical are analyzed and mapped into IEEE Suggested Upper Merged Ontology (SUMO). In addition, adopting SUMO allows Hantology to integrate and share with other ontologies like WordNet or the Academia Sinica Bilingual Ontological WordNet (Sinica BOW, Huang et al. 2004).

The senses represented by each Chinese character are mapped to SUMO to formally account for their conceptualization and their dependency relations. The derived lemmas are organized by their different senses in order to express the morphological context. Since some senses are dependent on

their pronunciations, the relation between pronunciations and senses are described. In Chinese writing, there are lots of variants which are different orthographic forms of the same word or morpheme. A linguistic context is proposed to describe the relations of variants. To make knowledge easily sharable, we establish a model expressed by Web Ontology Language-Description Logic (OWL-DL). This model integrates General Ontology for Linguistic Description (GOLD, Farrar and Langendoen 2003) framework to provide linguistic meta-knowledge, such as orthography, morphology, and syntax, for natural language processing.

Lastly, a knowledge-based solution to the problems of missing characters encoding and interpretation, as well as robust information retrieval of character variants are given to exemplify Hantology's application in NLP. We propose an interchange framework and Missing Characters Description Language (MCDL) for describing missing characters. Experiment results show that the missing characters and variants retrieval problems can be solved successfully by Hantology and interchange framework proposed in this thesis.

In sum, Chou (2005) made substantial contributions in the following areas:

First, the proposal and construction of a new linguistic ontology describes the knowledge structure of Chinese writing system. Hantology is the first 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.

Second, we propose a linguistic context for describing the relation of character variants (Huang et al. 2005). 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.

Third, we propose a knowledge-based framework to describe language variation. 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.

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. 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.

3. Conceptualization and Classification of the Radicals System

Since the invention of the Chinese characters is the beginning of conventionalization of the Chinese language, there is no documentation of the principles governing the construction. Fortunately, we do have a classical text that is reasonably close to the origin of the Chinese characters. This is 'The Explanation of Words and the Parsing of Characters' ShuoWenJieZi (Xyu, 121). ShuoWenJieZi identifies 540 radicals (*bu4shou3,* literally 'head of classification') for Hanzi. Although later studies, including of excavated data, necessitate revisions of the *ShuoWenJieZi* system and interpretation, it is still the most widely accepted system that also happens to be the closest to the original interpretation.

3.1. Mapping Radical-based Semantic Classes to SUMO

In order to faithfully show the original conceptualization of Chinese characters, we take the complete set of radicals of the *ShuoWenJieZi* as semantic symbols. For formal representation as well as ease of comparative study, they are mapped to the SUMO upper ontology. The conceptual classification of each radical is determined based on two sources of information. The first is the dictionary explanation given in *ShuoWenJieZi.* However, this source may contain errors caused by the author's idiosyncratic interpretation or the lack of data. For instance, *ShuoWenJieZi* give distinctive account of

the two bird-related radicals. It is said that 鳥 (114 derived characters) refers to long-tailed birds, while 隹(38 derived characters). However, when these 152 bird-related characters are examined, the generalization cannot be attested. Hence we conclude that the concept for both radicals is simply 'bird'.

The second, and in fact the more authentic and reliable, piece of evidence we use is the shared meaning of the family of derived characters sharing that radical. That is, the most reliable evidence to identify the semantic class is to look at the shared semantics of the class members. The most productive radical among the 540 radicals is 水 'water', which has 467 characters derived from it. In is important to note that some radicals may represent more than one concept and hence require multiple inheritance, as allowed by SUMO. An example is 雨 'rain', which will be linked both to the ontology node of 'water' and 'WeatherProcess.' Sinica BOW (Huang et al., 2004) is used when looking up for the SUMO correspondences for some characters. Most mapping tasks rely on human analysis.

4. The Ontology of a Semantic Radical

One illuminating discovery that we made 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 'plants'.

