Hanzi, Concept and Computation: A Preliminary Survey of Chinese Characters as a Knowledge Resource in NLP

von

Shu-Kai Hsieh

Philosophische Dissertation
angenommen von der Neuphilologischen Fakultät
der Universität Tübingen

am 06.02.2006

Tübingen

10.04.2006

Gedruckt mit Genehmigung der Neuphilologischen Fakultät
der Universität Tübingen

Hauptberichterstatter: Prof. Dr. Erhard W. Hinrichs
Mitberichterstatter: Prof. Dr. Eschbach-Szabo
Dekan: Prof. Dr. Joachim Knape

Hanzi, Concept and Computation: A Preliminary Survey of Chinese Characters as a Knowledge Resource in NLP

Shu-Kai Hsieh

Acknowledgements

There are many people to whom I owe a debt of thanks for their support, for the completion of my thesis and supported me in science as well in privacy during this time.

First, I would like to sincerely thank my advisor, Prof. Dr Erhard Hinrichs, under whose influence the work here was initiated during my fruitful stay in Germany. Without his continuous and invaluable support, this work could not have been completed. I would also like to thank Prof. Dr. Eschbach-Szabo for reading this thesis and offering constructive comments.

Besides my advisors, I am deeply grateful to the rest of my thesis committee: Frank Richter and Fritz Hamm, for their kindly support and interesting questions.

A special thanks goes to Lothar Lemnitzer, who proofread the thesis carefully and gave insightful comments.

I would like to thank my parents for their life-long love and support. Last but not least, I also owe a lot of thanks to my lovely wife Hsiao-Wen, my kids MoMo and NoNo for their understanding while I was away from home. Without them, it would not have been possible to complete the study.

Abstract

This thesis deals with Chinese characters (Hanzi): their key characteristics and how they could be used as a kind of knowledge resource in the (Chinese) NLP. Part 1 deals with basic issues. In Chapter 1, the motivation and the reasons for reconsidering the writing system will be presented, and a short introduction to Chinese and its writing system will be given in Chapter 2. Part 2 provides a critical review of the current, ongoing debate about Chinese characters. Chapter 3 outlines some important linguistic insights from the vantage point of indigenous scriptological and Western linguistic traditions, as well as a new theoretical framework in contemporary studies of Chinese characters. The focus of Chapter 4 concerns the search for appropriate mathematical descriptions with regard to the systematic knowledge information hidden in characters. The subject matter of mathematical formalization of the shape structure of Chinese characters is depicted as well. Part 3 illustrates the representation issues. Chapter 5 addresses the design and construction of the HanziNet, an enriched conceptual network of Chinese characters. Topics that are covered in this chapter include the ideas, architecture, methods and ontology design. In Part 4, a case study based on the above mentioned ideas will be launched. Chapter 6 presents an experiment exploring the character-triggered semantic class of Chinese unknown words. Finally, Chapter 7 summarizes the major findings of this thesis. Next, it depicts some potential avenues in the future, and assesses the theoretical implications of these findings for computational linguistic theory.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Motivation

The story begins with a *tree*.
In 1997, I worked as a research assistant with the CKIP (Chinese Knowledge Information Processing) group at the Academia Sinica in Taipei. At that time, one of our tasks was to correct sentences extracted from a newly founded Chinese Treebank. My eyes climbed up and down the sentence trees labelled with syntactical tags, oftentimes descenting to the leaf nodes (well, the terminal symbols), I saw Chinese characters there *on the fly*. In our reports of formal syntactic (be it LFG or HPSG) analysis, they appeared everywhere, but were just like a waste dump that draws no one's attention. (Even worse was that in addition to translation, we had to transliterate each of them so that people could also read them aloud at the conference). The experience spurred me on to rethink about the meaning and usefulness of Chinese characters, particularly in the context of *language*, or more generally, *cognitive processing*.

Starting from that, this thesis presents a primary result. Briefly, what I would like to deal with in my thesis can be stated as follows: to reestimate the meaning of Chinese writing in linguistics, and, on the other hand, to propose that Chinese characters (Hanzi)[1] could be used as a useful knowledge resource

[1]Since some of the East Asian scripts like Japanese, Korean, and Vietnamese writings are all traceable in one way or another to Chinese characters as their source, I will use the term "Hanzi" instead of Chinese characters in this thesis, when their neutrality and if

in Chinese language processing.

1.1 Does Writing Matter?

Writing, the symbolic representation of language in graphic form, has not been the main concern among linguists. The neglect of the written language could be attributed to many reasons. One of these reasons might be due to the prevalent "surrogational model" in contemporary linguistics, which states, writing is to be viewed as a surrogate or substitute for speech, the latter being the primary vehicle for human communication (Harris 2000). Such an attitude has paved the way for the opinion that the written language is somehow inferior to the spoken language, therefore not warranting the serious attention of linguists.

As Coulmas (2003) noted, most scholars in language science do not believe that the invention or discovery of writing makes a difference, either with respect to what language is, or how we think about it. Until now, linguistic orthodoxy has still concurred with Ferdinand de Saussure's apodictic statement that made Aristotelian surrogationalism a cornerstone of modern linguistics:

> Language and writing are two distinct systems of signs: the second exists for the sole purpose of representing the first. The linguistic object is not both the written and the spoken forms of words; the spoken forms alone constitute the object (Saussure 1959:23).

However, in the recent book titled "Rethinking Writing", Harris (2000) offers an alternative reinterpretation of Saussure's view: the blind spot in traditional Western thinking about writing (and other forms of communication) is a failure to grasp the systematicity involved. Writing, for Saussure, was not just an *ad hoc* appendage to speech. Writing systems are *systems* in their own right, even though they subserve or supplement forms of oral

possible, universality are emphasized.

communication.[2] For Saussure, the beginning of wisdom here is grasping the (semiological) fact that writing is not just a copy or mirror image of the speech system, *and cannot be.*

Another issue concerning writing in linguistics is the assumption of the *linearity* of the linguistic sign. This can be traced back to another famous notion of Saussure, that is, that language is comprised of *signs*, which embody a *signifier* (sound) and a *signified* (meaning), and these signs are *linear*, namely, one unit follows another in sequence. The assumption of the *linearity of the linguistic sign* implies another assumption that linguistic forms obey nothing but the principle of simple concatenation of a chain of temporally successive elements.

As obvious manifested in the development of modern linguistics, ever since Saussure's above-quoted postulate, the primacy of speech and the linearity of writing are taken for granted in linguistic research and theory formation, In fact it indeed works reasonably well if the linguistic form is a spoken form. Aronoff (1992) even points out that, like Saussure, Edward Sapir, Noam Chomsky and Morris Halle appeal to alphabet-based writing in successfully developing their phonological theories.

However, on the other hand, we also see a quite different picture in the Eastern tradition of the scientific study of language. Geoffrey Sampson (1985) expressed some personal feelings, which turns out to be somewhat accurate: 'The axiom of Western linguistics according to which a language is primarily a system of spoken form, and writing is a subsidiary medium

[2]In the history of linguistics, the Copenhagen school of *glossematics* shares similar opinions. The founder of glossematics, Louis Hjelmslev, once held that [...] linguistic units are independent of their expression in speech, writing, or any other material form [...] thus the system is independent of the specific substance in which it is expressed.(Siertsema 1965). In this respect, we can also see the parity of writing with speech. The system of speech and the system of writing are [..] only two realizations of an infinite number of possible systems, of which no one can be said to be more fundamental than any other. The contemporary echo, Harris's integrational linguistics, takes a position opposed to the orthodox position on all possible counts with regard to the status and study of writing as well. Interested readers are referred to Harris's relevant works.

serving to render spoken language visible, is very difficult for an East Asian to accept.' An interesting and illustrative example worth mentioning here is that, even in the modern *Encyclopedic Dictionary of Chinese Linguistics*, the first topic is the writing system, and is treated at great length.

Therefore, a much closer look at writing, in terms of Chinese characters, seems urgently needed, if we are to reach more global conclusions about the nature of human writing.

1.2 Knowledge-Leanness: A Bottleneck in Natural Language Processing?

Let's take a look at the current status from a bird's eye view. Natural Language Processing (NLP, also called computational linguistics) has been a discipline for more than 50 years. This field has grown since the 1960's as a fertile combination of ideas from areas such as cognitive science, artificial intelligence, psychology and linguistics. At the beginning, things seemed to be easy. It became evident that the problem of natural language understanding was much harder than people had anticipated, until many kinds of programs had been written to process natural language. As of the present, one of the main obstacles still get remained: the phenomena of massive ambiguity (both syntactic and semantic). In addition, it then became clear to be grasped that, understanding natural language requires a large amount of *linguistic* and *general knowledge* about the world, and, of course, the ability to reason with *it*. Acquiring and encoding all of these knowledge resources is one of the fundamental impediments in developing effective and robust natural language processing systems.[3]

Over the past years, the NLP researchers have uncovered a great deal of *linguistic knowledge* in different linguistic areas such as syntax, semantics,

[3]Some computer scientists such as Rober Wilensky have even described NLP as an "*AI-complete*" problem, which means that if we would like to be able to solve NLP problems, we have to solve a large fraction of AI problems first.

pragmatics, and many formal models have been proposed as a result. However, natural language processing systems still suffer in most cases from the lack of combining extra-linguistic knowledge with the linguistic knowledge in an integrated way. The richness of natural language causes difficulties for researchers attempting to build manually a full-fledged system capable of handling anything close to the full range of phenomena. As a result, most NLP systems have been constructed to function only in limited domains.[4]

Until recent years, presumably simulated partly by the availability of large machine-readable text corpora, the use of statistical techniques has entered the scene, and probabilistic and data-driven models soon became a quite standard paradigm throughout current NLP technology. Instead of getting knowledge into the computer *en masse* by building complex rules manually, many machine learning algorithms summoned have begun to be statistical in the sense that they involve the notion of probability and/or other concepts from statistical theory.

There is a line of poem among the fragments of the ancient Greek poet Archilochus which says: "The fox knows many things, but the hedgehog knows one big thing". In a figurative way, Pereira (2000) compared the hedgehog and the foxes as statistical machine learning methods and structural knowledge descriptions separately.[5] Statistical techniques seem to be an cutting tool for doing linguistic research with surprising *accuracy*. It is, surely, a natural tendency for a few computational linguists to become overtly attached to the probabilistic model, and expect it to be the sole methodological consideration in all the language processing problems. Nevertheless, to yield natural language *understanding*, as Pereira put it, we have the strong intuition that we require *something else*.

For example, most statistical models have something limited in common.

[4]This especially holds for machine translation where so-called *niche applications* (where the focus is on a specific application) have become more and more important.

[5]Fernando Pereira, 2000. The Hedgehog and the Fox: Language Technology and the Knowledge of Language. Invited talk at the COLING 2000, Saarbrücken.

First, they share the need for relatively large amounts of training data, especially for the supervised methods, the "right" answer must be part of the training data.[6] Secondly, they are all knowledge-poor, in the sense that they require no *real world knowledge* - what might be termed *common sense knowledge* - for their implementation. However, these knowledge-poor supervised training regimes could be a significant bottleneck in the development of practically robust NLP systems, if they aim at dealing with real-world applications.[7]

The open question which Pereira posed at the end of the talk is worth pondering: "Can hedgehogs evolve into foxes" ? In the foreseeable future, does the answer seem pessimistic? Or, as the optimist concludes, are the different paradigms slowly merging ?

Vossen (2003) observes that there is a tendency in NLP to move from morphosyntactical modelling and applications to semantics, where well known statistical techniques can often be easily combined with semantic data. NLP is thus moving towards inferencing systems that exploit common-sense knowledge. We may take the current state of research for automatic anaphora resolution as an illustration of this. As Mitkov (2003)[8] mentioned, up to now, the results from experiments concerning automatic anaphora resolution strengthened with statistical techniques, are still very discouraging. But, as he predicted, the ways forward could be knowledge rich: the exploitation of different (linguistic) knowledge resources to enhance anaphora resolution,

[6]There is, of course, another learning scheme called "unsupervised" learning, which does not presume the existence of set of classification. However, Manning and Schütze (1999) pointed out that this term is not always clear out in the field of statistical NLP, and is not always used in the same way in the machine learning literature, so we'll skip it for now.

[7]Take as example from Lenat (2002): [...] while Mycin can be considered a medical diagnosis system, really what it does is decide which of five kinds of meningitis you are most likely to have. It does that better than most GPs. [...] However, if you ask it to help treat a broken bicycle it will tell you what kind of meningitis it is most likely to have.

[8]Ruslan Mitkov. (2003). A Final Word. In: *Crash Course on Anaphora Resolution*, Tübingen.

including annotated corpora, bilingual (and multilingual) corpora, and even ontologies (e.g., WordNet), and the most promising way forward might rest on encoding (or retrieving) real world-knowledge.

But, the question is, *how*? The *knowledge acquisition bottleneck* is notorious. It takes a long time to get the knowledge even from the heads of children into a machine-readable form. There have been a few attempts, mostly inspired by artificial intelligence (AI), to represent *world knowledge* and connect this knowledge source with linguistic organizations by means of specialized interfaces. Nevertheless, there is an enormous widely-recognized tradeoff between knowledge-poor and knowledge-rich approaches: Do we really want to build a time-consuming and labour-intensive knowledge base (both linguistic and ontological), in order to enhance the performance of real-world NLP systems? [9]

The dream of building a realistic natural language understanding system is still there, what could be the next step? Accumulated and revised over thousands of years, Chinese characters retain a unique common sense structure, which has proven effective over many generations. The core topic of this thesis is thus: Could *conceptual* and *ontological* knowledge naturally "*grounded*" in writing (at least in ideographic one like Chinese) be useful and low-cost in this context?

1.3 Writing Systems: The Missing Corner?

The above thinking leads us to further consider *what* and *how* writing systems can do for NLP?

Occupying the main stage of modern linguistics, the surrogational model goes easily hand in hand with the notion that the basic function of the signs

[9]Lenat (2002), the founder of Cyc - the world's largest common-sense knowledge base -, estimates that the work on building up a common sense knowledge base done in the present decade will take about 250 man years of efforts. http://www.leaderu.com/truth/2truth07.html

used in writing is *phonoptic*,[10] i.e., serving to make sound "visible". Writing is thus just regarded as an ingenious technical device for representing the spoken language. In this perspective, writing is valued chiefly for the ways it offers for replacing speech in dealing with and disseminating the kinds of information that are regardless important (Harris 2000).

It is thus not strange that the current and possible role of writing systems in modern (computational) linguistic theory and practice has been explored only within a limited domain, such as Optical Character Recognition (OCR) and Text to Speech (TTS). Until now, writing systems in general are not a main concern in the field of computational linguistics.[11]

In this light, we are glad to share the same basic position of the *integrational linguistics* proposed by Harris, presuming that writing systems are *systems* in their own right, even though they subserve or supplement forms of oral communication.

However, how can we grasp the *systematicity* involved in writing systems exactly? One possible answer is that *knowledge* representation is employed in a natural language. In his recent book, Sowa (2000:168) mentioned that natural languages could be the ultimate knowledge representation languages. More and more approaches focus on the relation between natural language and knowledge representation. A potential perspective is: Natural language *itself* can be treated as a knowledge representation and reasoning system, not just as an interface.[12]

But, the reason that a satisfying knowledge representation language has not yet appeared, lies perhaps in the historical-cultural background of language evolution. After thousands of years of evolution in intimate contact with every aspect of human experience, natural languages have attained a greater flexibility and expressive power than any artificial language, but these

[10]This term is used to contrast to *optophonic* devices that converts lights into sound, rendering what was visible audible.

[11]Sproat's recent book (Sproat 2000) is a new attempt.

[12]Wong (2004) also shares similar ideas.

"knowledge experiences" are now difficult to trace, especially for languages with phonetical scripts.

In this context, if we turn to the only "ideographic" script, - one of basic types of writing systems in the world, - namely, the Chinese writing system, we could see that it still displays a considerable amount of semantic information at the character level. Chinese characters have survived over thousands of years; some proposed that the whole set of Chinese characters can be viewed as an encyclopedia in essence, or in terms of knowledge representation, as a kind of ontological knowledge. This distinctive feature might also suggest that the system of Chinese characters might contain rich but concise system of inter-related concepts.

Another point I wuld like to mention here is that the relationships between knowledge and written language, in particular the dependencies between conceptual categories and linguistic expressions in terms of Hanzi, have been and will be the subject of much psychological and philosophical debate. And there have been several different argumentations that have been considered for that. To meet the need of computational intelligence, I world take a rather *pragmatic* stand on this issue, which will be referred to throughout this paper.

In summary, this thesis attempts to lay the foundations for a new field of Chinese character-based NLP studies, a field which might be called *computational hanziology*, the formal study of the characters of the Chinese language. Inspired by the abundant conceptual and semantic information frozen in the characters, the goal of this thesis is to achieve a theoretical synthesis of computational theory of Hanzi, with the following questions in mind: In what sense could we regard Chinese characters as a kind of knowledge resource? and how to represent this knowledge resource and how to make it operate in NLP systems? We believe that research in the field of computational hanziology might contribute to the finding of solutions to main problems currently plaguing in computational linguistics.

Given this goal, my expository strategy will be laid out as follows: First, for readers who are not familiar with the Chinese language and Chinese writing, we introduce the basic notions in Chapter 2. As background knowledge, Chapter 3 gives an overview of related work in Hanzi studies, both from traditional and contemporary viewpoints. Emphasis will be placed on a recently-proposed specific theorectical framework concerning with a Hanzi-triggered conceptual modeling. Chapter 4 gives discussions of mathematical models around Chinese characters. Chapter 5 introduces an implementation system. Chapter 6 presents a NLP experiment based on the theory and system proposed. Finally, Chapter 7 provides an concluding remark on the extensibility of the approach, as well as an overview of potential research directions in the future.

Chapter 2

A Brief Introduction of Chinese and its Characters

Before embarking on the theme of this thesis, this chapter outlines the desiderata of basic concepts of Chinese characters and their relation to Chinese. The aim is to provide readers with enough information which will serve as the backdrop to our discussion later on. In the following, a brief description of the Chinese characters is given in Section 2.1, structural descriptions of them and some special linguistic issues involved are summarized in Section 2.2.

