Hantology-A Linguistic Resource for Chinese Language Processing and Studying

Ya-Min Chou, Chu-Ren Huang+

Jin Wen Institute of Technology +Institute of Linguistics of Academia Sinica

Taipei, Taiwan

milesymchou@yahoo.com.tw + churen@gate.sinica.edu.tw

Abstract

Hantology, a character-based Chinese language resource is created to provide an infrastructure for language processing and research on the writing system. Unlike alphabetic or syllabic writing systems, the ideographic writing system of Chinese poses both a challenge and an opportunity. The challenge is that a totally different resources structure must be created to represent and process speaker's conventionalization of the language. The rare opportunity is that the structure itself is enriched with conceptual classification and can be utilized for ontology building. We describe the contents and possible applications of Hantology in this paper. The applications of Hantology include: (1) an account for the diachronic development of Chinese lexica (2) character-based language processing, (3) a study of conceptual structure differences in Chinese and English, and (4) comparisons of different ideographic writing systems.

1. Introduction

The appearance of WordNet has been adopted by many researchers to solve natural language processing problems (Fellbaum, 1998). Several Chinese WordNets based on WordNet also have been developed recently. However, these Chinese WordNets are not able to describe the features of Chinese characters (Huang et al., 2004; Chen et al.; Chang, 2003). In addition, many problems that rarely exited in other natural language processing can not be solved by Chinese WordNets. A typical problem in processing Chinese text is the missing characters problem which is the characters not been encoded in computer systems Chou, 2005).

Another problem is the computers can not process variants correctly. Chinese characters have lots of variants which are the different glyphs with the same word or morpheme. For instance, both characters '体' and '體' are the same word and morpheme with different glyphs. Actually, they are variants and can be replaced each other. These problems also cause information retrieval and interchange problems. The WordNet does not represent the relations of variants because the foundations of WordNet are the synonyms. Synonyms are different with variants. Synonyms are the same meaning with different words. Variants are the same word with different forms. The Sentences, phrases, words or characters all are different forms. For computer systems, it is very important to know the meaning and concept carried by this different form. For English, each character is just a writing unit without carrying any concept. Therefore, it does not have the requirement to build resources for alphabetic characters. However, each Chinese character is not only a writing unit but also a concept unit. Because there are lots of relations among concepts, the characters are not independent of each other.

The knowledge of languages represented in computers is not only able to solve the problems of natural language processing but also to provide resources for research of languages, so language resources are critical to computational linguistics. However, researchers who study Chinese characters lack resources and have difficulty to get benefits from computers. Although there are several Chinese characters databases have been created, these databases only focus on glyphs or pronunciations of characters and have sharing difficulty with other applications. The purpose of this paper is to introduce Hantology and its applications for computer systems and researchers.

2. Relative Works

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 & 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 proposed a HanziNet which represent Characters by 16 bits binary code (Hsieh, 2005). Chinese characters are classified into hierarchy categories. HanziNet can describe the upper layer concept of a character. These previous studies only conceded on one dimension of Chinese characters. However, each Chinese character consists of glyphs, scripts, pronunciations, senses, and variants dimensions. The previous studies can not provide enough knowledge for computer applications and researchers. Chou and Huang propose an ontology named Hantology to provide glyph, script, pronunciation, sense, and variants of Chinese characters (Chou, 2005; Chou & Hung, 2005)

3. The Contents of Hantology

Hantology describes orthographic forms, phonological forms, senses, variants, variation and lexicalization of Chinese writing system. The orthographic composition of a Chinese word is either ideographic or radical-phonetic pairing. 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 ShuoWen as basic semantic symbols(Xyu, 121). To enable the conceptualization 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 of each Chinese character also adopt SUMO to represent the conception and relation among various senses. The lexicons generated by different senses are constructed to express the morphological context. Since the senses depend on pronunciations, the relation between pronunciations and senses are described by Hantology. 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 illustrate the major contents of Hantology, we use the character '臭' as an example. The figure 1 shows the glyphs, pronunciations and variants for '臭'. The principle of formation is '會意'(ideographic compound). Glyph evolution shows the ancient script of '臭' is '臭'. '臭' is a verb when it used as to smell by nose. There are four variants for the sense of smell. '後作' in the figure 1 means '臭' is replaced by '嗅' to express the sense of smell. The first citation appears in period of '唐' (Ton dynasty, 619AD-907AD).