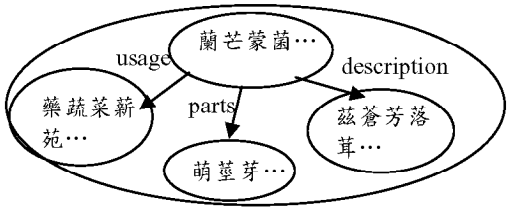

Figure 1. Conceptual Classes Represented by Radical 艸(*CAO3)*

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 surprising is that the conceptual clustering is not simply of taxonomy classification. As seem in figure 1, there are four productive relations described by the radical: being a kind of plant (e.g. orchid), being a part of a plant (e.g. leaves), being a description of a plant (e.g. fallen (leaves)), and being the usage of a plant (e.g. 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), 羊(goat) and 牛(cow) also could be divided into the same four classes.

We observe that this is similar with theory of generative lexicon (Pustejovsky 1995), where formal, constitutive, telic and agentive constitute the qualia structure of a word and provide the motivations for semantic changes and coercions. 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 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 金 radical, there is a substantial sub-set defined by how a metal object was made, as in Figure 2.

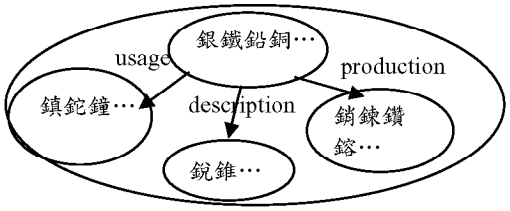

Figure 2. Conceptual Classes Represented by Radical 金 (*JIN1*)

It is also interesting to observe that there is no instantiation of the Constitutive aspect for the 金 metal radical. 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 is actually dependent on real world knowledge. Hence we find the conceptual structure of encoded by 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.

5. The Architecture of Hantology

Based on what described above, each radical is the head of a semantic class. Hence, each radical forms a small ontology itself, governing all the concepts and words derived from it. Each concept then can be mapped to a SUMO ontology node. Hence we have a matrix system linking two sets of hierarchically ordered ontology, as shown in Figure 3.

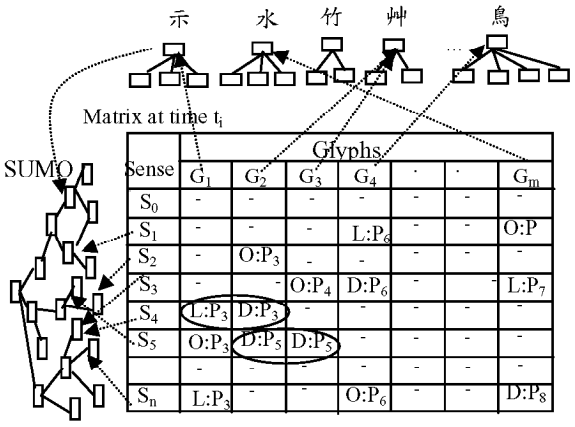

Figure 3. Mapping Hantology to SUMO at Time t_i

The content of Hantology is time sensitive. Since language and writing systems change over time, there will be variations of the Chinese writing systems too. With a temporal scale added to Hantology, we will be able to trace the form and meaning changes over time. For any specific time, as illustrated in fFgure 3, each radical is actually represented by a small ontology, which is the clustering of concepts related to the head concept. These concepts can be mapped to the shared upper ontology of SUMO. But most important of all, the linguistic coding space is a matrix of senses x glyphs. The figure shows some of the possible relations among the characters. In this diagram, 'O' stands for the original meaning, 'D' stands for derived meaning, and 'L' stands for loaned meaning; while 'P' stands for pronunciation. The design of Hantology allows each historical era to be represented; hence the evolution of the linguistic ontology can be observed.