2.1 What is Hanzi?

2.1.1 An Overview

Chinese writing has no alphabet. Instead, it employs Hanzi (read as hànzì, written as 漢字, literal meaning: Han-characters), which are named after the Han culture to whom it is largely attributed.[1]

Historically, Hanzi dates back to the late Shang Dynasty (about 1401-1122 BC). At that time, they were inscribed marks on tortoise plastrons

[1]They have been called characters, pictographs, pictograms, ideograms, ideographs, logograms, logographs, glyphs etc., based on different consideration of their nature. In this thesis, they shall be called "Hanzi" or "Chinese characters" interchangeably.

(the underside of the shell or armour) and ox scapulae (shoulder blades) – "oracle bones". Chinese writing has been in continuous use for well over three thousand years, though Hanzi forms changed with time,[2] from the point of view of the writing system as a whole, there have been no basic qualitative changes.

Early Chinese characters were mainly *symbols* and *pictographs* that represented some abstract concepts of daily life. In order to express more complex ideas and concepts, *pictographs* were developed and combined to form *ideographs* for multiple meanings. Today, these *ideographs* [3] form about 90% of the total Chinese characters in current usage (Ren et al 2001).

Chinese characters are written in a two-dimensional (2-D) quadrilateral format, which is why they are sometimes called 方塊字 (/fāng-kuài-zì/, 'characters in virtual square boxes'). In the following, we will introduce some of their major features: *various topological structures*, *a large amount of signary* and *an easy means of communication.*

The *topological structure* of a character means that the character is a combination of various components that can be shown in Figure 2.1 (Yiu and Wong 2003). The same component may appear in different characters, and may be located at different positions.

Chinese characters make up a large *signary*. The *"complete"* number of Chinese characters has grown tremendously over the millennia. Shang Dynasty inscriptions boast lightly more than 2,500 characters; Xu's classical work, a first large-scaled character dictionary, contains 9,353 characters; And 漢語大字典(/hànyǔ dà zídiǎn/, 'The Great Chinese Language Dictionary'),

[2]Throughout this paper, I will mainly focus on traditional character forms currently used in Taiwan and Hong kong. As for their variants used in Japan, Korea and Vietnam, and the simplified forms used in mainland China, these will be considered only when necessary.

[3]Some issues concerning the *ideographic* property of Chinese characters are controversial, these will be discussed in detail in the next chapter. The term *ideograph* or *ideogram* can be understood here as a symbol used in a writing system to represent an idea or a thing, not a particular word or phrase.

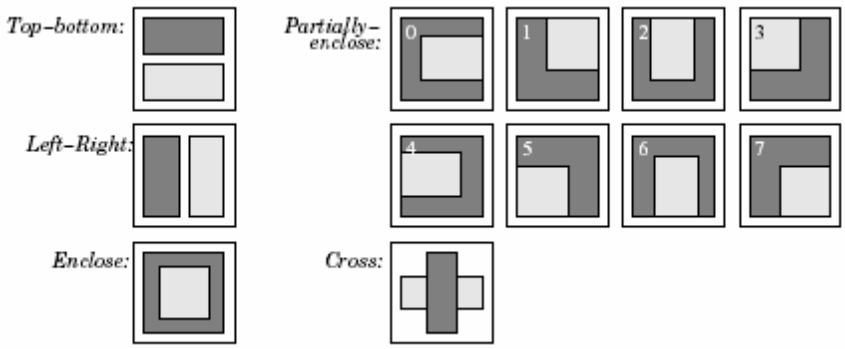


Figure 2.1: Some topological structures of Hanzi (adopted from Yiu and Wong (2003))

published in 1986, comprises 56,000 characters. A historical comparison of the Chinese signary is shown in Table 2.1. It is noted that the *actual* comprehensive number of Chinese characters can only be estimated approximately.

As for the Chinese characters used in common *vocabulary* today, different statistical studies have demonstrated different results. In general, 3,000 characters have been defined for daily use, 7,000 characters have been determined to be necessary in writing, and there is a total of 60,000 characters which include complex and simplified styles. Interestingly, most characters listed in a Chinese dictionary are rarely used.[4]

It is an unfortunate fact that all commonly used Chinese character encoding schemes assign character codes to only a limited number of characters. Although Unicode was designed as an attempt to represent "all" characters in the world, its two-byte form can represent at most 65,536 characters. In mainland China, two character sets, containing 3,755 characters and 6,763 characters, respectively, were announced as the National Standard GB2312-80 (the first is a subset of the second one). In Taiwan, 5,401 characters are

[4]Some researchers estimate that the vocabulary of Chinese characters is roughly equivalent to Western words in total. See Govindan and Shivaprasad (1990), *Character recognition: A review*. In *Pattern Recognition* Vol.23, No.7.

Table 2.1: Chinese signary: A historical comparison

Name of Dictionary	Year of Publication	Number of Characters
Shuo Wen Jie Zi 說文解字	ca.100 A.D.	9,353
Zi Lin 字林	4th C, A.D.	12,824
Yu Pian 玉篇	543 A.D.	250
Guang Yun 廣韻	1008 A.D.	26,000
Lei Pian 類篇	1039 A.D.	31,000
Ji Yun 集韻	1067 A.D.	201
Zi Hui 字匯	1615 A.D.	33,179
Kangxi Zidian 康熙字典	1716 A.D.	47,035
Zhongwen Da Cidian 中文大辭典	1968 A.D.	49,905
Hanyu Da Zidian 漢語大字典	1986 A.D.	56,000

Table 2.2: How many Hanzi does a computer recognize?: A code scheme comparison

Character Set	Num. of Characters
GB2312-80 (China)	6,763
BIG-5 (Taiwan)	5,401
JIS-2 (Japan)	3,390
Unicode	65,536

included in a standard set called BIG-5. In Japan, 3,390 characters are included in the JIS level-2 standard. Table 2.2 shows a current code scheme comparison.

As will be discussed, Chinese characters function differently than a purely phonetic script mainly in that they the former carry *stronger* semantic weight in and of themselves. Such graph-semantic feature makes efficient *communication* possible between people who speak different languages (or dialects) within- and outside of China. In the so-called *Hanzi Cultural Circle* which includes Japan, Korea and Vietnam, Hanzi were adopted and integrated into their languages and became Kanji, Hanja, and ChuHan respectively. Today, Japan and South Korea still use Hanzi as an integral part of their writing systems.

It is widely believed that the trend in the evolution of human scripts is

a process which began with the pictogram, evolving through the ideograms to the phonogram. The writing system of Chinese, however, constitutes the only exception to this statement, and is the only ideographic script still in use today. It is at the same time the oldest one in use (around 3400 years).[5]

2.1.2 The Relation Between Chinese and Hanzi

We will now introduce some specific features of Chinese which are quite relevant to an understanding of how Hanzi works.[6].

- *Abundant in homophones*

Chinese is a language with a small number of syllables which are clearly demarcated from one another phonologically. Mandarin Chinese, for example, which is the official language in the Chinese world, distinguishes only 411 different syllables, each of which may theorectically have five tones at most. This amounts to no more than 1,284 actually used syllables (Wang 1983), while various dictionaries put this number between 888 and 1,040 (Coulmas 2003:57). Such (relative) phonetic poverty can therefore yield an extraordinary number of *homophones*, which has an ambiguous effect on the spoken language. It is commonly believed that the pervasive homophones constitute the *main* reason for the retention of ideographic script in Chinese.

- *Tripartite articulation*

In general, 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*. Each morpheme is exactly one syllable long, so there are no cases such as *vorstellen* or *Solidalität* in German, where a

[5]The other famous one was the Egyptian system of hieroglyphs, which supported an astounding civilization for 3600 years, but vanished about 1800 years ago.

[6]Most of this section is based on Sampson (1985:147)

single meaning-unit spans more than one syllable, or *ausnehmen*, where the /aus/ is a meaning-unit corresponds to only a fraction of a syllable. For the sake of simplicity, the following formula might be clear at a glance:

$$\text{Hanzi} \approx \text{Syllable} \approx \text{Morpheme}$$

For example, the word 路 is a character that represents both the meaning "road" and the syllable /lù/; and the word 徜徉 is a word made up of two characters, where each character represents one syllable, /chǎng/ and /yáng/ respectively, and each character contributes to the compositional meaning of "to roam aimlessly".[7]

- *Morphological isolating*

From the comparison of morphology across languages, (ancient) Chinese has traditionally been considered to be a member of the family of *isolating* languages in which every word consists of a single morpheme. Syntactically, Chinese does not have any noun declination or verbal conjugation. In many cases, it is difficult to clearly differentiate between word compounds and syntactic phrases.[8] However, as will be seen in the following section, there is a new trend in morphology of modern Chinese.

- *Trend of disyllabification*

In modern Mandarin Chinese, there is a strong tendency toward dissyllabic words, while the predominant monosyllabic words in ancient Chinese remain more or less a closed set.

This tendency could be explained in view of some historical linguistic viewpoints, for since the Yuan Dynasty (1206 AD), Chinese has gradually

[7]It is noted that though Chinese writing is primarily syllabic, it is not a syllabic writing system due to the fact that most characters posses semantic radicals, (sense identifier). Thus, some propose that Chinese is best described as a "morpheme-syllabic writing system" in which radicals and phonetic parts serve mutually diacritical functions.

[8]It is impossible to see or to hear if a word is a noun, a verb or an adjective, and in ancient Chinese most words could be used as noun, verb, adjective or adverb.

lost its consonantal endings like -t,-k,-p,-m, retaining only -n and -ng.[9] This has greatly reduced the number of monosyllables that can be used while, on the other hand, new characters had to be created to cope with the more and more sophisticated requirements of verbal communication. This state of affairs inevitably brought about an increase in homophonic clashes. For the language to find a way out of this dilemma, *disyllabification* has naturally become the device needed to resolve these homophonic clashes.

Once this dissyllabic tendency set in, it greatly influenced the development of the Chinese lexicon:

(1). This tendency causes disyllabification not only of monosyllables but also of polysyllabic constructions. For instance, a quadrisyllabic nominal phrase like 國家安全 /*guó jīa ān quán*/ "national security", in which *guó jīa* means 'nation', and *ān quán* means 'security, safe'; It is shortened to a disyllabic compound word 國安 /guó ān/, taking the first and the third character of the original structure. Similarly, a quadrisyllabic verbal phrase like 互相勉勵 /*hù xiāng miǎn lì*/ "mutually encourage", in which /*hù xiāng*/ means "mutually", and /*miǎn lì*/ "encourge". It is abbreviated as a disyllabic noun 互勉.[10] Thus condensation works hand in hand with expansion to disyllabify every possible lexical structure that comes its way (Yip 2000).

(2). This trend also plays a dynamic role in creating new words in modern Chinese. Unlike the monosyllabically oriented ancient Chinese, the increase in words is reflected in the increase in written symbols, whereas in modern Chinese, the increase in words corresponds directly to the increase in disyllabic combinations. For example, the 非典型肺炎 (atypical pneumonia) in China is often expressed as 非典 in newspaper and other media.

[9]The following descriptions are extracted from Yip (2000).

[10]Note that polysyllabic constructions with more than 4 syllables are also often shortened to trisyllabic constructions. for example, 資訊科學系 (department of information science) condenses to 資科系. As for which characters should be retained for the new structure, this is an interesting theme to be explored.

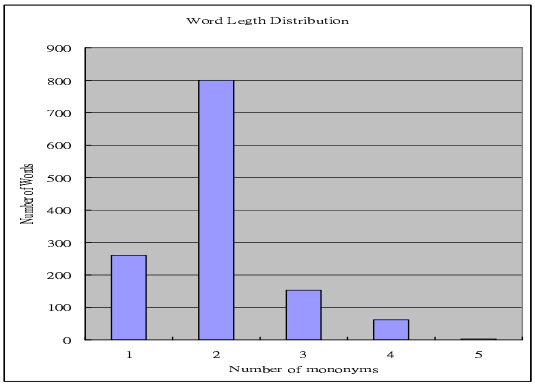


Figure 2.2: The word length distribution of Chinese characters

One set of statistics[11] shows the proportion of different polysyllabic structures (Figure 2.2). This diagram should give some idea of disyllabic predominance both in actual usage and in the lexicon proper.

- *Relative flexible semantic constraints in compounding*

In practice, we often find that a single word in a European language often translates into modern Chinese as a sequence of two *morphemes*, that is, as two characters. However, it is difficult to identify these unambiguously as single compound words akin to English examples like *blackbird, interview,* or *overthrow*, because the borderline between *morpheme* combinations is vaguer for Chinese than it is for English. To put it simply, a Chinese speaker is relatively free to group *morphemes* into different combinations.

In many cases, a Chinese speaker sees these *morphemes* as the blocks of *conceptual combination* which the language-system supplies, and utilizes these building blocks to express things and ideas even within the domain of individual language-use. Of course, this should not be exaggerated, for there certainly are very many cases where a particular compound word of two *morphemes* is strictly constrained within its own fixed and idiosyncratic semantics; but in relative terms, we may say that Chinese morphemes have

[11]Yi, Xiwu (1954:10-11)

more freedom of combination than do the morphemes of English or other European languages.

Since *morphemes* are relatively free to combine with one another grammatically, the result is that for Chinese there is no very clear-out notion of "word" as a unit larger than the *morpheme*. This constitutes one of the main linguistic issues which will be discussed in next section.

2.2 Character Structure Units and Linguistic Issues

This section introduces the structural units of Chinese characters and one of the most fundamental linguistic issues in Chinese computational linguistics.

2.2.1 Constituent Units of Chinese Characters

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, 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 hierarchy: *character, component and stroke*.

- Character

  A character can be seen as a pattern with the appearance of a rectangle or square. Its appearance is basically determined by the shape of basic strokes and their combination.

  Roughly, characters can be divided into two categories: 獨體字(/dú tǐ zì/, 'independent Hanzi') and 合體字(hé tǐ zì, 'combined Hanzi'). The only difference lies in the fact that the latter is composed of at least two components, while the former stands alone as a complete unit. For instance, in Figure 2.3, 語(/yǔ/, 'language') is a combined Hanzi, because it consists of two components: 言(/yán/, 'speech') and 吾(/ú/,

'we'). The component 言, as an independent Hanzi, has similar meanings but is clearly distinguished from the meaning of the Hanzi 語.

According to the investigation of characters in common use (Scurfield 1991), 90 % of characters belong to the *combined Hanzi*, while only the remnant 10% as independent characters. Most of these 10% of characters come from original characters (*pictographs*) in the historical development of Chinese characters. Some of the original characters continue to be used as individual characters, while some of them are now only used as components of a character.

- Component

  A component can be regarded as a minimal *meaningful* unit within a character. A component of a character may be a character itself, or any structural part of another character.

  In the traditional view, the overwhelming majority of characters contain two kinds of components: a *phonetic* component and a *signific* component (also termed "radicals"). The former indicates with fair accuracy the syllable of Chinese for which the characters stand, while the latter identifies a broad semantic category of the character in question. Radicals have always been used as a lexicographic ordering principle, and are used in almost all indexes to facilitate finding characters in *dictionaries*.

  A limited number of components, provides the stock for forming a potentially infinite number of Chinese characters. This has led some scholars to believe that such a system of radicals contains not only clues regarding the evolution of characters, their pronunciation and meaning, but also the secret to the logical structure that they presume underlies the language as a whole (Porter 2001).

  It is worth mentioning here that there has not been a commonly ac-

Table 2.3: Number of radicals: A comparison

dictionary name	num. of radicals
說文解字 (An Analysis and Explanation of Characters)	540
康熙字典 (KāngXī Dictionary)	214
辭海 (Sea of Words)	250
新華字典 (New Chinese Dictionary)	180
漢語大字典 (Great Chinese Character Dictionary)	201
漢語大辭典 (Great Chinese Word Dictionary)	201

cepted idea of how many components there are and how to classify them semantically. Table 2.3 shows a comparison of the number of radical components in different dictionaries.

According to the Cang-Jie Theory which will be elaborated on in Chapter 3, the component layer consists of two parts: 字首 (character head component) and 字身 (character body component). A character head component, similar to a radical, originates from pictographs in Chinese characters. This component provides a major semantic category and occupies an independent position in a character. A character body component, like a signific component, gives a phonetic clue; but it also contributes to the refinement of meaning of the character.

- Stroke

  No matter what kinds of components, they can be ultimately further simplified to simple combinations of basic strokes. A stroke is defined as a dot or a continuous line confined to a square-shaped area, such as 一 (horizontal stroke), 丿 (falling to the left), 乙 (narrow angle open right), 乛 (double hook) and so on. In some literature, a stroke is defined as a *grapheme*, the minimal graphic unit of Chinese characters, for the structure of their shape can be defined entirely by these primitive strokes. All characters are composed of a combinaation of eight basic strokes (though calligraphers note as many as 64). The general rules of

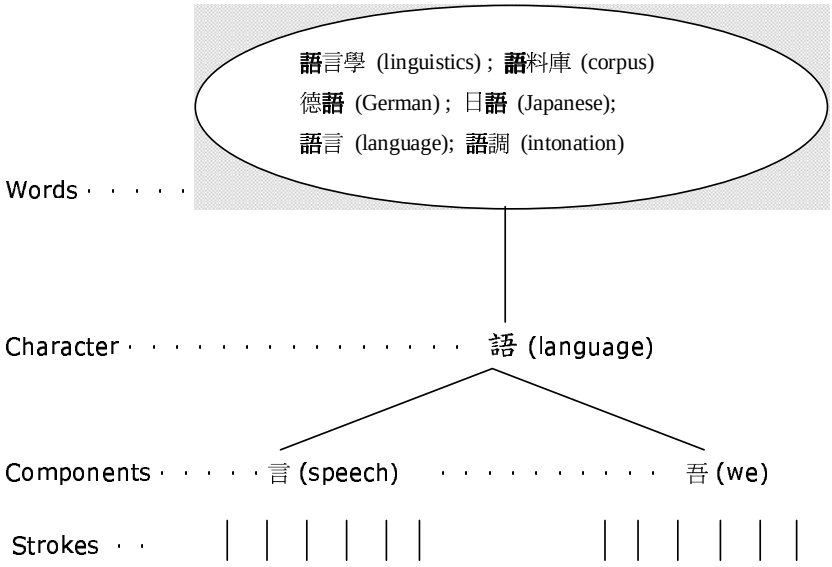


Figure 2.3: A three-layer hierarchy of the Hanzi lexicon structure

stroke orders are from top to bottom, from left to right and from the outside to the inside.