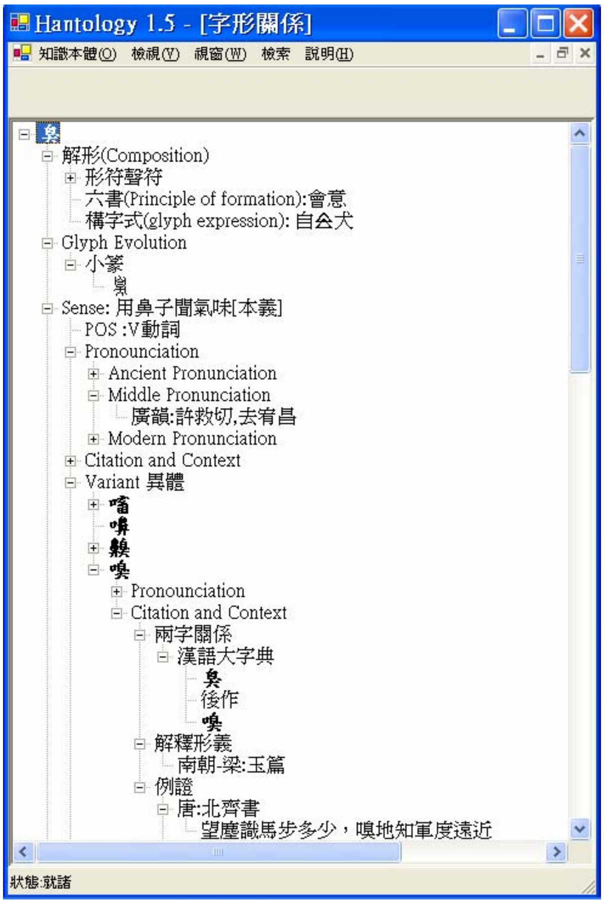

Figure 1: The Glyphs and Variants Knowledge for '臭'

The figure 2 shows the senses and generated words of '臭'. The original sense is to smell by nose. All senses are mapped into SUMO. For instance, the sense of '臭名' (a bad reputation) is mapped into the concept '主觀評價屬性'(Subjective Assessment Attribute) in SUMO. The generated word from sense of '臭名' is '遺臭萬年' which appeared in period of '元' (Yuan dynasty, 1280AD-1368AD).

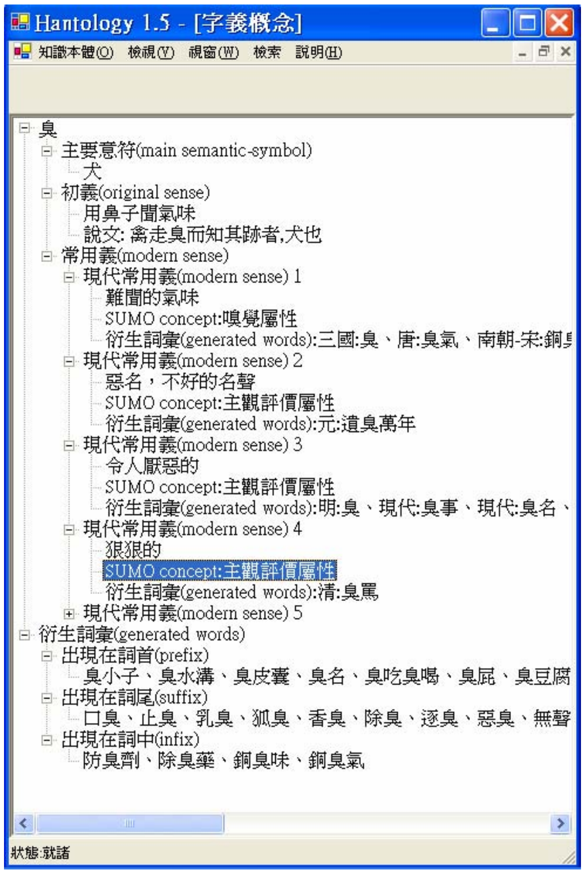

Figure 2: The Senses and Generated Words for '臭'