Most of Chinese characters represent concepts. The same concept may be represented by many different Chinese characters. After thousands years, there are very complicated variant relation among Chinese characters. The design of all computers' encoding systems, including Unicode, is based on alphabetic writing systems. For computers encoding system, each character is assigned a unique code. If two codes are different, then, they are assumed to be different characters. However, for Chinese writing system, this assumption is too strong. In Chinese writing system, different characters codes may be the same characters. For example, 說 and 説 are the same Chinese characters but encoded with different Unicode. Actually, 說 and 説 are variants. Variants are the main reasons that Computer applications are not able to process Chinese characters properly. Because variants relations are important features in Chinese writing system, we develop a framework to describe variants relations. This framework constitutes with several dimensions including sense, pronunciation, time, place, and constraints of derived words.

(1) sense

Most Chinese characters are morphemes or words. The sense dimension concerned what senses both characters can represent the same concepts.

(2) pronunciation

The sense of Chinese characters depends on the pronunciation. If the pronunciation changes then the sense are also different.

(3) time

Variant relation is not stable because the sense changes over time. It is important to describe the dynamic features of variant relation.

(4) place

Because Chinese characters don't represent directly the pronunciation of words, they are used by different place.

(1) constraints of derived words

Although the same concept can be represented

by variants, some words only use specific character.

In Figure 3, glyph G2 and G1 are variants in sense 4 at time t_i. Glyph G2 and G3 are variants in sense 5 at time t_i. Figure 4 illustrates the interaction among sense, time, and place. 亨 and 享 were variants of the same character before Qin. However, they did have variant relation after Tang. On the other hand, 門 and 閔 were variants of the concept door in the Yen region, but, 閔 did not have the door meaning in the Qi region.

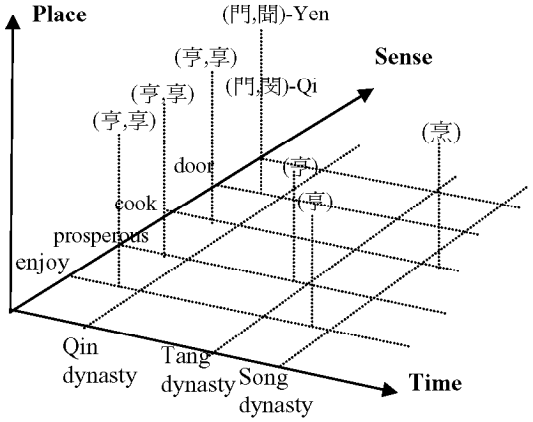

Figure 4. Character Variants: Temporal and Locational Depencies.

Chinese characters can generate words. In modern Chinese language, most words consist of two characters. To reflect this feature, the words generated from each character are described in Hantology. Figure 5 illustrate the relation among characters, words and synonyms.

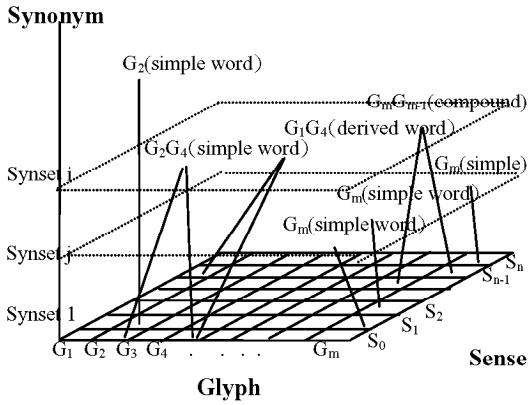

Figure 5. Words Generated from Chinese Characters.