Figure 2.3 shows an example of the 3-layer structure of Hanzi.[12] The character 語 can be divided into two components: 言 and 吾, which can be further decomposed into basic strokes, respectively. And the character 語 (language) can become one part of various semantically related *words* such as 語言學 (linguistics), 德語 (German), 語調 (intonation) and so on.

2.2.2 Word/Morpheme Controversies in Chinese

In this section we introduce two crucial linguistic issues: the definition of the terms *word* and *morpheme* in Chinese.

Having a clear definition of what a word is seems like a prerequisite for all kinds of linguistic study. However, the question "What is a word?" in Chinese still has no definition which is universally acceptable at present.[13]

[12]For technical reasons, the strokes are not demonstrated.

[13]As Packard (1997) noted, there was no term in Chinese for "word" as distinct from "character" until the beginning of the twentieth century. The current expression for

For speakers of English, a typical *word* has a pronunciation, a spelling, a meaning, and a "part of speech", which can be found in the "dictionary". Words in written texts are bounded by spaces or punctuation on either side. On the contrary, it is somewhat different in the case of Chinese. The writing system is non-alphabetic, written syllable by syllable, and it is not easy to identify a word in the text due to the fact that there are no separators (like *spaces* in written English texts).[14] For example, is 未婚 /wèi hūn/ (lit: not-married; "single") a word with two characters, or two words with one character respectively? Should 土地公有政策 /tǔ dì gōng yǒu zhèng cè/ be separated into 土地 (/tǔ dì, 'ground') 公有 (/gōng yǒu/, 'public-owned') 政策 (/zhèng cè/, 'policy'), or 土地公 (/tǔ dì gōng/, 'Kobold') 有 (/yǒu/,'has') 政策 (/zhèng cè/, 'policy') or something else entirely?

Even native speakers of Chinese disagree on what a word is. There are multiple studies[15] (Wu and Fung, 1994; Sproat et al., 1996; Luo and Roukos, 1996) showing that the agreement between two (untrained) native speakers is about upper 70% to lower 80%. The agreement between multiple human subjects is even lower (Wu and Fung, 1994).[16] Proper names, number, measure units and compound words constitute the main factors that human subjects differ in word segmentation, although these ambiguities do not change a human being's understanding of a sentence. In the area of NLP, such low agreement among human judges affects directly the evaluation of machines' performance as it is hard to define a gold standard, which leads to the well-known knotty *word segmentation issue* in some Asian languages.[17]

"word", namely, 詞(cí), is a learned term, used mostly only in linguistics.

[14]Theoretically, for a given sentence $C_1C_2C_3...C_n$, there are 2^{n-1} possibilities of segmentation, where n stands for the number of characters.

[15]These are quoted from Luo (2003).

[16]Many psycholinguistic studies have also reported that there is disagreement among the word-marking responses. See Hoosain (1991).

[17]Currently, there are three proposed segmentation standards widely adopted in Chinese NLP: the Mainland Standard of China (GB-T 13715-92, 1993), the ROCLING Standard of Taiwan (Huang et al., 1996), and the University of Pennsylvania-Chinese Treebank (Xia, 1999).

There have been different definitions of words from the viewpoints of theoretical linguistics definitions, the most complete recent linguistic discussion of this topic is given by Packard (2000), where he discusses the notions of the Chinese word from orthographic, sociological, semantic, syntactic, psycholiguistic perspectives ... and so on. A detailed review is beyond the scope of this chapter. Generally, this thesis inclines to agree with Sproat's review in that the linguistic theoretical approach proposed tends to be more principled at an abstract level, but harder to pin down when it comes to specifics.[18] In addition, Sproat's pragmatic "moral" in doing Chinese NLP is also taken here: "There *correct* segmentation depends on the intended purpose of the segmentation."

The issues discussed above are indirectly related to another fundamental problem: What is a *morpheme* in Chinese? Again, it is not as clear-cut for Chinese as it is for European languages.

By defining morphemes as the smallest linguistic units in language to which a relatively stable meaning may be assigned, most linguists find it easy to think that Chinese words are made up of one or more morphemes, and the common acceptable notion of the Chinese morpheme is further defined as a single syllable, or a single character. As introduced previously, a morpheme in Chinese is something that is written with a single *character*, and pronounced as a single syllable. At first glance, this seems to be true, but such a position comes up against some intractable problems which, although not central, deserve to be discussed:

1. *Disyllabic morphemes*

   Sproat (2000) extracts a list of disyllabic morphemes that occur more than once from a 20 million character corpus, such as 踉蹌 /làngqiāng/ (hobble), 躊躇 /chóuchú/ (hesitate). These pair of characters can only

[18] See Sproat and Shih (2001); Sproat: Review of Packard. (2001). In LINGUIST LIST 12.897).

co-occur with each other, that is, they are disyllabic morphemes which therefore violate the notion of the "monosyllabic morpheme".

2. *Borrowed polysyllabic morphemes*

   Another problem raised by Sproat and Shih is the foreign names that have been borrowed into Chinese, such as 路透社 /lù tòu shè/ (Reuter's News Agency) and 阿拉法特 (Arafat) even have three and four syllables, which constitute an exception to the aboved-mentioned notion as well.[19]

To sum up, the aim of this chapter was only to provide a bird's-eye view of Chinese characters. Viewed some of the distinctive features of the Chinese lexicon, which differs from the alphabetic and non-tonal system of most European languages, will enable us to undertake a more penetrating study of the lexicon in general. With this in mind, in the following Chapters, we will inaugurate a new survey of Chinese characters.

[19]In order to avoid the inclomplete explanation of the "morpheme" notion, and facilitate the analysis of the lexical structure of Chinese, Yip (2000) tries to propose an alternative notion: *mononym*. According to her, this is a set of *monosyllabic word-building primes* in Chinese, which differs from a morpheme in the sense that (i) it can be not only a meaningful morpheme, free or bound, but also a sub-morpheme lacking meaning (like '-ceive' in 'receive'); (ii) it is exclusively monosyllabic; (iii) it is always potentially separable from other mononyms and formally deployable in its own right. However, this alternative notion seems to avoid dealing with the three main points mentioned here. For more details about Chinese mononyms, please refer to Yip (2000).

Part II

Background

Chapter 3

A Review of Hanzi Studies

The purpose of this chapter is set out to review the ancient and current study of Chinese characters (also called Hanziology). It begins by outlining some of the historic developments from the ancient Chinese lexicography and philology, to contemporary discoveries in the linguistics. Generally, characters can be viewed from different angles. This chapter concentrates mainly on discussions about the structural descriptions of the Chinese characters as well as the psycholinguistic observations. As this thesis is primarily about Hanzi and concept, emphasis will be placed on the issue concerning the semantic information which Hanzi "carries". This sets the scene for an exploration of a theoretical framework currently proposed in the contemporary Hanzi study. In the process I hope to clarify some of the crucial issues in the literature about Chinese writing system, and lay out the foundation of our survey in the coming chapters.

3.1 Hanziology: A Definition

A science begins with the identification and definition of its object of study. In the case of a writing system, it can be defined as "*a system of more or less permanent marks used to represent an utterance in such a way that it can be*

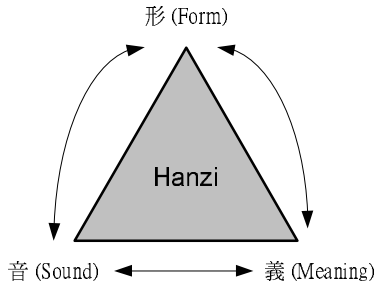


Figure 3.1: Hanzi triangle

recovered more or less exactly without the intervention of the utterer."[1] Due to the specifics of Hanzi, the studies of Hanzi is difficult to anchor. For the reader with alphabet script background, there could be some perplexity if confused with some similar disciplines, such as Orthography, Etymology, and Onomasiology.[2]

In China, the study of Chinese characters in ancient times was called "Xiǎo-Xué"(小學), which means "the teaching of characters to children". Before the 19th century, most *linguistic* study in China centered on the characters: their closely interconnected relationship between *forms, meanings and phonological structures* (see Figure 3.1). Correspondingly, there have been three traditional branches of Chinese philology: Wénzì-Xué (scriptology) Xùngǔ-Xué (text-based semantics) and Shēngyùn-Xué (historical phonology), which target these tripartite properties of Hanzi respectively.

In the 1950s the scriptology was referred to as 漢字學 /Hànzi-Xué/ (Chinese characterology or Hanziology). Since the 1980s, a new synchronic discipline called modern Hanziology (or Sinographemics) has emerged in China. In order to differentiate this from traditional Hanziology, which focused mainly on

[1]By this definition from Daniels and Bright (1996), writing is bound to language, consequently, pictograms that are not couched in a specific linguistic form are excluded.

[2]With its German synonym *Bezeichnungslehre*, Onomasiology is the subdiscipline of semantics which starts out from "Sachverhalten und Begriffen der realen Welt" looking for appropriate linguistic expressions (Wörter/Wortformen) to denote them. (Bußmann 1990: 545,672)

the diachronic aspects of the forms, meanings and pronunciations of Chinese characters, modern Hanziology deals with various uses of characters today including their use in information processing for computers. It is a multifaceted script science which, in addition to traditional methods of analysis, such as the "Six Writing" approach, (which to be introduced later) also employs theory and resources from contemporary linguistics, statistics, cognitive science and computer technology. In the following, we will review these by selecting some main points in the context of Hanziology.

3.2 Indigenous Frameworks

Before pursuing the theme any further, this section delineates first what may be called the traditional view of Chinese characters. Specifically, I'll focus on two classical theories: XǔShèn's six writings of Hanzi classification, and Wáng Shèn-Měi's right-hand side assumption which could be of great benefit to the understanding of later discussions.

3.2.1 Six Writing: Principles of Character Construction

During the Han Dynasty, around 120 A.D., a philologist named *Xǔ Shèn*, compiled the earliest "dictionary" of Chinese characters 說文解字 /Shuō Wén Jiě Zì/ ("An Analysis and Explanation of Characters and their Components") with a compilation of 9,353 characters. He divided them into six categories according to the way they were constructed, and called them Liù-Shū, ("the six writings" or "the six principles"). It can be seen as an early classification system that intended to assign every character to one of the six categories.

These six principles were not devised by Xǔ Shèn himself, but were merely his induction and summary of the ancient ways of creating characters. Strictly speaking, only the first four of these categories are true methods of character construction; the last two categories are just methods of expanding the range of use of an existing character. In the following, we illustrate these

six categories mostly based on Yip (2000):[3]

1. Xiàng-Xíng (象形 : lit. "resembling shape", the picto-graphic principle): Characters made by this principle are simple drawings aimed at representing real-life concrete objects, later, these characters were stylized into a squarish pictograph. For example, 日 /rì/ ("sun") was originally the drawing of a circular sun with a dot inside; and 母 /mǔ/ ("mother") originally the drawing of a woman whose breasts have been highlighted.

2. Zhǐ-Shì (指事 : lit. "pointing at situations", the picto-logic principle): This principle indicates that a stroke can be added to a pictograph posing as a logical reference point for the matter under discussion. e.g. 血 /xǐe/ ("blood") - a dot is seen above 皿 /mǐn/ ("vessel" especially that which was used during an oath ceremony) to mean the blood itself; and for the character 本 /běn/ ("fundamental"), a horizonal line is added to 木 /mù/ ("tree") below, indicating where the "root" is.

   A subset of this types is composed of characers whose component strokes designate purely abstract notions such as number. e.g. 一 (one), 二 (two), 三 (three). Characters of this kind make up the smallest proportion of Chinese characters.

3. Huì-Yì (會意 : lit. "assembling meaning", the picto-synthetic principle): Two or more pictographs can combine together to form a new character. In this case, the meaning of the resulting character is a function of the meaning of the pictographs of which it is composed. e.g.明 /míng/ ("bright") is composed of 日 /rì/ ("sun") and 月 /yuè/ ("moon"); 信 /xìn/ ("trusting") is composed of 人 /rén/ ("man") and 言 /yán/ ("speaking").

4. Xíng-Shēng (形聲 : lit. "integrating shape and sound", the picto-phonetic principle):

[3]More information in English can be found at http://www.chinaknowledge.org

A meaning component (also known as a "radical", "determinant" or "classifier") combined with a sound component which serves as a pronunciation guide. e.g.擎 /qíng/ ("lift up"), where the component 手 /shǒu/ ("hand") signifies the basic semantic category, and the component 敬 /jìng/ gives a clue of pronunciation. The overwhelming majority of modern Chinese characters belong to this category.[4]

As we can see, the four principles of character formation described above do not totally deviate from a *pictographic* stance. However, in the development of the writing system, there are some abstract ideas for which pictographic devices would not work well. The following two principles came into scence to fill the vacancy.

5. Zhuǎn-Zhù (轉注 : lit. "mutually interpretation", the mutually interpretive symbolic principle):
   A new character is created based on the borrowing of an existing character with a similar meaning. For example, "it so happened that the meaning of 'deceased father' finds a semantic overlap in the character 老 /lǎo/ ("old man") and, deciding to borrow part of its form and pronunciation, comes up with 考 /kǎo/, retaining the top written element and vowel quality of the borrowed form."[5]

6. Jiǎ-jiè (假借 : lit. "false borrowing", the phonetic loan principle):
   This time the borrowing procedure may not be thinking in terms of the similarity of meaning but that of sound. For example, if we want to assign a character to the notion of "want" (/iào/), we find that this is not easily depicted by a pictograph. A convenient way to solve this problem would be to *borrow* a homophone (要 iāu, "waist") among

[4]It has been claimed that, as many as 97% of Chinese characters can be analyzed as this type. See DeFrancis (1984).

[5]Over the years there have been various interpretations of these two principles. I adopt the explanation and examples chosen by Yip (2000:42).

existing characters for the purpose, despite that fact that there is no connection of meaning between the two.[6]

Another important contribution of Xǔ-Shèn worth mentioning here is the invention of the *semantic radical classification system*. In "Shuō Wén Jiě Zì", characters can be composed from a range of 540 *semantic radicals (bùshǒu)*.[7] These "radicals" became integral parts of characters and describe fields of meaning. The modification of extant characters by additional (semantic) radicals can lead to an enormous increase of their number. For example, characters such as 抱 (embrace), 採 (pluck), 抓 (grasp) all share the radical 扌 (hand).[8]

3.2.2 Yòu Wén Theory

As already mentioned, characters following the picto-phonetic principle constitute the major part of Chinese characters, and these characters have become the main object of the study of Chinese writing. However, in the Song Dynasty, a philologist named Wáng Shèng-Měi raised the so-called Yòu Wén Shuō (右文說, lit: "right component theory"). The kernel of this theory is an alternative interpretation framework concerning the meaning composition of characters as a whole (合文為字). Based on his analysis, not only the radical gives a clue about the meaning of the combination as a whole, but the phonetic parts which mostly stand on the right hand side, contribute to the

[6]To avoid the proliferation of such homonyms, there are other mechanisms involved. This is, however, beyond the scope of this thesis. Interested readers are referred to Yip (2000).

[7]The second Qing Emperor KāngXī (who reigned 1662-1722) commissioned the compilation of a dictionary of Chinese characters. In 1716 the Kāng-Xī ZìDiǎn or Kāng-Xī's Character Dictionary appeared. The work collected a total of 47,035 character and reduced the numbers of radicals to 214. Modern dictionaries today still use this system, sometimes adding their own radicals.

[8]After a long standing evolution, many scholars believed that the meaning-indicating function of the radicals has became variable. Coulmas (2003:56-57) even claims that, considering the great diversity of radicals, "it is impossible to see in them anything resembling a logically consistent and comprehensive system of semantic categories," others disagree. The next chapter will go more in detail.

meaning composition as well (聲符兼義). Though this theory took a back-seat to traditional Hanziology, since its emergence, there has always been a polemical undertone along this line, till the present time, many echos are still easy to be found. This background knowledge can provide a basis for the controversial issue which will be discussed later.

3.3 Contemporary Linguistic Studies

Chinese characters have been one of the main concern in some European philosophical writings (e.g., Leibniz(1971), Humboldt(1979), Derrida(1970)). This section gives an overview of Chinese writing in the linguistic study of writing systems: its status in the classification of human writing; and fundamental controversies involved in respect of structural descriptions and psycholinguistic viewpoints.[9]

3.3.1 The Classification of Writing Systems

In the post-Saussurean era, most linguists working in the area of global writing systems have paid much attention to nomenclature and typology. The tripartite scheme of ideogram, syllabogram, alphabet was the most notable one and has remained the most popular, but has led to some suggestions about the nature of certain scripts, and several alternatives have been offered (Daniels and Bright 1996).

In Sampson's classification scheme (Sampson 1985:32), writing as a whole can be *semasiographic* or *glottographic* (see Figure 3.2). He used the term *semasiographic systems* for systems of visible communication which indicate *ideas* directly, without tying them to any one spoken language. In contrast, *glottographic systems* provide visible representations of spoken-language utterances. Writing systems, as Sampson saw them, and in keeping with the widely accepted meaning of the term, are all glottographic. Note that in Fig-

[9]This section is written based on a previous paper, Hsieh (2003b).

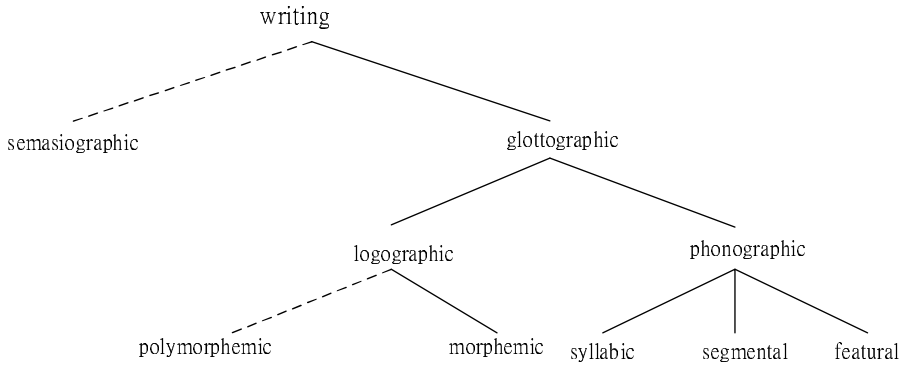


Figure 3.2: Sampson's classification scheme

ure 3.2, the dotted lines indicate, for example, that the problem of whether semasiography is deemed to be true writing or not, is open to question. Among glottographic systems, the major division is between *logographic* and *phonographic* scripts. *Logographic writing systems* make use of symbols that represent words or concepts. The shape of the symbols employed is often closely related to the meaning of the respective concept. In contrast, in *phonographic writing systems*, the meanings of the objects denoted by the symbols does not play a role. They establish a relationship between the symbols employed and the pronunciation of the elements denoted.