To make knowledge be shared easily, we establish a model expressed by Web Ontology Language-Description Logic (OWL-DL) and integrate with General Ontology for Linguistic Description (GOLD) to provide the writing, morphological, syntactical knowledge for natural language processing (Farrar &Langendoen, 2003). The figure 3 shows the part of OWL-DL of Hantology.

```
<?xml version="1.0"?    encoding="UTF-8"?>
<rdf:RDF xmlns=http://www.ntu.edu.tw
/2005/01 /Hantology.owl#
xml:base="http://www.ntu.edu.tw
/2005/01/ Hantology.owl">
xmlns:gold="http://www.emeld.org/gold.owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-
rdf-syntax-ns#"
xmlns:sumo="http://reliant.teknowledge.com
/DAML/SUMO.owl#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#
xmlns:goldlinguistics="http://www.owl-ontologies.com/u
nnamed.owl#"
<owl:Ontology rdf:about="">
<owl:imports
rdf:resource="http://protege.stanford.edu/plugins
```

```
/owl/protege"/>
<owl:imports
rdf:resource="http://reliant.teknowledge.com
/DAML/SUMO.owl"/>
<owl:imports
rdf:resource="http://www.owl-ontologies.com
/unnamed.owl"/>
<owl:versionInfo>2005-01-01, edited by Ya-Min Chou
</owl:versionInfo>
</owl:Ontology>
<owl:Class rdf:ID="Glyph">
    <rdfs:label>字形</rdfs:label>
</owl:Class>
<owl:Class rdf:ID="Pictograph">
    <rdfs:subClassOf rdf:resource="#LiuShu"/>
    <rdfs:label>象形</rdfs:label>
</owl:Class>
  <owl:Class rdf:ID="Indicative">
    <rdfs:subClassOf rdf:resource="#LiuShu"/>
    <rdfs:label>指事</rdfs:label>
</owl:Class>
<owl:Class rdf:ID="IdeographicCompound">
    <rdfs:label>會意</rdfs:label>
    <rdfs:subClassOf>
      <owl:Class rdf:ID="LiuShu"/>
    </rdfs:subClassOf>
  </owl:Class>
<owl:Class rdf:ID="SemanticPhoneticCompound">
    <rdfs:label>形聲</rdfs:label>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#LiuShu"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="MutuallyInterpretive">
    <rdfs:subClassOf rdf:resource="#LiuShu"/>
    <rdfs:label>轉注</rdfs:label>
  </owl:Class>
<owl:Class rdf:ID="PhoneticLoan">
    <rdfs:label>假借</rdfs:label>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#LiuShu"/>
    </rdfs:subClassOf>
  </owl:Class>
```

Figure 3: The part of OWL-DL for Hantology

4. The Applications of Hantology

As a linguistic resource, there are many possible applications of Hantology:

(1) As an important resource for studying the development of Chinese lexicalization

In ancient Chinese, most words are only mono-syllabic and are represented by one character. In contrast, words with two syllables have dominated in modern Chinese. The words with two syllables need to be represented by two characters in Chinese writing system. The study of the development of Chinese lexicalization is critical to researchers. However, the progress of lexicalization is difficult to study without Chinese characters resources and proper computers application tools. Hantology can represent the character in a specific word is a morpheme or a word. First citation for each word is described in Hantology, so researchers can study Chinese lexicalization easier.

(2) As an important resource for natural language processing

One important feature of Chinese is that most words are compound words and the sense of Chinese words is correlated with the sense of each character represented in words. Therefore, Chinese characters are the basic knowledge to understand the semantic meaning of Chinese. Because the knowledge structure of Chinese characters has been properly represented in Hantology, it is an important resource to solve some problems of Chinese language processing. One critical problem in natural language processing is the unknown words problem. Because most words are compound words in Chinese, the meaning of unknown words can be predictable from characters represented in unknown words.

By using Hantology, the knowledge of characters represented in unknown word can be used by computers to decide the possible senses of unknown word. Hsieh also use the same idea to solve unknown words and ambiguous words problems by using HanziNet(Hsieh, 2005). However, Hantology provide more knowledge, so it is more useful for natural language processing. There are many variant words in Chinese. Variant words are the same words with different variant characters. Actually, some unknown words are variant words. If computer applications can detect the unknown words with variant characters, then unknown words can be processed more properly.