6. Formal Representation of Hantology

The Semantic Web initiative not only underlines the need for automatic semantic processing by the web and highlights the crucial role of ontology as the infrastructure of knowledge representation. Hence, the fact that there is a widely used linguistic ontology with overt encoding of semantic classes is significant. It is worthwhile to see covert this linguistic ontology to a formal representation that can be accessed in the Semantic Web. Since OWL (Web Ontology Language) has been designated as the Web Ontology Language for W3C, we adopt OWL-DL to formally represent Hantology. The successful implementation is significant in two ways (see Figure 6). First, it will facilitate exchange and processing of knowledge represented in Chinese texts ad well as allow web-based applications. For instance, since Japanese texts are also encoded in Chinese characters (i.e. kanji), Hantology can serve as an infrastructure for exchanging Japanese to Chinese information. Second, converting the Hantology to formal representation allow us to check the consistency of the ontology.

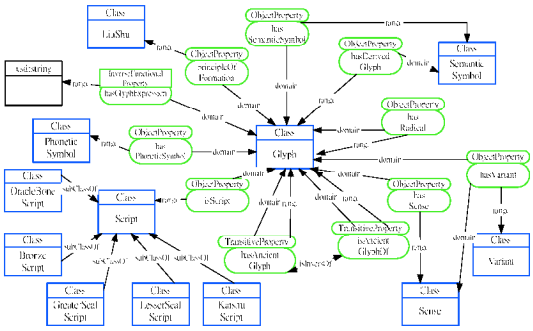

Figure 6. OWL Semantic Model of Glyph in Hantology.

Unicode differentiates many characters which only have micro difference at glyph level. Actually, they are only variants of the same characters. This glyph-based design may cause many problems. For instance, these variants may be treated differently even though they should have the same semantic content. This problem cannot be solved with encoding systems. It would be necessary to explicitly describe the relations among these characters. Adopting OWL-DL can solve this problem. OWL-DL has inference function that will allow computers to identify Chinese characters properly the character variants:

If hasGlyphInUnicode(G_k,$Unicode_i$) and hasGlyphInUnicode(G_k, $Unicode_k$)
then $Unicode_i$=$Unicode_j$

for examples：

if hasGlyphyInUnicode(G_i, 說) and hasGlyphyInUnicode(G_i, 説)
then 說＝説

if hasGlyphyInUnicode(G_j, 研) and hasGlyphyInUnicode(G_j, 研)
then 研＝研

if hasGlyphyInUnicode(G_k, 眾) and hasGlyphyInUnicode(G_k, 衆)
then 眾＝衆

In addition, because Chinese characters have been used for thousands years, 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 7):

if hasAncientGlyph(G_i, G_j)
then isAncientGlyphOf(G_j, G_i)

if hasAncientGlyph(G_i, G_j) and
hasAncientGlyph(G_j, G_k)
then hasAncientGlyph(G_i, G_k)

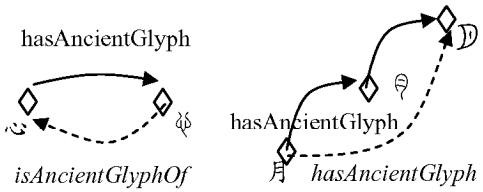

Figure 7. Ancient Glyphs and Inferred Glyphs

The name space of Hantology on the web is http://www.ntu.edu.tw/2005/Hantology.owl# . We give a semantic model of the part of Hantology that describes Glyph as an example in figure 6.

7. Results

7.1. Towards a Knowledge System based on Chinese Characters

The knowledge structure of Chinese radicals has been built. There are 3000 high frequent characters described in Hantology, including their orthographic forms, senses, variants and generated words. The whole knowledge structure formed by radicals is large, so only a part of results are shown in this paper. Figure 8 shows the knowledge structure of radicals about animals. This draft ontology can be pruned and re-arranged later to show more faithfully the system of conceptualization as encoded when Chinese characters were invented.