Sampson asserts that Chinese writing is a *logographic (- morphemic)* rather than a phonographic system (as marked in Figure 3.2). This is because Chinese characters do *not* encode phonological information, but represent morphemes directly. As introduced in the previous section, the units of script are co-extensive with syllables, which are phonological units. For this phenomena, Sampson argues that this is merely ".. an accidental consequence of the fact that in Chinese the minimal meaningful units, or morphemes, happen always to be one syllable long" (Sampson 1985:148).

It is remarkable here that Sampson opposes the notion supposing that Chinese writing is semasiographic, what he called a very widespread misunderstanding among Western intellectuals. Such a viewpoint, he continues,

=3pt

Major types	Subordinate types	Examples
syllabic systems	"pure" syllabic systems	Linear B, Kana, Cherokee, Yi
	morphosyllabic systems	Chinese, Mayan
consonantal systems	"pure" consonantal systems	Phoenician, Hebrew, Arabic
	morphoconsonantal systems	Egyptian
alphabetic systems	"pure" phonemic systems	Greek, Latin, Finnish
	morphophonemic systems	English, French, Korean

Table 3.1: DeFrancis's classification scheme

is reinforced by the common use of the word 'ideogram' to refer to Chinese graphs, suggesting that they stand for ideas rather than words.

Such accounts of script types are not wholly without controversy. DeFrancis (1989) has even made a stronger claim that *all* full writing systems are largely phonographic, that no true logographic scripts exist. He even contends further that we need to throw out the term "ideographic" altogether.

In fact, DeFrancis's argument is simply based on the mainstream attitude toward the structural interpretation of Chinese characters we have introduced: The large majority of Chinese characters that have been created throughout history are "semantic-phonetic compounds" (形聲字), where one element in character gives a hint of the meaning, and the other element gives a hint about the pronunciation. For example, of the 9,353 characters that had been developed up to the second century A.D., about 82% of these characters were semantic-phonetic compounds (DeFrancis 1989:99). Thus, for DeFrancis, Chinese writing is not logographic at all, but rather what he terms *morpho-syllabic* – It is basically a phonographic writing system with additional logographic information encoded. Figure 3.1 shows DeFrancis's classification of writing systems.

As Sampson observes,[10] it is logically possible that a writing system could encode various levels of linguistic information. However, the fault of these

[10]Quoted from Sproat (2004) in his course "Writing Sytems", Spring semester 2004. In http://catarina.ai.uiuc.edu/L403C/lecture1.html.

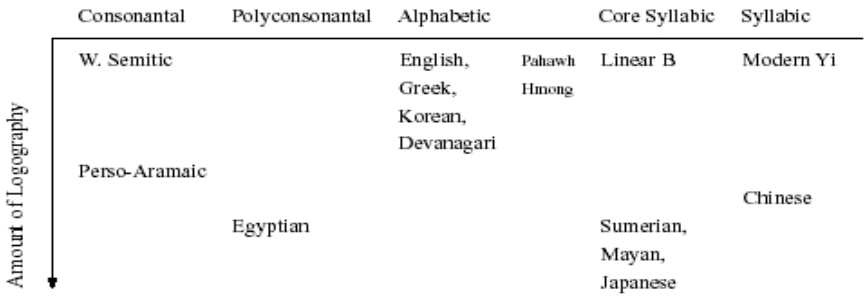


Figure 3.3: Sproat's classification scheme

taxonomic settings lies in their dependence on the notion of "purity", that is, any given writing system must be *either* meaning units (thus, logography) *or* sound units (thus, phonography).

In reality, a *pure* phonographic system is impossible, for writing does not equal phonetic transcription; neither does a *pure* logographic system exist today.[11] The writing systems in current use are rather of mixed types to a certain extent. In his later article, Sampson (1994) also suggests that assigning a writing system to a particular category must depend on a subjective judgement as to how *close* and *regular* the relationship between pronunciation and written form needs to be before one treats that relationship as the central organizing principle of a script.

Therefore, we are in agreement with Sproat (2000:142) in stating that it is important to realize that *all* writing systems probably have some *degree* of logography - defined as the graphical encoding of nonphonological linguistic information. Writing systems can thus be classified minimally in a two-dimensional space according to what types of phonological elements are encoded and what degree of logography they encode. Figure 3.3 shows Sproat's classification scheme.

Recently, following Sproat's basic notion, Fukumori and Ikeda (2002) pro-

[11]The pictographs (象形字) apearing in the early stages of Chinese script might be called logographic.

posed a new system of classifying writing systems based on the language they convey, the linguistic unit(s) they represent (i.e. words, syllables, consonants, segments, and semantic classes), their graphic shapes, and if known, their genealogy. According to them, writing systems can thus be represented in the following format:

language-(式 *shiki*) unit-(型 *gata* (genealogy-)) shape-(系 *kei* writing)

So Chinese characters in this theme can be represented as 中國語式表語音節型漢字系文字 (Chinese-shiki logographic-syllabic-gata Kanji-kei writing). And Ugartic script as: (Ugaritic-shiki consonantal/syllabic-gata Ugaritic-kei writing). Unfortunately, in this proposed format for representation, no explicit explanations for these predefined items are provided; on the other hand, the mixing of *language* and *writing* classification could result in more difficulties.

To sum up, we have introduced three main models proposed for the study of writing systems, which could be called the taxonomic model (Sampson and DeFrancis), the probabilistic model (Sproat) and the type-feature model (Fukumori and Ikeda). Among these, the probabilistic model seems to be the best model in flexibly explaining the complexities of writing systems. However, as a reasonable explanation model, it suffers a crucial deficiency: in spite of the fact that the term *degree* is used; since we have no commonly accepted quantitative measurement for the *amount of logography*, to what degree, for example, do we say that Japanese is *more* logographic than Chinese (as depicted in Figure 3.3)?[12]

3.3.2 Ideographic or logographic?

Actually, the core of the continuing debate mentioned above can be reformulated as: What kinds of information does Hanzi represent? or in linguistic

[12]Bosch (1994) proposed a quantitative measure of the orthographic depth. But unfortunately, it can only deal with relation between spelling and phonology. See Bosch et al. Measuring the complexity of writing systems. In *Journal of Quantitative Linguistics* No.1.

parlance, Do Chinese characters constitute an existing of logographic, phonographic or ideographic writing? This question gives rise to the main standing controversy in the study of the Chinese writing system. Though Sproat (2000) has *resolved* it successfully in terms of the notion of "degree", there seems to be *something* missing.

Before we embark on this something, two terms *logogram* and *ideo-gram* should be defined. A *logogram* (or *logograph*) is a single symbol that represents a complete grammatical *word* or *morpheme*; while an *ideo-gram* (or *ideograph*) is an pictorial element expressing not sound, but an *idea*.

Leibniz (1971) stood on the *ideographic* side, while Saussure, Sampson and even Chao, a famous Chinese linguist, all insisted that Chinese characters constitute a *logographic* writing system, which represents spoken words, not ideas. Chao even claims that this position has been recognized by Sinologists ever since the first half of the nineteenth century. The representative paragraph from Chao is as follows:[13]

> [...] from very ancient times, the written characters have become so intimately associated with the words of the language that they have lost their functions as pictographs or ideographs in their own right and become conventionalized visual representations of spoken words, or "logographs". They are no longer direct symbols of ideas, but only symbols of ideas in so far as the spoken words they represent are symbols of ideas.

For Sampson, the commonly used term "ideographic" should be avoided because it is not clearly defined, and furthermore, it blurs the crucial distinction between semasiographic and logographic systems in his classification scheme. There are also others who have been severe in criticizing those who think that Chinese writing is either ideographic or that Chinese characters always function logographically. For example, DeFrancis (1984) even holds the view that Chinese characters today, just like alphabetic writing, serve only as written tokens of the spoken language. He called the notion that

[13]This paragraph is quoted from Harris (2000:145).

Chinese characters represent meaning directly, without reference to speech, the *Ideographic Myth*. For DeFrancis, the "logographic" symbol proposed by Sampson, still corresponds to what he called an "ideogram".

This discusion seems not to be fair to the notion of the "Chinese script being an ideographic script", for most of them do not reveal comprehensive knowledge of Sinology. By sticking to certain mainstream doctrines, thus neglecting some alternative indigenous theories like the Yòu Wén Theory, many researchers have no difficulty convincing themselves that Chinese writing is more or less of this type or that type.

3.3.3 Word-centered or Character-centered?

As we have seen, the current theoretical linguistic models of writing systems seem to fail to be in accord with each other in classifying Chinese writing system into the taxonomy of human writing systems.

In contemporary Chinese linguistics, in addition to the controversy of *word* and *morpheme* introduced in section 2.2.2, how to anchor the *character* in linguistic theory construction of Chinese is controversary, too. For instance, character-centered (in contrast to the mainstream word-centered) approach claims that Chinese characters could not solely be regarded as writing units. Due to the tripartite property, they should be at the center of the study of Chinese linguistics, cognition, conceptualization, all these are intricately bound up with the way in which Chinese classify and convey their experience of the world through Chinese character.[14]

We believe that such debates might be illuminated from the results of psycholinguistic experiments. Sometimes the conclusion appears quite different when psycholinguistic data are considered, for psycholinguistic research does not deal with specifying the interrelationships of the elements that constitute

[14]The detailed discussion of this debate between character-centered (字本位) and word-centered (詞本位) is beyond the scope of this thesis. Interested readers are refered to Xu (2004), Pan (1997).

a writing system, but instead, how a writing system works in terms of actual perception, processing and production. So now we turn to the domain of psycho-neurolinguistical studies of Chinese characters.[15] The questions considered here are reformulated by Packard (2000:284) as follows:

- What is the Chinese "mental lexicon" ?
- Are words or characters "stored" in and "access" from the lexicon as "gestalt wholes" ?
- What role does a character play in accessing the Chinese lexicon?

In general, the *lexicon* of a natural language may be understood as all of its words and their synchronic uses. It is popularly known as *vocabulary*. The *mental lexicon* is a repository with long-term memory where a language user's knowledge of the lexicon is stored in a flexible manner. And *lexical access* is understood here as, given orthographic or phonological input, the "route"of access to find the best match "listed" in the mental lexicon.

In the research of the Chinese mental lexicon, some experiments have found that, as in the case of English, the Chinese mental lexicon takes *words* as its basic constituents.[16]

However, it must be pointed out that though the lexicons of all languages might share some similar properties, they do not necessarily contain similar lexical items or operate in the same way. For example, the "word superiority" hypothesis which is well tested in alphabet script-based languages has not been by all any means strictly ruled out for ideographic script-based languages like Chinese, since the lexicalization of the objective world with

[15]Literatures on the relavant psycholiguistic study of Hanzi are plentiful. This subsection will focus on two subjects: the *mental lexicon* and *lexical access*. For an in-depth coverage of specific theorectical controversies, relevant texts are cited in Li et al.(eds). (2004). Handbook of East Asian Psycholinguistics.

[16]In addition to psycholinguistic experiments, Huang et al (1998) also adopt a corpus-based investigation of the Chinese mental lexicon. The result shows that *words* reflect psychological reality.

its multifarious phenomena from different speech communities with subjective imaginations can be an arbitrary language-specific and culture-specific process.

In their introduction to a special issue of *Language and Cognitive Processes* devoted to issues of processing East Asian languages, H.C. Chen and Zhou (1999) expressed feelings of uncertainty similar to those of Chinese psycholinguists about the concept of the *word*.

> For instance, contemporary theories of language processing unexceptionally consider words as the basis of complex comprehension processes [...] This is not surprising, because, after all, words are transparent units for speakers of European languages. However, it is not obvious whether the same arguments and conclusions relating to word processing that have been reached through psycholinguistic studies with European languages can be generalized to other languages, such as Chinese, in which words are not transparent units (pp. 425-426). [17]

In fact, in the case of Chinese speech comprehension and production, the question of interest is: what kind of unit is stored in the mental lexicon? Many experiments have been done to attempt to answer this question: Every word the speaker knows (Butterworth 1983); only morphemes and some morphological rules (Taft and Forster 1975); words are stored in decomposed fashion which are accessed via characters (Zhang and Peng 1992)... and so on.[18] Hoosain (1992) and Elman (2004) have showed that, a larger portion of Chinese multimorphemic words, in contrast to English, are not listed in the lexicon but rather 'have meanings that are arrived at in the course of language use', and that the Chinese lexicon contains a large number of individual morphemes and a "lexical tool-kit" which allows for the creation and understanding of multimorphemic words (Hoosain 1992:126).

Another dimension of psycholinguistic research that might shed light on the classification of writing is the study of the *reading process* which is a

[17] This paragraph is quoted from C-H.Tsai.(2001). Chapter 2.
[18] Interested readers can refer to Packard (2000:297).

hot topic in the research of the lexical access (Coulmas 2003). Some previous researches concerning the neural organization and processing of Chinese characters, have used fMRI to compare brain activation, and suggested that the underlying neuroanatomical mechanism of Chinese reading is unlike that of English word reading.

Whereas alphabetic systems are based on the association of phonemes with graphemic symbols, Chinese writing is based inherently on the association of meaningful morphemes with graphic units. Zhou and Marslen-Wilson (1996) even argue that in the reading process of Chinese, *direct visual-to-meaning* is the only way to access information stored in the mental lexicon.

Over the years, such contentions have been vehemently rebuted by opponents.[19] Some reports show that there is no clear experimental evidence supporting the hypothesis that reading a word written in Chinese involves processes different from those involved in reading a word in an alphabetic system (Giovanni 1992); other research has found that phonological information is available and used by readers of Chinese as well, while semantics, rather than phonology, is delayed at the character level (Perfetti 1999). After reviewing the current main works in this field, Sproat (2004) made a cogent concluding remark: "One must conclude that Chinese characters map, in the initial stages of processing, to a level of representation that is basically *phonological*.

Based on the current results of psycholinguistical studies, we seem to come to a conclusion: Every writing system, including character-based ones like Chinese, is basically phonographic. However, one hidden point here, nonetheless, concerns the *logic* of theoretical development. Just like most studies in the classification of writing systems, most psycholinguists presume that the concepts of *ideographic* and *phonographic* are mutually exclusive.

The experimental design sets the *ideographic property* as a target at

[19]See Tzeng et al. (1992) and Kuo (2004) for recent reports.

first, then experiments are performed to support, *to what degree* Hanzi carry phonological information, or *to what degree* phonological information is involved in the reading process of Chinese. To my knowledge, there has been no theory or experiment done in a converse way: That is, assuming that every script functions phonologically, an experiment is made to see *to what degree* these scripts carry ideographic information. For example, in English words with *dis-* usually denote negation; in Chinese, characters with the radical 忄 usually denote some mental state in respect of emotion.

In this context it is interesting to discuss the notion of *the Orthographic Depth Hypothesis* (ODH) which is closely utilized in the psycholinguistic research we have discussed. The ODH in its weak form [20] states:

> "all written languages allow for both a grapheme-to-phoneme correspondence route (route A), and for a lexical access route (route B-D, or perhaps C-D) (See Figure 3.4). But the cost of each route directly relates to the type of orthography (deep or shallow) involved (Katz and Frost 1992)"

According to the ODH, the Chinese language, which has strong irregularities between orthography and phonology correspondence, should be called orthographically "deep" languages. Readers of Chinese in the naming experiments[21] might take the route B-D or the "deepest" route C-D.

Again, the notion of ODH is phonologically oriented. Experiments and measuring algorithms for the *Logography depth* (or *ideography depth*) on the other hand are expected.

[20] Quoted from Sproat (2004).

[21] This is a kind of lexical-access experiment, where subjects are presented with a written stimulus on a screen, and are asked to read the stimulus aloud. The time between the presentation of the stimulus and the onset of vocalization is measured.

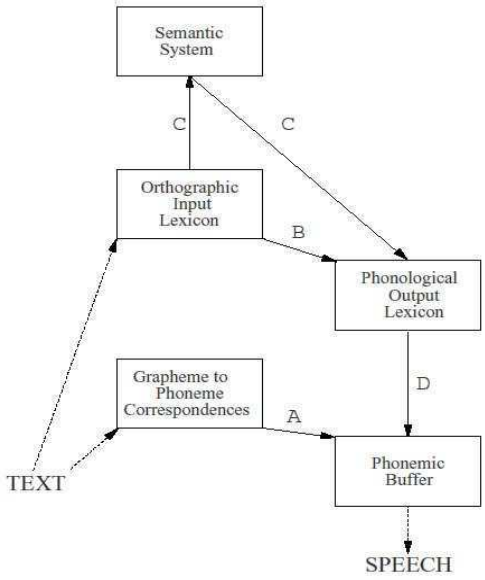


Figure 3.4: Orthographic Depth Hypothesis

3.3.4 Critical Remarks

In sum, as we have reviewed, previous literature reflects divergent views in regard to the nature of Hanzi. While some linguists argue against the traditional understanding of Hanzi and prefer a word-based phonography-oriented linguistic setting, there are also some experimental psycholinguistic data which support the claim that is consistent with the character-based ideography-oriented linguistic setting, at least in so far as the notion that characters have an independent status in the Chinese mental lexicon. More specifically, a wide range of researchers from several different disciplines have all, in their various ways, sustained or to rejoindered the proposition that "Chinese writing is ideographic". These discussions call for some general considerations.

First, as to the question of whether Chinese writing is ideographic or logographic, DeFrancis misguided the direction of discussion in that, when we say that Chinese writing is a sophisticated ideographic system, this does not mean that it conveys only meaning without regard to sound. For those re-

searchers who speak of ideographic writing, they do not necessarily content that concept or idea writing is possible. What they want to emphasize is rather that the signs of the systems they are concerned with are to be interpreted primarily not in terms of sounds but of meaning (Coulmas 2003:41). Admittedly, from a semiotic point of view, there exists neither a purly logographic nor a purly phonographic writing system in today's world. there are probibilistic regularities rather than strict rules in describing the structure of writings.