(3) As an important resource for studying the conceptual structure of Chinese and English lexicons

Hantology has the links between the concept of Chinese characters and SUMO (Standard Upper Merged Ontology). Hantology contains words generated from Chinese characters which are mapped into SUMO by Sinica BOW. It will be possible to study what concepts are derived from characters to words. The senses of characters on different periods are described in Hantology, we also can study the distribution of concepts represented in Chinese characters on different periods. The culture of different periods has the relations with the concept represented in usage of words. Hantology and Sinica BOW are resources for the researchers. Because WordNet also has been mapped into SUMO, the distribution of conception of Chinese and English lexicons can be compared and analyzed. These comparisons and studies mentioned in this paper are difficult to be made without Hantology.

(4) As an important resource for comparing different ideographic writing systems

There are other ancient writing systems that have the same features with Chinese. For instance, Cuneiform and Hieroglyphika characters also use semantic symbols or classifiers to represent the sense of characters. However, the conception system among different ideographic characters has not been studied. Using the approach adopted in Hantology, the conception system of Cuneiform and Hieroglyphika characters can be represented and compared each other.

5. Conclusions

Chinese has many different features. One important feature is Chinese use ideographic writing system instead of alphabetic or syllabic writing systems. For decades, we use the methods for alphabetic or syllabic writing systems to process Chinese characters in computer systems. How to represent Chinese characters properly in computer systems are ignored by researchers. That's why we must face the missing characters and information interchange problems for decades since computers were invented. The computers have strong impact on linguistics. The best evidence is the computational linguistics. However, researchers of Chinese characters still not get the same benefits from computers. There are only little databases and tools for Chinese characters. To solve these problems, we propose and create a linguistic resource named Hantology for Chinese characters. Hantology has been used successfully to solve the missing characters and Chinese information interchange problems. Besides, we propose four applications for natural language processing and studying.

Hantology has been studied for many years. However, it still needs to make extensions. The further extensions of Hantology are as following:

(1) Increasing the knowledge of Hantology

For instance, the pronunciations of Hantology are classified into ancient, middle-ancient and modern. Actually, the pronunciations of characters are varied in different place. This knowledge needs to be added into Hantology. In addition, the senses of characters depend on the context. The context is difficult to describe, so Hantology use cited sentences to express the context. To describe the context more precisely, the structure of context should be developed for Hantology.

(2) Mapping Hantology to MILO

In Hantology, the concepts of Chinese characters are mapped into SUMO. SUMO only contains about 1000 concepts. Therefore, it is not enough to distinguish the concepts expressed by different characters because there are more than 50000 Chinese characters. The MILO (MId-Level Ontology) which is the extension of SUMO has been developed. The MILO adds many new concepts into SUMO and makes it easier to integrate with other ontologies. By mapping Hantology to MILO, computer applications can distinguish the concepts expressed by characters much more easily than if only mapped into SUMO.

6. References

Chang, J.S., Lin, T., You, G.N., Chuang,T.C., and Hsieh, C.T.(2003)"Building A Chinese WordNet Via Class-Based Translation Model", *Computational Linguistics and Chinese Language Processing*, Vol.8, No.2, pp.61-76.

Chen, H.H., Lin, C.C. and Lin, W.C.(2002), "Building a Chinese-English Wordnet for Translingual Applications", *ACM Transactions on Asian Language Information Processing*,Vol.1, Iss.2, pp.103-122.

Chou, Y.M. and Huang, C.R. (2005), "Hantology: an Ontology based on Conventionalized Conceptualization, *Proceedings of Ontologies and Lexical Knowledge Bases*, Oct. 15.

Chou, Y.M. and Huang, C.R. (2005), "Analysis and Construction of Context in Chinese Variants", *ROCLING XVII: Conference on Computational Linguistics and Speech Processing*, Taiwan, Sep.15-16.

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

Farrar, S. and Langendoen, T.(2003) "A Linguistic Ontology for Semantic Web", *GLOT International*, vol.7,no.3, pp.97-100.

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*, Cambridge: MIT Press.

Hsieh,S.K. (2005) HanziNet: An Enriched Conceptual Network Of Chinese Characters", *6th Chinese Lexical Semantics Workshop*, April.20-24.

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.

Ishikawa, T. (1999) Mojikyo font database, http://www.mojikyo.org/html/abroad/index_e.html

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, January 28-30.

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

Yung, S.F.(2003) *The Retrieval System for ancient and modern pronunciations of Chinese Characters*, NSC 90-2411-H-002-039-A9, National Taiwan University.