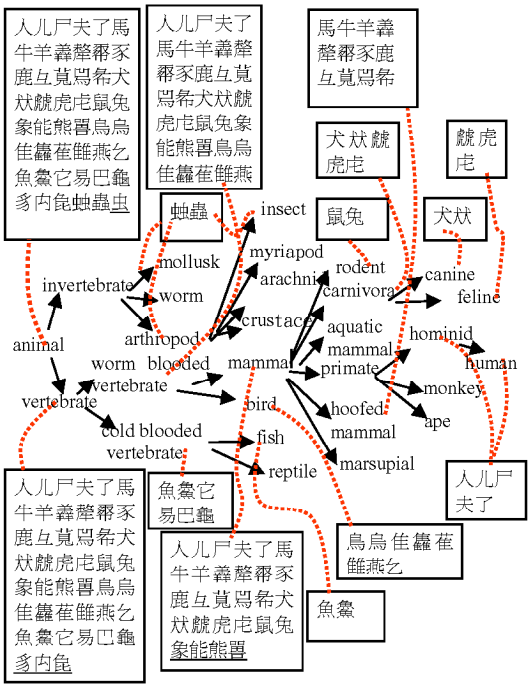

Figure 8. Knowledge Structure of Animal-related Radicals

7.2. Accessing Hantology

As a knowledge base that has a rich time-depth as well as glyph representations, we designed an interface such that the many-layered Hantology knowledge can be effectively accessed. With this interface, a user can browse by form, including the semantic components of a character, by meaning, or by variants. A web-based version is being constructed now for wide access.

Figure 9. Radical Classification by SUMO

In the above figure, we show one of the features of our interface system. For this application, a user can identify all radicals according to an ontological concept as defined in SUMO. A possible application, of course, is for a web-based user to execute a meaning-based web-search without having to know the Chinese language and also achieve much higher conceptual recall without having to list all related lemmas.

8. Conclusion

Chinese characters explicitly encode conventionalized conceptualization. However, this knowledge structure has never been utilized in language processing before. 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. 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. This knowledge is an important part of Hantology. Hantology is a formal representation of the linguistic ontology conventionalized with the Chinese writing system. We show that the radicals, the semantic symbols, do form a robust and well-accepted conceptual system. In addition to explore the possibility of representing a conceptual system that has been implicitly followed by users of the same writing system, we also tried to explicitly define the relations within the system and make the information useful. The historical depth of Hantology will allow

us to examine how knowledge systems evolve through time.

References

Chou, Ya-Min. 2005. Hantology-The Knowledge Structure of Chinese Writing System and Its Applications. Unpublished Dissertation. National Taiwan University.

Chou, Ya-Min and Chu-Ren Huang. 2005. Construction of a Knowledge Structure based Chinese Radicals.[In Chinese] Proceedings of the Sixth Chinese Lexical Semantics Workshop. Xiamen. April 21-24.

Farrar, Scott, and Terry Langendoen 2003. A Linguistic Ontology for Semantic Web. GLOT International. 7.3.97-100.

Fellbaum Christiane.1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.

Huang, Chu-Ren., Chang, Ru-Yngm and Sian-Bing Lee. 2004 "Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO." Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal, pp. 1553-1556.

Huang, Chu-Ren. 2005. Knowledge Representation with Hanzi: The relationship among characters, words, and senses [In Chinese.] Presented at the International Conference on Chinese Characters and Globalization. January 28-30. Taipei.

Hung, Jiafei, Chu-Ren Huang, and Yiching Wu. 2005. Towards a Study on the Lexical Semantics of Character- and Word-Variants [In Chinese]. Proceedings of the Sixth Chinese Lexical Semantics Workshop. Xiamen. April 21-24.

Niles, I., and Adam Pease. 2001. "Toward a Standard Upper Ontology". Proceedings of the 2nd International Conference on Formal Ontology in Information Systems, Ogunquit, Maine.

Niles, I., and Adam Pease. 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, 412-416.

Pustejovsky, James. 1995. The Generative Lexicon, Cambridge: MIT Press.

Xyu, Shen. 121. ShuoWenJieZi 'The Explanation of Words and the Parsing of Characters'. Cited Edition. Beijing: ZhongHua(2004).