Secondly, for the psycholinguistic aspects, we have briefly reviewed the Chinese case in the study of the mental lexicon and lexical access via reading. There seems to have two camps in these repects as well. However, it should be carefully verified that the two parties are focusing on the same facets of a linguistic phenomenon (e.g. words, characters or one-character words ?). In addition, though most experiments have shown that for the most part, Chinese writing represents these phonetically, so do all real writing systems despite their diverse techniques. The conceptual dominance of characters results in words in Chinese not behaving as words in inflecting or agglutinative languages. Aside from orthographic depth, a method of measurement of logographic/ideographic depth is expected in the future.

Both structural descriptions and psycholinguistic modelling seem to presume that the notions of *ideography* and *phonography* are mutually exclusive. We will argure that, at least in the case of Chinese script, the polyfunctionality of signs is inevitable. Like Egyptian hieroglyphs, they serve as word signs, phonetic determinatives and semantic classifiers at the same time (Coulmas 2003:60). To break the theoretical impassé, this thesis will take a pragmatic position based on the tripartite properties of Hanzi: Chinese characters are **logographic** (morpho-syllabic) in essence, function **phonologically** at the same time, and could be interpreted **ideographically** and implemented as conceptual primitives by computers. We will elaborate on this proposal in the next chapter.

3.4 Contemporary Hanzi Studies

3.4.1 Overview

Broadly conceived, linguistics is the study of *language*. Hanzi, though a core topic in traditional Chinese philological studies, has only gained recognition in the field of Chinese character teaching of the current time. Although many excellent scientific works concerning with Chinese character have been made, such as Stalph (1989), Bohn (1998), Harbaugh (2004), etc., most of them focused mainly on the elaborate analysis of shape structure of character: its component and distribution. Only very few of them, such as Ann (1982), tried to give a systematic explanation of the ideographic trait of Chinese characters. But the weakness of these works, as was criticized by Guder-Manitius (1998:114), lies in their "Volksetymologie"- oriented approach, and therefore came across as being rather impressionistic rather than scientific.

In the following, I will dwell on the semantic issues around Chinese characters by sketching a new theorectical framework called *Hanzi Genes Theory* (HG theory, hereafter) proposed by B.F. Chu and his colleagues.[22] This theory is based on the discovery of the interpretation of the conceptual dimension of Chinese characters. In particular, it also tries to reveal the *common sense knowledge* carried by Hanzi.

[22]Chu's team at Culturecom Holdings Limited (www.culturecom.com.hk) has tried to construct the blueprint of the general theory of Chinese characters. They have developed many ideas and products concerning Hanzi-based Chinese information processing, such as the vector glyph generation device and the CāngJié Input method, the latter has become one of the current standard Chinese input method. Other software and hardware solutions for the Chinese IT market and the lesser developed world, include Chinese E-books, voice and character recognition programs, Hanzi-embedded CPU (Culturecom 1610 and 3210 processors), a Chinese character driven multimedia platform (text-to-scene system), and so on. In this section, I will deal only with language processing / understanding theses explained in his book: Engineering the Genes of Chinese Characters. The laboratory website is at : http://www.cbflabs.com (only in Chinese). For more information in English about the CāngJié Input method, see http://www.cjmember.com

3.4.2 Hanzi Gene Theory: a Biological Metaphor

It would be useful to informally first introduce some of the essential biological terms which might be closely associated with the coming discussion.

All living organisms are composed of cells, and each cell contains the same set of one or more chromosomes – strings of deoxyribonucleic acid (DNA) – that serve as a "blueprint" for the organism. A chromosome can be conceptually divided into genes – functional blocks of DNA, each of which is assembled into a chain to make a protein. Very roughly, one can think of a gene as a special sequence in a strand of DNA which encodes a trait, such as eye color. The different possible "settings" for a trait (e.g., red, black) are called alleles. Each gene is located at a particular locus (position) in the chromosome.

The motive why biological terms have been used as an inspiration for constructing NLP theories and systems for Chinese is that, to a certain degree, the language of DNA and natural languages share similar traits.[23] The biologist R. Pollack (1994) discovered out, the DNA language of the cell, and the way it is made manifest in protein, has its parallels in the Greek and hieroglyphic Egyptian inscriptions found on the *Rosetta Stone* unearthed in 1799. DNA and the stone (as the French linguist Champollion decoded) both carry a linear representation of a text into a sculptural one. In both, information is "translated" from an alphabetic sentence of many letters (base pairs or the Greek alphabet) into a sculptured figure (protein or hieroglyph).

Hanzi, like the hieroglyphs on the Rosetta Stone, could be thus analogized

[23] Actually, in recent years, with the rapid advancing of *bioinformatics*, many quantitative linguistic theories and language processing techniques have been contributed to this new research area. By regarding information stored in *DNA sequences* as information stored in *natural language*, a number of statistical methods successfully employed in NLP are now being applied to the study of DNA sequences. See Mantegna et al (1994). Linguistic features of noncoding DNA sequences. *Physical Review Letters 73(23)*, and Ng and Li (2003). Statistical linguistic study of DNA sequences. *The 7th Conference on AI and Applications*.

as a protein[24] which undergoes the process of biological *translation* from genes to proteins. While we cannot yet unveil the meanings of a gene the way a protein does each time it folds into its active ("meaningful") form, we have learned how a protein is put together from the information in a gene (Pollack 1994).

The definition of "character" in genetics is: "A structure, function, or attribute determined by a gene or group of genes." We also find similar parallels in defining "Chinese character (Hanzi)": A form, sound, meaning determined by a Hanzi gene or group of Hanzi genes.

Nonetheless, the *Hanzi Gene* still seems unclear at first sight. For Chu (1992),[25] *genes* are the basis of understanding and the most basic analyzable unit within the system of the universe. Being built upon genes, all ideas would be easy to figure out when their interfaces are clearly defined. Chu even made a strong argument that ".... constructed from the *genes* of the objects and events forming the cosmos, Hanzi comprises a limited set designated as the 'genes of common sense that could be used to express limitless understanding. Namely, the construction of Hanzi-encoded concepts is an artificial form of 'genetic engineering' in which the 'meaning' could be *inferred* from the 'shape' of the character."

From here on, Chu starts to lay out the main tenets of his theory. In short, HG theory proposes that the knowledge representation of the essence, qualities, and properties of the objects and events surrounding us in the world *are* all embedded in the characters. In this sense Chinese characters parallel the genes of biology, which condense the life forms of all living organisms. Like biological genes, a limited number of Chinese linguistic genes allow the codification of all of the capabilities of language. The most significant

[24]Interestingly, the proportion of estimated number of proteins divided by the number of genes $\frac{1000{,}000}{30{,}000}$ is similar to the estimated number of currently-used Chinese characters (in GB code) divided by the number of the radicals $\frac{6763}{214} \simeq 33.3$.

[25]Due to the difficulties of translation, the following brief explanation draws heavily on the article and notes of Chu's student Walter van Patten.

implication is that a limited set of characters can create a kaleidoscope of meaningful expressions.

In detail, this theory aims to capture the basic units (called *Hanzi Genes*) by analyzing the six dimensions of Chinese characters on the computing environment: *order, code, shape, argument, voice and meaning* (Chu 1992). Not by just boosting the theory, Chu claims that a limited number of *Hanzi Genes* have been found. Through the classification and sequencing of these Hanzi Genes, the ultimate goal is to attain computational intelligence in processing and understanding Chinese.

It seems that the theoretical ambition is grandiose. The scope is intended to span the full range of interests from classical problems involving the philosophy of the mind, through issues in cognitive science and life philosophy to ideas related to artificial intelligence and computer science.[26] The following will not provide an exhaustive treatment of the entire panorama, our primary emphasis will be placed upon the Order Genes, Code Genes and Meaning Genes of Hanzi which could be of crucial importance in the field of Chinese lexicography, computer processing and natural language understanding, respectively.

Order Genes (for the Character Sorting Problem)

The analysis of "order genes" is set to solve the sorting and indexing of characters. As widely recognized, one of the powerful applications of the Latin alphabet and its various derivatives lies in the ease and clarity of sorting.

[26] According to the explanation from the website, the systems and hardwares implemented based on the Hanzi Gene Theory were not designed only as Chinese input method but to propose a global humanities system. A clear spell-out of the theory and its various implementations in greater detail is beyond the scope of this thesis. Readers who are interested in the theoretical part are asked to refer to the website of Chu Bong-Foo Laboratory, (unfortunately most of the content are written in Chinese); readers who are interested in the implementation part, such as the claimed world's most cost-effective embeddable multilingual computing architecture, and detail about the first 32-bit Chinese CPU, jointly developed with IBM by embedding the MCGE and Midori Linux into the PowerPC microprocessor, please refer to http://www.culturecom.com.hk .

Large amounts of alphabet-based information can be sorted, searched, and classified at great speed. In contrast, Chinese employs no alphabet. On first observation, one might surmise that a systematic sorting method would not be feasible in the Chinese lexicographical praxis, due to the influence of the "alphabetless".

Indeed, the arrangement of characters in a Chinese dictionary[27] and methods for looking up these characters are not so clear at a glance. Basically, there are two *indexing schemes* designed to order characters in dictionary:

- *via Transliteration Systems*

  If the pronunciation of a given character is known, one may use either the romanization systems (transliteration systems using the Roman alphabet like Hànyǔ Pīnyīn, or non-romanization one like Zhùyīn Fúhào (known as ㄅㄆㄇㄈ or BPMF) to locate the character.[28]

- *via Radical Systems*

  If the pronunciation of a given character is unknown, another common practice is to use the *214 KāngXī Radicals* to first locate the main cluster of characters, and then count the strokes left over to obtain its pronunciation and gloss. For instance, if one wants to look up 信 (/xì/, "*faith*"), one character consists of the radical 人(/rén/,"*man*") and seven additional strokes. First, one finds the radical of this character ("man") in the index page of the dictionary, and then skims through one additional stroke, two additional strokes, etc., until one finds entries for seven additional strokes.

[27]The differences between "character dictionary" and "dictionary" will be discussed in Chapter 5.

[28]The Hànyǔ Pīnyīn system was developed by the People's Republic of China in 1958 (Zhùyīn Fúhào was used before then), and is now the only transliteration system used in mainland China. It uses 25 of the 26 English letters (except "v") to represent the sounds of Mandarin. In 1981, the International Standardization Organization (ISO) decided to use Pinyin as the official system to transcribe Chinese proper names and relevant phrases. And as for the BPMF system, whose graphs are derived directly from Chinese characters, is now nonetheless the predominant system used padagogically in Taiwan.

Cosmology		Writing strokes		Human Nature		Transformations	
日	A	竹	H	人	O	尸	S
月	B	戈	I	心	P	廿	T
金	C	十	J	手	Q	山	U
木	D	大	K	口	R	女	V
水	E	中	L			田	W
火	F	一	M			卜	Y
土	G	弓	N				

Figure 3.5: The 24 main Cang-Jie signs. The 4 rough categories here are designed for the purpose of memorizing.

The first method presumes one is either familiar with the Roman alphabet and the resulting transliteration, which is therefore easy only for European learners. The non-romanization one like BPMF, takes a lot of time to learn, both for European and Chinese learners. In addition, as stated, if one has no idea about the pronunciation of the character, there is no way to find them. Many modern dictionaries do not even have such information about characters. The second method, which has been developed over a thousand years, has disadvantages as well. For example, it is sometimes difficult to determine the radical; the counting of strokes is occasionally a probelm, too. For a native speaker it often takes 5-6 minutes to find out a particular entry.

This background brings up to Chu's invention, the *Cāng-Jié* system. In this system, Chu proposed a set of 24 main condensed shapes of characters - called Cāng-Jié signs (Figure 3.5)- that can reproduce all the patterns forming Chinese characters[29] which are condensed from the shapes of Chinese characters.

With two additional exception keys, Cāng-Jié signs correspond to 26 English alphabet keys. And with the rules that determine the selection of signs, each character has its own Cang-Jie code and therefore characters can be sorted and searched as well as words in alphabet-based languages. These

[29]Readers may imagine these signs as a sort of "Chinese alphabet"

Cāng-Jié signs constitute what are called "order genes" here.

Meaning Genes (for Character Semantic and Conceptual Representation Problem)

Before we go to the detail, let us first probe into some relevant background. Traditionally, among Chinese linguists, it has been asserted that (as it still is in today's mainstream) the overwhelming majority of characters that have been created throughout history, are so-called Xíng-Shēng Zì ("*semantic-phonetic characters*"), where one element (also called a radical) in the character gives a clue of the meaning, and the other element (a phonetic determinative) only provides information about the pronunciation corresponding to the character. For example, 賙 (/zhōu/,"to help in charity") is composed of the radical 貝 (meaning:"money"), and the phonetic determinative 周 (pronounced as "/zhōu/").

Alternative views assert that the mainstream views are somehow misguided by Xǔ Shèn. They argue that the Chinese writing system is almost totally ideographic in the sense that the vast majority of characters, except for proper names and a few instances of onomatopoeia, can be interpreted as Huì-Yì Zì ("*compound ideograms*"), where each element (be it a meaning component or a sound component) in the characters contribute to the overall meaning. If we probe into the relationship between various forms and their derivations and combinations, we shall find even more regular and systematic correlations in terms of meaning and form.

The latter view has been an undercurrent of the study of Hanziology, as we have introduced previously in section 3.2.2, and contemporary echoes are easily found.[30] Though these two main views all copiously quote the classics, no one seems to be able to give a systematic explanation that answers the question with certainty: "*Do Chinese characters really "encode" a sophisticated system of meaning?*" For such classical dispute, HG theory claims that

[30]See T.K.Ann (1982).

such a question should not be treated as a problem of *etymology-proper*, but as an *interpretive* problem. In the context of the information age, the possibility of implementation and verification turns out to be a more convincing way of thinking about this issue.

Among the "Hanzi Genes" proposed by HG theory, the "meaning genes" are the most controvervial issues. He proposed that there are 512 such "meaning genes": 256 common sense genes + 256 concept genes, which could be extracted and induced from Chinese characters. The concept genes will be discussed in the next section. Here we will first introduce what he claims as one of the unique features of Chinese characters, namely, the "common sense classification structure" encoded in Chinese characters. We can retrace the "common sense classification" of Chinese characters to their classical division into *head* and *body* components (Chu 1998: D4:14)

According to Chu, most of the characters could be interpreted as Hùi-Yì Zì ("compound ideograms") that can be decomposed into two parts, namely, 字首 (部首) (character head (CH); head component), and 字身 (聲旁) (character body (CB); body component). The CH part means the basic semantic category, while the CB part points to a complementing meaning and to the approximate sound of the character. There are 256 components in total (CH + CB), which are referred to as 常識基因 (common sense genes).

The following table shows some examples of character heads and character bodies, together with their combinations. For example, the CH 心 (compressed as 忄) can combine with different CB such as (半 (half)、吾 (I)、曷 (expel)、周 (close) ...) constituting characters such as 怑 (feel ill)、悟 (comprehend)、愒 (rest)、惆 (sad).... All these characters carry the CH 心, so they share similar semantic categories which relate to human feelings or a mental state. Examining CBs individually gives further information about the fine differences between them. We can also examine them starting from the CB. For example, the CB 曷 (expel) can combine with different CH like 人 (human)、日 (sun)、心 (heart)、水 (water)、言 (speech)...., and constituting characters with

a core or derived meaning of the CB, such as 偈 (rushing)、暍 (heatstroke)、愒 (rest)、渴 (thirsty)、謁 (call on) ... respectively.

	人 (human)	日 (sun)	心 (heart)	水 (water)	言 (speech)	手 (hand)	口 (mouth)
半	伴	·	怑 (not all right)	泮	·	拌	·
吾	·	晤	悟 (comprehend)	·	·	·	唔
亢	伉	·	·	·	·	抗	·
召	·	昭	·	沼	詔	招	·
曷 (expel)	偈 (rushing)	暍 (heatstroke)	愒 (rest)	渴 (thirsty)	謁 (call on)	揭 (hold high)	喝
門	們	·	·	·	·	捫	·
周 (close)	倜 (unconventional)	·	惆 (sad)	·	調 (tone)	·	啁 (chirping)
俞	偷	·	愉	渝	諭	·	喻
旨	·	·	恉	·	·	指	·
齊	·	·	懠	濟	·	擠	·
戈	伐	·	·	·	·	找	·
足	促	·	·	浞	·	捉	啶
亡	·	·	忘	·	·	·	·

* Some examples of character heads (in the across row), character bodies (in the down column) and their affiliations. Note that the dots mean such combinations do not exist in a modern Chinese dictionary, though they remain possible combinations. In addition, due to space limitations, not all characters are glossed with English translations.

3.4.3 Hanzi, Concept and Conceptual Type Hierarchy

This subsection explores the Hanzi-driven *Concept Genes* and their relationships proposed by HG theory. The meaning of words, the relationship of words to concepts, and how concepts are structured in the mind have been disputed since before Aristotle's time. Different considerations have been discussed from *logical, cognitive semantics* and *psychological* point of view.[31]

Relevant questions could be rephrased as follows:

- What are concepts?
- How can concepts be organized?
- How does Hanzi *represent* concept?

Hanzi and "Concept Genes"

The question "what exactly is a concept?" has bothered semanticists for more than two generations. Roughly, concepts embody our knowledge of the world, and we use them to make approximations and as a simplified method of communicating with others. For HG theory, concepts are symbols that are outlines of our understanding of the internal representation of an individual's complex experience. Due to the probabilities involving all of the possible combinations of every individual, one individual's personal experience cannot be identical to that of any other individual.

Human experience is subjective and complicated, for the convenience of communication, concepts are specially defined symbols and information is utilized to represent the related *approximations* of our understanding. Since concepts are mental information units, and we cannot look into our own or others' minds, *concepts* are destined to be subject to speculation. But the

[31]For concept theories in general, the interested readers might consult Smith and Medin (1981) and Murphy (2002).

private nature of concepts does not prevent them from being the basis of communication.

As an organization of neural networks, our brain remembers and processes the signals that are transmitted from the sensory organs. "Remembering" involves taking the phenomena that occurred in a dynamic state of time and recording it separately into a multi-layered static structure. "Processing" means restoring the recorded static structures to resemble the original dynamic phenomena, through the connections of the neural network within the brain. During the thinking process, we utilize these representative *symbols* to access the *interface* of our memory network. Once accessed, we expand the factors of our comprehended experiences one after another.[32]

Looking into the recent development in cognitive science, HG theory's theoretical specification shares many hypotheses with the *holistic* approach within cognitive linguistics. In contrast to the *modular* approach, this view does not regard language as an autonomous cognitive faculty, but as an ability which can be explained by general cognitive principles (Langacker 1987; Croft and Cruse 2004). Thus the basic properties of a language result from general processes of conceptualization which relate to various areas of human experience. Under this assumption, linguistic knowledge - knowledge of meaning and form - is basically a *conceptual structure*. Linguistic phenomena, then, could be explained by general conceptual principles.

However, the central issue here concerns the relation between language and cognition in general. For the relation between **writing and cognition**, only rare literatures were appeared. In linking Chinese writing with cognition, Chu philosophized the relation as follows:

> "[...] Throughout their cultural development, Chinese emphasized 象 ("symbols") and ignored 數 ("numbers"). A "symbol" is the microstructure of an idea and can also be called the "connecting point" of

[32]Chu, Discourse 4: Thinking Process.

> a network. Once the neural network of the brain and the conceptual network combine, every "connecting point" serves as the core of a "symbol", allowing us to achieve complete understanding through the expansion of this conceptual network. [...]"

This sets the scene for the representation of *Concept Genes* via Hanzi.

Hanzi and Concept Class Hierarchy

According to the HG theory, concepts should be ordered according to some conceptual classification scheme and presented in a systematic structure. But how does one elicit, organize and explore hierarchical relations of concepts? There have been many answers to this question proposed in philosophy, artificial intelligence and database design. In the following, as a background knowledge, we will at first clarify the difference between the considerations, then discuss the proposal of HG theory.

- Kinds of Concept Organisations

In the study of concept organization, we can find many ways in which an object can be categorized. In general, the basic building blocks of concept organization are *concepts* and *relations*. Concepts can represent either *types* or *roles*. The basic difference between these two is that types are said to be semantically rigid, i.e. their instances are such that they always belong to the type, while an instance's roles may change. For example, a person's gender is a type because it *cannot* change during the lifetime of an individual. On the other hand, a student is an individual's role as she/he ceases to be a student when she/he graduates (Guarino 1998). Generally, the notion "type" is central to concept organization.

Types can be organized into different kinds of **concept type hierarchies**. The most common concept type structures used in computational linguistics

are tree, and various lattices and semi-lattices,[33] if the types are ordered by a *partial order* (also called the *subtype relation*).[34]

Lattice structure is a kind of *closed hierarchy structure*, which has exactly one top-concept and one bottom-concept. In contrast, both a *tree* and a *semi-lattice* structure are kinds of *semi-closed hierarchies*, which can be top-closed or bottom-closed. That is, in a lattice-structure concept type hierarchy, any two concepts (types) can have common subtypes (so-called *the greatest lower bound*) and common supertypes (*the least upper bound*), while in a tree-structure concept type hierarchy, the structure is restricted so that any two concepts necessarily have a common supertype, but they have no common subtype.

In a lattice-structure concept type hierarchy, there are two types that always exist. The Entity (or "the universal type"), and the Absurdity (or "the absurd type"). The Entity type is the type that is a supertype of every other type in the type hierarchy, while the Absurdity type, being the lowest element, is a subtype of every other type in the type hierarchy, and nothing exists which is an instance of Absurdity.[35]

In a tree and semi-lattice-structure concept type hierarchy, take the *top-closed hierarchy* for example, which is a concept hierarchy where there is one top-concept and several bottom-concepts. In this kind of structure, any two concepts are top-connected. There can be some bottom-connected concepts with each other. In this case, the structure is a semi-lattice. If any two

[33]Their formal descriptions will be given in the next chapter, and some core ideas in the realm of ontology construction in the current NLP systems will be discussed in chapter 5.

[34]Generally, we might say that a set of concepts and their binary relations form a special kind of network called *hierarchical structure*, in which nodes (concepts) connected by the hyponym-hypernym relation (commonly called the IS-A relation). Although the hierarchical structure appears to be a universal property of all cultures' categories of the natural world, how exactly it is mentally represented is still not clear (Murphy 2002). However, the controversial psychological status of the hierarchical structure is not the main concern here.

[35]The reason why we need an Absurdity is that it makes for certain theoretical conveniences, which are deeply rooted in lattice theory.

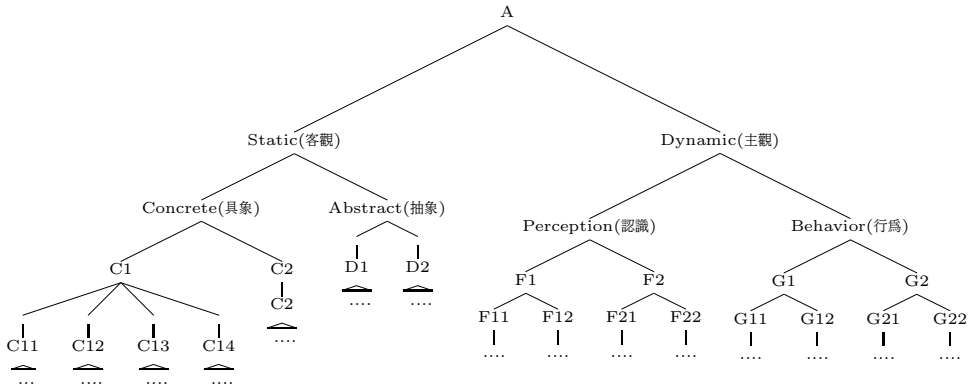


Table 3.2: Chu's tree-structured conceptual hierarchy (truncated for brevity)

concepts are not bottom-connected, then the concept structure is a tree.[36]

Another building block of concept organisation is the set of relations, which are used to connect one or more concepts types to each other. In principle, there are no restrictions concerning the arity of these relations, but in general, uniary and binary relations are adequate enough to build *ontology*. If desirable, relations of greater arity can be expressed simply by means of an additional concept representing the relation itself. An important idea related to the lattice-structure type hierarchy is the *multiple inheritance*. That is, a type may be inherited from more than one type.

In the following, we will introduce the tree-structured concept hierarchy proposed by HG theory, as well as the underpinned philosophical consideration.

Chu's proposal for the concept type hierarchy is based on both *binary* and *quaternary* classification methods (Table 3.2). The binary classification method, he claims, is the simplest and most effective tool for conceptual data analysis. Its earliest recorded use was in an ancient work of Chinese philosophy, *The Book of Changes*, which was the origin of binary numerals.

[36]Sowa's type hierarchy (Sowa 1984:80) is an example of a lattice. An example of a semi-lattice-structure concept hierarchy can be found in one axiomatization of Kauppi's (1967) concept theory.

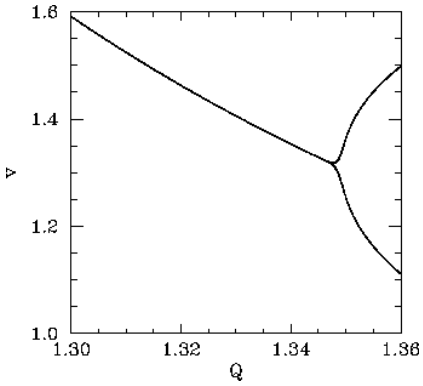


Figure 3.6: First period-doubling bifurcation

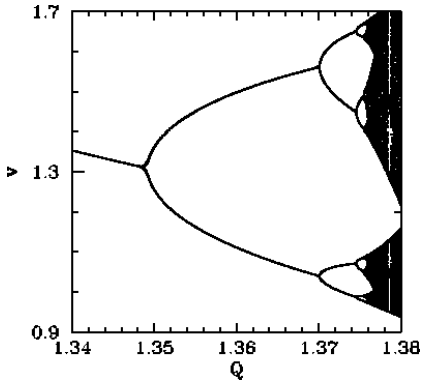


Figure 3.7: Second period-doubling bifurcation and third bifuration

To observe any phenomenon, the sensory organs must perceive stimulae and compare the first stimulus they receive with the subsequent one; the observer will derive an idea of the opposites forming this phenomena.[37]. This is the principle of the binary method: A classification based on opposite phenomena. All of the ideas of one category of phenomena can be used to obtain phenomena of an opposite nature.

Chu presented a system of concept classification that illustrates an application of the binary method. Used to systematically classify Chinese character concepts, the binary method can produce a *"conceptual network"*. However, Chu argues, the binary classification method has its limits. When concepts are classified at the *third* level, they lose their opposing nature. He illustrates such thinking by resorting to the findings of modern Chaos Theory, which posits that after *third period doubling bifurcation*, everything becomes a random value. In Figures 3.6 and 3.7, we can see that the second bifurcation converts a period-2 attractor into a period-4 attractor (hence, two curves split apart to give four curves). Likewise, the third bifurcation converts a period-4 attractor into a period-8 attractor (hence, four curves split into eight curves). Shortly after the third bifurcation, the various curves in

[37]For detailed discussion please refer to "Discourse 6 – Cognition"

the figure seem to expand explosively and merge together to produce an area of almost solid black. This behaviour is indicative of the onset of *chaos*.

Based on this classification scheme, Chu proposes his conceptual hierarchy as follows:

The first criteria in classifying concepts relates to the standpoint of the observer, which corresponds to the first dimension, and in this case involves the basic distinction between the `objective` world and the `subjective` one. The term `objective` corresponds to the "entities" of the static state, and the term `subjective` correspond to the "applications" of the dynamic state. The second dimension refers to "subjects observed"; the third dimension relates to "understanding". These can each be divided into two categories.

The objective domain includes the `abstract` sub-domain of ideas and the `concrete` sub-domain of material things. The `abstract` domain generates two categories of ideas: the first category is `definitions` derived from the understanding process; the second category is `applications` originating from human needs. In the second sub-domain called `concrete`, things can be classified either as material and existing in the `natural` world or as `artificial` (in the sense of man-made). The `subjective` exists in the minds of humans; it starts from the outside and moves inside; it belongs to the domain of `perception`; another example of the `subjective`, but which starts from the inside and moves towards the exterior, is `behavior`. `Perception` is divided into two categories: `sensory`, which includes the various stages of the perception process; and `states`, which includes the circumstances that exist after events occur. `Behavior` can be divided into the inherent `basic instinct` and `social behavior` acquired during life.

If we continue to the classify at deeper levels, we have gone beyond three levels, and should not feel constrained to persist using the binary classification method. A system based on four categories is now ideal, and these four can be divided into eight categories. *The Book of Changes* also follows this structure.

It is noted here that, like the *taxonomic thesaurus*, the only binary relations between the nodes are specified by two arcs, namely, the IS-A relation.

- Conceptual Encoding of Chinese Characters

Now, the last question to be answered is: How do Hanzi represent concepts?

HG theory proposes that in such a tree-like structure, all characters are positioned at the leaf level. Characters found in the same node are assumed to carry similar conceptual information, and a systematic approach to represent conceptual information is by selecting characters in the *binary* alphabet $\{0, 1\}$.

In code theory, it is convenient to use words of the same length in the message transmission. If there are 2^n binary words of length n, then the complete set of them is usually denoted by $\mathcal{V}^n$; for example,

$$\mathcal{V}^3 = \{000, 100, 010, 001, 110, 101, 011, 111\}$$

As known, each symbol in such a word is called a *bit* (an abbreviation for binary digit). A binary *code* of *length* n is simply a subset C of $\mathcal{V}^n$. In the case of Hanzi encoding, we have 2^8 binary *characters* (concept types) of length 8, that is,

$$\mathcal{V}^8 = \overbrace{\{00000000, 00000001, 00000010, 00000011, ..., 11111111\}}^{256}$$

Table 3.3 shows some examples of these characters.

Take 高 (/gāo/, "high") for example, in this box, we can interpret the characters as follows: Its concept belongs to the subjective domain (1), is the effect or product of perceiving (10); Of or relating to the senses or sensation (100), and transmitting impulses from sense organs to nerve centers. It is something which can be seen (10000), and distinguished from its surroundings by its definite, distinctive form (10000101).

Table 3.3: A self-synchronizing code of Chinese characters

characters	codeword	characters	codeword
火 (fire)	00000000	天 (sky)	01011000
禾 (grain)	00001000	語 (language)	01100000
骨 (bone)	00010001	法 (law)	01110000
衣 (clothes)	00100110	漢 (man)	01111100
城 (town)	00101000	民 (folk)	01111110
桌 (desk)	00110110	憂 (worry)	10010110
類 (genus)	01000000	古 (ancient)	10101000
定 (stable)	01001000	勝 (can bear)	10111110
宙 (infinite time)	01010000	行 (go)	11001000

```
高 1 0 0 0 0 1 0 1
1 :  Subjective domain
1 0 :  Perception
1 0 0 :  Sensory
1 0 0 0 0 :  Vision
1 0 0 0 0 1 0 1 :  Shape
```

Take as another example, 語 (/yǔ/, "language"). It is arranged in such a position that we can interpret it as follows: It is classified as belonging to the objective domain (0); considered apart from concrete existence (01); it is something applied to a special use or purpose (011); the substance of a communication (01100); and a collection of knowledge about a specific event or situation (01100000).

```
語 0 1 1 0 0 0 0 0
0 :  Objective domain
01 :  Abstract domain
011 :  Applications
01100 :  Message
01100000 :  Information
```

To make it useful for real-world applications, Chu goes further in proposing that a character - as an analogy for a *chromosome* - can be encoded with a 32-bit long binary string of *DNA sequences* representing four different *genes*, including information about a concept, common sense, symbolic and properties, respectively. A sequence of the character chromosome layout is shown in Figure 3.8. The position of genes with their associated codings are

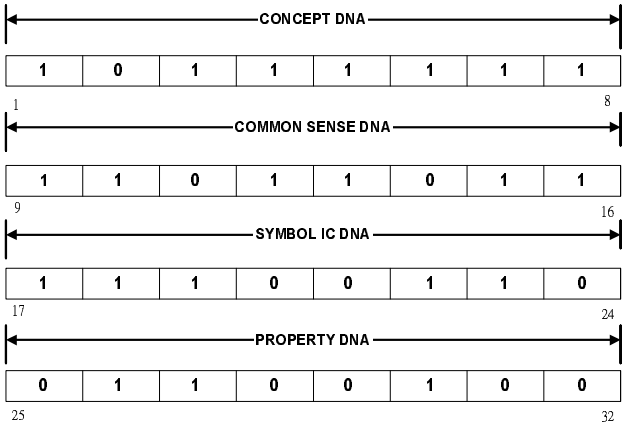


Figure 3.8: A complete code definition of a character

listed below:

- `bits 1 - 8` (CONCEPT DNA): The first eight bits in the character chromosome are used to represent conceptual information. A total of 256 different concept types can be selected.

- `bits 9 - 16` (COMMON SENSE DNA): These eight bits are assigned to the representation of common sense knowledge. A total of 256 different component types (CH + CB) are available.

- `bits 17 - 24` (SYMBOLIC DNA): In order to provide a wide range of applications, there are 256 types of symbolic DNA concerning mainly the proper names are designed. However, this information has not been made public.

- `bits 25 -32` (PROPERTY DNA): These are the descriptions of the features of the character in question, at the moment, this information is not accessible to the public either.

Take the character 簫 (/xiāo/, "China flute") for example. Its concept code is 00111010. By searching the concept hierarchy, from the first bit 0, we know that it is static, and belongs to the objective external which can be *used* by humans. The second bit 0 depicts that it is a concrete thing with a body, form, color, quality and quantity which can be *recognized* by humans. The third bit 1 means that it is an artifact, having a function and a geometric form. It is something that people can *know*. The fourth and fifth bits mean that it relates to human work. We know that, from the last three bits, it is an instrument.

The common sense gene of this character[38] is represented by CH (竹) and CB (肅), which depict the common sense knowledge concerning with this character: it is solid, straight, and tube-like, and so on.

By encoding information in this way, according to HG theory, the *complete* definition of conceptual and semantic information of *each* character could be obtained.

3.4.4 Critical Remarks

In sum, this section has sketched some of the fundamental issues of Chinese characters under dispute in the context of contemporary Hanzi studies, in particular, in regard to the HG theory. As a theory proposed and expounded by a non-linguist, HG theory does not proclaim it adhering to any current linguistic theories. As the founder claims, in fact, this theory could be seen as continuing the traditional lore of Chinese scriptology with some "technical" terms.

For a scientific theory, it would be premature to impose the requirement of understanding solely based on specific tradition. In this final subsection, I would like to make some remarks from a linguistic viewpoint. In the process I hope to clarify to some extent how HG theory's position relates to other

[38] Unfortunately, instead of a code, only a description is given.

approaches to semantic and conceptual studies, and particularly, how it bears on certain issues such as *conceptual primitives*, *the relation between (word) meaning and concept*, and the *conceptual hierarchy* in the context of modern Mandarin Chinese.

Conceptual Primitives

At first sight, HG theory could be located in the approach of *Primitive Decomposition*. In particular, it resembles the **semantic primitives** approach by Wierzbicka (1996), which in recent times has come to be known as the "Natural Semantic Metalanguage" (NSM) approach to semantic analysis.

NSM theory starts with some assumptions that language must have an irreducible semantic primitives, and the semantic analysis must be conducted in natural language, rather than in terms of technical formalisms (abstract features, logical symbols, etc.); [39] And according to Goddard and Wierzbicka (2002), a good exponent of a primitive meaning may be a word, phraseme, a bound morpheme, or other linguistic expressions, just so long as it expresses the requisite meaning.

In some respects NSM approach are compatible with HG theory. For example, about 60 firm candidates for semantic primitives have been proposed so far (Wierzbicka 1996), and interestingly, all of the *classes* (such as Substances, Determiners, Mental predicates, Existence, etc) that these semantic

[39]From a broad perspective of theorectical development, the search for "conceptual/semantic primitives" might be traced back to Roger Schank's landmark work in the early 1970s. Schank tried to find the *primitive symbols* that one might use to represent the meaning of what we perceive (that is, Conceptual Dependency), and tried to define the processes that could interpret our understanding of sentences (Conceptual Analysis). A parser under this scheme means something that associate a linguistic input with what Schank called a *conceptual structure*. During the ensuing years, several theories for representing conceptual structures have been developed. For example, the Lexical Conceptual Structure (LCS) theory proposed by Jackendoff, also claims that there is a single set of universal conceptual primitives which might manifest itself in both lexical and grammatical domain. But in Jackendoff's view, the "primitives" are 'abstract' and not to be identified with ordinary word-meanings, and he believes that these "primitives" are non-exhaustive, which are different from NSM and HG theory discussed here. See Goddard (1998).

primitives belong, can be found in conceptual hierarchy of HG theory.

There are also differences. For HG theory, it would be reasonable to identify Chinese characters as the proper exponents of conceptual primitives. This idea has not been realized in the NSM research, even in the search of conceptual primitives in Chinese language. Another crucial difference between HG theory and NSM theory lies in that the conceptual primitives in HG theory constitute a hierarchical system, where every conceptual primitive is systematically organized.

Though widely well known and with abundant literatures in linguistics, *Primitives Decompositional approach* has suffered from many criticisms. Anti-primitives auguments are drawn from a broad range of sources. However, as Goddard (1998) comments, it is more difficult to characterize what would count as a good argument against the entire concept of *Primitives Decompositional approach*, but when we settle down to a detailed discussion, a good argument against the status of any particular element as a semantic primitive is, an argument that the element in question is either definable (not the basic), or language-specific, or both. Since HG theory does not claim to be able to apply to other languages, a cross-linguistic validation would not be necessary here. But like NSM theory, HG theory has worked with a specific set of semantic primitives (*concept genes* and *meaning genes*), which are vulnerable by empirical disconfirmation. To make this point more concrete, in the following, some examples will be adduced.

For the part of *concept genes*, HG theory made two presumptions: (1), there are a limited number of basic concept types (called "concept genes"), and (2) Hanzi could be regarded as the *instances* of these concept types. These might need more empirical surveys to testify. But how this theorectical setting can be in harmony with the problem of "meaning variants" such as *homonomy*, *polysemy* and *metaphor*, which might be one of the most intractable issues in primitives decompsition approach, explanations are unfortunately not given in public.

In addition, the relation of character and meaning is not always inherent. Considering the *onomatopoeia* example, whic is to be understood here as linguistic sounds being used to copy sounds in nature or in the human world. In the English lexicon, we have onomatopoeia like "drip and drop" or "splash and splotch". In Chinese, 嘰哩咕嚕 (jiligulu, "talking indistinctly"), 幾哩括拉 (jiliguala, "talking loudly") are good examples. In some casese, onomatopoeia is not only employed for the imitation of natural sounds, e.g. 喔喔喔 (/wowowo/, "cock-a-doodle-doo"), but also for the coinage of names for sound-related entities: e.g. 布穀鳥 (/buguniau/, "cuckoo") is supposed to be a sound representation of that bird.[40] For the characters in these examples of onomatopoeia, it would be improper to assert that they denote to concepts.

For the *meaning genes*, HG theory asserts that there are only a limited number of basic character components, which constitute a set of basic meaning units; and every character can be decomposed into two components (namely, two meaning genes): *Character Head* and *Character Body*. But in some cases, the criteria of decomposition is not clear at al. E.g., why the character 貳 is decomposed to 二貝 and 戈, instead of 貝 and 二戈.

Another criticism might go to the the general advocacy that "self-completeness" of *meaning composition via CH and CB* within a character. When we say that a conceptual system has primitives, we usually implies the principle of semantic composition to a certain extent, which build up more complex concepts from less complex ones (Lakoff 1987:279). Similar to the principle of semantic composition, HG theoy presumes the 會意法 (the picto-synthetic principle)[41] as the main semantic composition among *meaning genes*. Though by resorting to the classical argumentation of traditional scriptology concerning the correlation between meaning composition within a character, it would be relatively speculative to conclude that all characters in modern use are bound to the principle alone. In HG dictionary, examples are not difficult to

[40]These examples are taken from Yip (2000).

[41]Please refer to the explanation in section 3.2.1

enumerate. E.g., why the meaning of 掀 (lift) can be *induced* from the meanings of its two components 手 (hand) and 欣 (glad); why the meaning of 揶 (ridicule) can be *induced* from the meanings of its two components 手 (hand) and 耶 (question particle), .. and so on. For such examples, notwithstanding a far-fetched explanation offered by HG theory, it would be hard to assert any inherent semantic composition principle within character components.

Conceptual or Semantic?

As a proposed general theory of concept, one crucial deficiency of the HG theory might lie in that, the relation between (word) meanings and concepts, especially in the context of Chinese writing, is shunted aside. The term "semantic" and "conceptual" are used interchangeable throughout the HG works.

Indeed, drawing clear-out distinctions between meanings, concepts and their linguistic expressions (be they words or characters) is not an easy task because they are so intimately interwoven. This is also an enormous topic that has been attracting researchers for a long time. Philosophers, psychologists and linguists have argued as to whether there is an abstract layer of concepts which is separate from word meaning or whether the word meanings and the concepts are identical (Aitchison 2003).

In my opinion, HG theory seems to resemble the view that semantic primes can be thought of as linguistically embodied conceptual primes (Wierzbicka 1996; Goddard 1998), and thus semantic analysis is by its nature a conceptual inquiry. Throughout the current available version of HG theory, it does not provide discussions about linguistic meanings from (referential) semantics, which argues that words get their meanings by *referring* to real objects and events, but rather places great weight on a **conceptual view of meaning** from the *cognitive* perspective.

For instance, the cognitive psychological approach assumes that we have some sort of mental description that allows us to pick out examples of the

word and to understand it when we heard it. Murphy (2002) claims that *word meanings* are psychologically represented by mapping words onto *conceptual structures*. Namely, a word gets its significance by being connected to a concept or a coherent structure in our conceptual representation of the world. This resembles Langacker's model which regards the meaning of a word as an *access node* into the knowledge network (Langacker 1987:163).

Murphy (2002:391) suggests three principles for this conceptual view of word meaning, which might be in accord with the position of HG theory. The three principles are quoted as follows:

1. Word meanings are made up of pieces of conceptual structure,

2. An unambiguous word must pick out a coherent substructure within conceptual knowledge (while ambiguous words choose n coherent structures, one for each of their n meanings).

3. When an unambiguous word has multiple related senses, which is called *polysemy* in lexical semantics, the meanings are overlapping or related conceptual structures. For instance, the two senses of *foot* in the sentence "We are standing at the *foot* of the mountains", and in "One of her shoes felt too tight for her *foot*" are related by both containing conceptual information about "at the bottom part".

He argues further that important parts of the psychology of word meaning can be easily explained by referring to the psychology of concepts following from these principles. That is, principles of concept use carry over to become principles of word meaning. In addition to this, there are two corollaries that follow from these principles that are important as well.

- First, semantic content entails conceptual content.
  Namely, if a word we know means something, that something must be part of our conceptual structure.

- Second, no semantic distinctions can be made that are not distinguished in conceptual structure.

  For example, we couldn't distinguish the words *chair* and *stool* if we didn't perceive the difference between these kinds of things and have that difference represented in our concepts of furniture.

This been said, linguistic complexities are more elaborate and rich in detail. Aside from the **polysemy** and **ambiguity** phenomena, it has always been a difficult problem for linguistic theory as well as lexicographic practice, as to what criteria we should set in defining **homonymy** (or homographs). Surely, such linguistic complexity leads to more specific psycholinguistic discussions and models, which aim to bridge the gap between the static representation of words in the head and the dynamic process of comprehension.[42]

In dealing with problem of ambiguity, HG theory does not oversimplify matters by assuming that characters are associated with only single concept type. In its design, a character can be assign to more than one (the maximal number is four) concept type. Nevertheless, HG theory does not provide a convincing criteria or linguistic consideration in the *assignment* of concept type, when characters (i.e., exponents of "concept genes") have secondary, or polysemic *conceptual meanings*?

Conceptual Hierarchy vs. Ontology

Now we draw to the last point. On the whole, we would agree that, giving a set of concepts and a set of relations, associating them with each other remains a *subjective* process. The quality and quantity of hierarchically organized ontologies rather depends on the author's hand-crafted guidelines, and on her/his interpretation of these guidelines. In the Hanzi-driven conceptual hierarchy, HG theory relies mostly on Chu's personal philosophical introspection in presupposing a realm of concept types as abstract entities.

[42]The interested reader is referred to Murphy (2002).

The *psychological reality* of these concept types needs experimental proof. Many characters do not fit neatly into an ontological hierarchy at all, while others fit rather well (or badly) at many places.

In fact, this involves the long-standing problem of the *relativity* and *subjectivity* of conceptual classification. As we know, concepts are inventions of the human mind used to construct models of the world. Wong (2004) points out some examples of arbitrariness in conceptual classification observed in several existing lexical databases like WordNet 1.5 and EuroWordNet 2. However, as Sowa (1984:339) claimed, the world is a *continuum* and concepts are *discrete*, therefore a network of concepts can *never* be a perfect model of the world. At best, a kind of workable approximation is desired. As far as Chinese characters are concerned, Wong (2004) argues, the semantic relatedness displayed by Chinese characters provides a means to concept classification which might be more *objective*, more explicit and, hence, easier to capture.

In addition, in many aspects, conceptual hierarchy of HG theory parallels the *ontology* in the recent development of NLP and knowledge information processing. But, the reason why only the *INSTANCE-OF* (i.e., certain characters are instances of certain concept types) and *IS-A relations* (i.e., certain concept type is a kind of certain concept type) are permitted in the conceptual hierarchy is not clearly exposited either, as other kinds of links could specify properties known about each concept, and still other connections might be made between related concepts or properties. It would be more convincing and interesting to compare how the conceptual hierarchy of HG theory fit with other proposed ontologies using the existing tools and algorithms concerning with ontology mapping, ontology alignment, consistency checking, .. etc, which have been proposed.

To conclude this brief remark, a number of points about the theorectical conception of HG theory deserve to be reiterated.

In many cases, the interpretation of some of the data presented in the Hanzi Gene Dictionary and the application of the principles governing charac-

ter formation to the analysis of specific characters turn out to be less than clear-out. In addition, in many places, there are only "blueprint" available,[43] and the lack of enough empirical data also results in misinterpretations of this theory.

To be fair to HG theory, regardless of the problems discussed above, and though within the framework of HG theory, many ideas are originatedlly drawn from a broad range of sources, treating Chinese characters as conceptual primitives, and presuming a tree-like hierarchy for the representation of these conceptual primitives are new tries in the field of Chinese NLP as well as Hanziology. In my opinion, with more theorectical refinements discussed above, introducing ontology as a locus for establishing a rich set of conceptual primitives could be a work serving as a testbed to get a better grip on Chinese language processing in general, and on the other side, as a remedy to the arbitrariness in the design of knowledge representation. In addition, it would also be interesting to see how different lexical knowledge sources come together to signify the value of Hanzi in use.

Having outlined the discussion of Hanzi studies in this chapter, we can now cover some insights into a working model of natural language processing resources which will serve as a basis for our discussion henceforth. In the next chapter some formal models will be discussed.

[43]For example, only the 256 concept genes are available, other kinds of "genes" (common sense, symbolic and properties) have not been published due to the commercial reasons.

Chapter 4

Mathematical Description

In the previous chapter, we reviewed some of the fundamental issues in the study of Hanzi both from the linguistic and hanziological viewpoints, especially with regard to the conceptual/semantic information of Chinese characters. In this chapter, I would like to review some formal descriptions of them. There are a number of different formal ways to characterize Hanzi, and different characterizations have led to different models. Our main focus is on choosing the appropriate model to represent the conceptual and common sense knowledge "wired" in Hanzi. A comprehensive survey of the mathematical theory of Hanzi is therefore beyond the scope of this chapter. Some models having a bearing on the notions discussed in previous chapters will be chosen, and mathematical preliminaries will be provided as needed, in order to facilitate understanding the models we will discuss. The final section lists and expounds on possible answers to the major questions, with a proposed tentative model which aims at describing the semantic and conceptual structure of Hanzi.

4.1 Introduction

The first to be cited is that of the **generative-grammar-based model** (Wang 1983). One of Wang's goals was to construct a *grammar* of the structure and writing order of Chinese characters. Only within the generative grammar framework, he claimed, can we give an descriptively adequate account of the native Chinese's writer's intuitive knowledge of her/his writing system. He developed a procedure for predicting the stroke order, namely, the relative placement of semantic "classifiers" and phonetic "specifiers" of Chinese charcters within the framework of generative grammar. However, what he called by *intuitive knowledge* was mainly concerns with shape structure, and due to the lack of systematic explanations for the conceptual and semantic information carried by Hanzi, this model can be only applied to the task of character recognition.

In the following, some basic terms are defined, which are necessary to understand the proposed *formal language* models of Chinese characters.

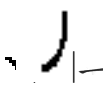
Definition 4.1.1. *(**Symbol, Alphabet, Strings, Formal Grammars and Languages**)*
A string *is a finite sequence of elements. The elements are typically called* symbols*, and the set of all symbols under consideration, including the empty string* Λ*, is denoted* Σ*. The set of symbols may also be called an* alphabet *especially when each symbol is a single letter, even in the case of Chinese characters. And the term* formal language *is simply defined as a set of strings. The formal rules which are needed to specify a language, i.e., to produce legal strings of a language are called the* (formal) grammar *of that language.*

For example, if the *alphabet* is Σ = { 不, 快, 樂 }, then the instances of "strings over the alphabet Σ " can be: Λ(the empty string), 快樂, 不樂, 快樂不快樂 and so on.[1] And $\mathcal{L}$ = { 快樂, 不樂, 快樂不快樂 } is a language of three

[1]The principal operation on strings is called *concatenation*. The concatenation of

strings, that is, $|\mathcal{L}| = 3$. With these basic notions, we can then propose a 4-tuple Grammar to define the *formal language* of Chinese characters.

Definition 4.1.2. *(**Formal Grammar of Chinese characters**)*
A formal grammar of Chinese characters is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{T}, \mathcal{S}, \mathcal{P})$, *where* $\mathcal{V}$ *and* $\mathcal{T}$ *are finite sets with* $\mathcal{V} \bigcap \mathcal{T} \neq \phi$: $\mathcal{V}$ *is the set of nonterminal symbols, and* $\mathcal{T}$ *is the set of terminal symbols;* $\mathcal{S} \in \mathcal{V}$ *is the start symbols; and* $\mathcal{P}$ *is a finite set of productions, or grammar rules. They are expressed in BNF as follows:*

$\mathcal{T}$ = {  | ... }

$\mathcal{V} = \{<Character> | < Component > | < Stroke >\}$

$\mathcal{S} = \{<Character>\}$

$\mathcal{P} = \{P_1 | P_2 | P_3\}$

$P_1 : <Character> := <Stroke> | <Component>$

$P_2 : <Component> := <Stroke> | <Component>$

$P_3 : <Stroke> :$  | ...

In a similar manner, Feng (1994) also proposed 15 tree-style construction types for the analysis of Hanzi.

4.2 The Finite-State Automata and Transducer Model

In moving to the more general level of the language model, discussions in this section are directed not only toward the modelling language, but also toward two abstract machines called *automata* and *transducers*.

strings x and y, usually denoted xy, is a string consisting of the characters of x followed by the character of y.

The formal properties of the mapping between linguistic information and orthography have been rarely explored. Until recently, Sproat's pioneering work (Sproat 2000) in this field of computational theory of writing system is a breakthrough. In his book, Sproat asserts that there must be *some* reasonable relation between the writing itself and the linguistic information it encodes.

One of the central claims in his formal theory of writing concerns the *regularity*: $M_{ORL \longrightarrow \Gamma}$ *is a regular relation.*[2] Informally, it states that the mapping relation between linguistic and orthographic elements is *regular*. In the technical sense, $M_{ORL \longrightarrow \Gamma}$ can be implemented as the finite-state transducer.

4.2.1 Finite-State Techniques: An Overview

Finite-state methods have been used extensively in language research. For the sake of simplicity, the following gives some basic notions mostly modeled on those of Sproat (Sproat 1992;2000), including a brief description of *regular languages and relations*, and their associated abstract machines, namely, *finite-state automata* (FSAs), and *finite-state transducers* (FSTs).[3]

Definition 4.2.1. *(**Regular Grammars**)*
A regular language is one in which every production rule conforms to one of the following patterns:

$$X \rightarrow xY \qquad X \rightarrow y$$

[2]Sproat (2000:14) introduced the notion of the Orthographically Relevant Level (ORL) as being the level of linguistic representation encoded orthographically by a particular writing system. In addition, he denotes the output of the mapping from the ORL to spelling itself as Γ. Note that the notion of the *Orthographically Relevant Level* proposed here is not the same as the notion of the *orthographic depth hypothesis* discussed earlier, since the former relates to how abstract the encoded phonological information is, not the regularity of the encoding.

[3]For general introduction to the theory of automata, please see Hopcroft and Ullman (2000), Introduction to Automata Theory, Languages, and Computation; further discussion in detail about finite-state transducer please consult Mohri (1997, 2000) and so on.

where X and Y are each single non-terminals, x is a terminal, and y is either the empty string (ε), or a single terminal.

It is common to define a *regular language* using a recursive definition such as the following:

Definition 4.2.2. *(**Regular Languages and their Closure Properties**)*

1. *$\emptyset$ is a regular language.*
2. *For all symbols $a \in \Sigma \cup \Lambda$, $\{a\}$ is a regular language.*
3. *If $\mathcal{L}_1, \mathcal{L}_2$, and $\mathcal{L}$ are regular languages, then so are*
   (a) *$\mathcal{L}_1, \mathcal{L}_2$, the* concatenation *of $\mathcal{L}_1$ and $\mathcal{L}_2$: for every $w_1 \in \mathcal{L}_1$ and $w_2 \in \mathcal{L}_2$, $w_1 w_2 \in \mathcal{L}_1 \cdot \mathcal{L}_2$;*
   (b) *$\mathcal{L}_1 \cup \mathcal{L}_2$, the* union *of L_1 and L_2;*
   (c) *$\mathcal{L}^*$, the* Kleen closure *of $\mathcal{L}$. Using $\mathcal{L}^i$ to denote $\mathcal{L}$ concatenated with itself i times, $\mathcal{L}^* = \bigcup_{i=0}^{\infty} \mathcal{L}^i$.*

As seen, regular languages can be constructed from an alphabet of symbols using only the operations of concetenation ($\cdot$), ($\cup$) and ($*$). While the above definition is complete, regular languages observe additional *closure* properties:

- Intersection: If $\mathcal{L}_1$ and $\mathcal{L}_2$ are regular languages, then so is $\mathcal{L}_1 \bigcap \mathcal{L}_2$.
- Difference: If $\mathcal{L}_1$ and $\mathcal{L}_2$ are regular languages, then so is $\mathcal{L}_1 - \mathcal{L}_2$, the set of strings in $\mathcal{L}_1$ that are not in $\mathcal{L}_2$.
- Complementation: If $\mathcal{L}$ is a regular language, then so is $\Sigma^* - \mathcal{L}$, the set of all strings over Σ that are *not* in $\mathcal{L}$.(Of course, complementation is merely a special case of difference).

- Reversal: If $\mathcal{L}$ is a regular language, then so is $Rev(\mathcal{L})$, the set of reversals of all strings in $\mathcal{L}$.

RE are set of strings, and they are usually notated using *regular expressions*. A fundamental result of automata theory are the so-called *Kleene's theorems*, which states that regular expressions are equivalent to FSA. This can be defined as follows:

Definition 4.2.3. *(**Finite-state automata (FSAs)**)*
A FSA, $\mathcal{M}$, is a quintuple, $(\mathcal{Q}, \Sigma, q_0, \delta, \mathcal{A})$, where

- *$\mathcal{Q}$ is a finite set of states,*
- *Σ is a finite set of symbols,*
- *$q_0 \in \mathcal{Q}$, where q_0 is the start state,*
- *$\mathcal{A} \subseteq \mathcal{Q}$, where $\mathcal{A}$ is the set of accepting states, and*
- *$\delta : \mathcal{Q} \times \Sigma \rightarrow \mathcal{Q}$.*

A FSA has a finite numbers of states and has no other form of memory; this is why it is called "finite". Alternatively, we can also define regular languages in terms of FSAs: A language $\mathcal{L}$ is a regular language iff there exists an FA $\mathcal{M}$ such that $\mathcal{L} = \mathcal{L}(\mathcal{M})$.

Definition 4.2.4. *(**Regular relations**)*
Regular n-relations can be defined in a way entirely parallel to regular languages.

1. *$\emptyset$ is a regular n-relation.*
2. *For all symbols $a \in [(\Sigma \bigcup \Lambda) \times \ldots \times (\Sigma \cup \Lambda)]$, $\{a\}$ is a regular n-relation.*
3. *If $\mathcal{R}_1, \mathcal{R}_2$, and $\mathcal{R}$ are regular n-relations, then so are*

(a) $\mathcal{R}_1 \cdot \mathcal{R}_2$, *the (n-way) concatenation of* $\mathcal{R}_1$ *and* $\mathcal{R}_2$*: for every* $r_1 \in \mathcal{R}_1$ *and* $r_2 \in \mathcal{R}_2, r_1 r_2 \in \mathcal{R}_1 \cdot \mathcal{R}_2$;

(b) $\mathcal{R}_1 \cup \mathcal{R}_2$;

(c) $\mathcal{R}^*$, *the n-way Kleene closure of* $\mathcal{R}$.

As seen, regular relations can be constructed from an alphabet *pair of* symbols using only the operations of concatenation$(\cdot)$, $(\cup)$ and $(*)$. These are implemented with finite-state transducers. We define them thus as follows:

Definition 4.2.5. *(**Finite-State Transducer**)*
A FST is a 6-tuple $(\sum_1, \sum_2, \mathcal{Q}, i, \mathcal{F}, \mathcal{E})$ *where:*

- $\sum_1$ *is a finite alphabet, called the input alphabet.*
- $\sum_2$ *is a finite alphabet, called the output alphabet.*
- $\mathcal{Q}$ *is a finite set of states.*
- $i \in \mathcal{Q}$ *is the initial state.*
- $\mathcal{F} \subset \mathcal{Q}$ *is the set of final state.*
- $\mathcal{E} \subset \mathcal{Q} \times \sum_1^* \times \sum_2^* \times \mathcal{Q}$ *is the set of edges.*[4].

A finite-state transducer (FST hereafter) can be seen as a FSA with symbol pairs as labels for each arc. However, with success in application to the word segmentation (Sproat and Shih 2001), such grammar formalisms neglect the problem of the topological structure of Chinese characters, that is, the ways in which *graphemes* concatenate.

[4]Note that FSTs may be weighted: They are then referred to as weighted finite state tranducers (WFST's).

4.2.2 Topological Analysis via Planar Finite-State Machines

As previously mentioned, unlike most alphabetic writings such as English, which is predominantly linear, Chinese writing is two-dimensional, both up and down and left to right. In considering this problem, it becomes obvious that the usual notion of a regular language, where the catenation operator '·' denotes simple left-to-right concatenation, will not suffice here. Sproat thus proposes a more powerful notion: *planar grammars*.[5] To put it simply, planar (or "two-dimensional") languages and relations differ from string-based regular languages and relations only in the definition of a richer set of concatenation operations.

Let's take an example to illustrate this. Suppose that Chinese characters are a set of two-dimensional figures that can be arranged in a predetermined layout, such as the four rectangles labeled $\gamma(\alpha), \gamma(\beta), \gamma(\zeta)$ and $\gamma(\delta)$ shown in Figure 4.1. By assuming the "stroke ordering principle" taught in the elementary school, we start with the rectangle on the left-hand side, then we say $\gamma(\alpha)$ *left catenates* with $\gamma(\beta)$; then this pair *downwards catenates* with the pair $\gamma(\zeta)\gamma(\delta)$; and $\gamma(\zeta)$ *left catenates* with $\gamma(\delta)$. An example of a character that fits this pattern is 蹦 (/bèng/,"leap"), which consists of the components, 足, 山, 月, 月, arranged as : 足 $\rightarrow$ [山 $\downarrow$ [月 $\rightarrow$ 月]].

One point which must be noted is that, as Sproat points out, planar catenation operators, unlike those in string-based concatenation, are *not* in general *associative*. The use of brackets in Figure 4.1 is one of the possible solutions to this problem.

Now the formal definition of planar regular language can be given based on Definition 5.2.2, with only one modification: splitting the concatenation operations "·" into five operations. Namely, *Left* $\rightarrow$, *Right* $\leftarrow$, *Downwards* $\downarrow$, *Upwards* $\uparrow$ and *Surrounding* $\odot$ Catenations. The relevant closure property

[5]The definitions and descriptions given here are mostly modeled on those of Sproat (2000), for more in-depth formal treatments on two-dimensional languages please refer to Giammarresi and Restivo (1997).

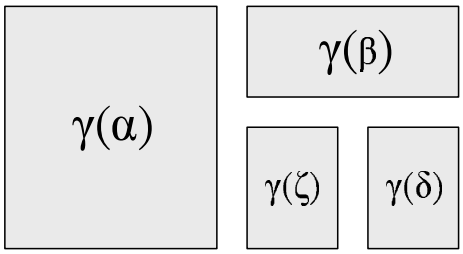


Figure 4.1: One of the topological structures of Chinese characters described by $\gamma(\alpha) \rightarrow [\gamma(\beta) \downarrow [\gamma(\zeta) \rightarrow \gamma(\delta)]]$.

in the Definition 5.2.2 of regular language now only needs to be somewhat modified:

Definition 4.2.6. *(**Planar Regular Languages**)*
3. If $\mathcal{L}_1, \mathcal{L}_2$ are planar regular languages, then so are
(a) $\mathcal{L}_1 \rightarrow \mathcal{L}_2$; $\mathcal{L}_1 \leftarrow \mathcal{L}_2$; $\mathcal{L}_1 \downarrow \mathcal{L}_2$; $\mathcal{L}_1 \uparrow \mathcal{L}_2$; $\mathcal{L}_1 \bigodot \mathcal{L}_2$.

The abstract machines to planar regular languages and relations are *planar finite-state automata and transducers* (2FSA and 2FST), respectively. The 2FSA can be defined along with Definition 5.2.3, by simply adding to the definition a starting position in the planar figure p, a set of directions d, and a set of grouping brackets $\mathcal{B}$.

Definition 4.2.7. *(**Planar finite-state automata**)*
A planar finite-state automata is an octuple $\mathcal{M} = (\mathcal{Q}, q_0, p, d, \mathcal{B}, \mathcal{A}, \Sigma, \delta)$ where:

- *$\mathcal{Q}$ is a finite set of states,*
- *$q_0 \in \mathcal{Q}$, where q_0 is the start state,*
- *p is the starting position for q_0, chosen from the set {left,top,right,bottom},*
- *d is the set of directions {R(ight), L(eft), D(own), U(p), I(nwards) } ,*
- *$\mathcal{B}$ is the set of grouping brackets { [,] },*
- *$\mathcal{A} \subseteq \mathcal{Q}$, where $\mathcal{A}$ is the set of accepting states,*

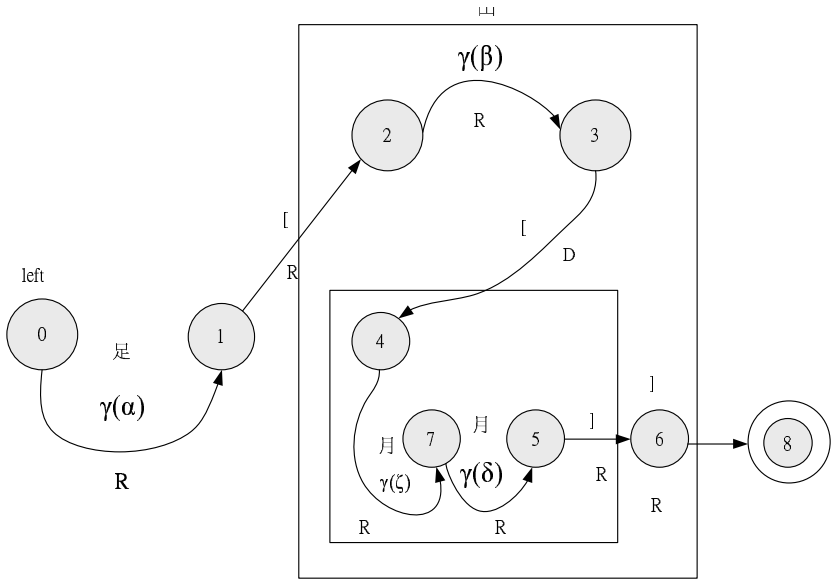


Figure 4.2: A planar FSA that maps the expression $\gamma(\alpha) \rightarrow [\gamma(\beta) \downarrow [\gamma(\zeta) \rightarrow \gamma(\delta)]]$ (the planar figure of "蹦") given in figure 4.1. The labels "R" and "D" on the arcs indicate the recognizing direction (Right and Down); the label "left" on the starting state 0 indicates the position where scanning starts.

- Σ *is a finite set of symbols, and*
- δ *is a transition relation between* $\mathcal{Q} \times (\Sigma \cup \epsilon) \times (\Sigma \cup \epsilon \cup \mathcal{B}) \times$ d *and* $\mathcal{Q}$.

For instance, to recognize the character "蹦"(leap), whose planar figure is shown in Figure 4.1, we need a planar FSA to recognize the description $\gamma(\alpha) \rightarrow [\gamma(\beta) \downarrow [\gamma(\zeta) \rightarrow \gamma(\delta)]]$. That is, at the beginning, the automaton should be able to scan the left-hand side of the figure, then proceeds in reading $\gamma(\zeta)$ rightwards (足), reading rightwards across one grouping bracket, rightwards again across $\gamma(\beta)$ (山), then downwards across one grouping bracket, rightwards across $\gamma(\zeta)$ (月), reads once again rightwards across $\gamma(\delta)$ (月), and reads rightwards across two grouping brackets at the end. Figure 4.2 shows how the 2FSA works.

As for *planar finite-state transducers*, it can be defined in a similar way as 2FSA. We only need to change the eighth item in the above definition:[6]

[6]In order to implement the central claim $\mathcal{M}_{ORL \rightarrow \Gamma}$ in a given writing system, it is

- δ *is a transition relation from* $\mathcal{Q} \times (\Sigma \cup \epsilon) \times (\Sigma \cup \epsilon \cup \mathcal{B}) \times$ d *to* $\mathcal{Q}$.

In sum, finite-state techniques are well-understood, and inherently efficient and compact mathematical models which have gained great success in many Chinese NLP tasks, such as text analysis and Chinese word segmentation. In this section, the formal properties of finite-state automata and finite-state transducers are briefly introduced, in particular, we present a planar formalism proposed by Sproat (2000), which is more than sufficient to allow for an exhaustive structural analysis of the most complex Chinese characters. Indeed, some text-processing applications are now based on planar finite state model.[7]

However, a point needs to be made here is that, via the FSA model, such as the generative grammar model, the regularities of Chinese characters as graphic patterns – without any explicit reference to sound and meaning – can be explicitly explored. However, for our purpose here we are more interested in representation models that can formalise conceptual and semantic information. In the next section, we will turn to the (semantic) network models which are closely related to the FSA model.

interesting not only in planar regular languages, but more generally in planar regular relations and their computational devices. However, since we only want to illustrate the formal models with respect of Chinese characters, such concerns are outside the scope of the present study.

[7]See Chuan, De-Ming (1999). Project in dealing with "missing characters". Chinese Document Processing Lab. http://www.sinica.edu.tw/~cdp/

4.3 Network Models

As mentioned earlier, most of the mathematical descriptions of Hanzi have focused on the shape structure, and the formalization of semantic-conceptual information encoded within Hanzi has therefore been neglected. In the remaining sections of this chapter, we will turn to this largely unexplored aspect.

The section that follows is devoted to a discussion of network models. By highlighting some well known formalisms, the aim is to make clear what could be the model most fit Chinese ideographic structure. Before beginning, I would like to introduce some basic notions.

4.3.1 Basic Notions

The following definitions are taken from Watts (2004) and Aldous (2000).

Definition 4.3.1. *(**Graph**)*
A Graph G refers to a structure composed of sets of a nonempty set of elements, called vertices, and a list of unordered pairs of these elements, called edges. The set of vertices is denoted by $V(G)$, and the list of edges is called the edge list of G, denoted by $E(G)$. The number of vertices in $V(G)$ is termed the order (n) of the graph, and the number of edges in $E(G)$ is termed its size (M). If the vertices are jointed by directed edges, such graph is called a digraph.

Definition 4.3.2. *(**Network**)*
Graphs or digraphs can be used to represent all kinds of networks, where the vertices represent some network elements (depending on the particular application under consideration), and the edges represent some predefined relationship between connected elements. Networks with undirected edges are called undirected networks, networks with directed edges are directed networks. In directed networks, the total number of connections of a vertex is

called its degree k (also called "connectivity"); while in undirected networks, the degree of a vertex $k = k_i + k_o$*, namely, the sum of its in-degree* k_i *(the incoming edges) and out-degree* k_0 *(the outgoing edges).*

As a kind of *directed network* labelled on both vertices and edges, the idea of a **Semantic Network** representation [8] for human knowledge is generally acknowledged in the field of computational lexical semantics. Semantic netwroks were proposed to represent *meaning* and relationships of natural language words. A graph is constructed where nodes represent *concepts* and they connect to other *concepts* by a particular set of arcs called *semantic relations*.

Semantic networks have been used for *knowledge representation* since the early days of artificial intelligence research. In fact, the earliest work in this area was done by Charles Sanders Peirce (1839-1914). He developed a graphical system of logic called *existential graphs* and used it to systematically record observations of the world around him. Contemporary semantic networks bear great resemblance to Peirce's existential graphs, and his graphs have been an inspiration for many researchers in fields of AI and philosophy.

In the field of psychology, graph representations have also been used to represent structures of *concepts and associations*. Otto Selz (1881-1943), a German psychologist from the University of Würzburg, used graphs to represent different concept hierarchies and the inheritance of properties. Lindsay and Norman (1977) conclude to the same idea of representing the human brain and its information storage as a semantic network: Concepts; generalizations; specializations; defaults; exceptions and their properties can be described in a simple yet expressive way (Lindsay and Norman 1997).

Generally, a semantic network is composed of three basic elements:

- *concepts* are abstract, universal entities that serve to designate a category or class of entities, events or relations. However, the content of

[8] This idea originated in the work of Quillian (1968).

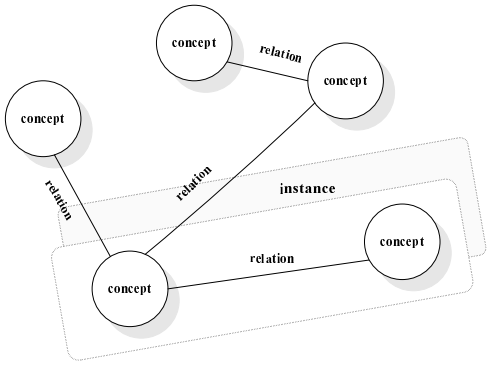


Figure 4.3: Elements of a Semantic Network

concepts varies, and depends on the theoretical setting of semantics proposed.

- *relations* describe specific kinds of links or relationships between two concepts. Every relation is *bidirectional*.

- *instances* (of a relation) consist of two concepts linked by a specific relation. An occurrence of two concepts linked by a relation is called an *instance* of that relation.

Figure 4.3 illustrates the relations between these elements.

Because Semantic Network models are powerful in modelling various things such as expert knowledge, sentences, chained causalities, narratives, and semantic constraints bearing on a lexicon (Findler 1979), a wide variety of different models have been introduced. The content of the structure represented in the semantic network depends on the applications intended.[9]

[9]More recently, semantic networks have been subject to an interesting area motivated by the search for methods to organize and display larger and more complex knowledge bases. New interest in object-oriented programming and object-oriented databases has also focused attention on the object-centered aspects of semantic networks, especially **type hierarchies and inheritance**. In general, the term "semantic network" encompasses an entire family of graph-based visual representations. They all share the basic idea of representing domain knowledge in the form of a graph, but there are some differences concerning notation, naming rules or inferences supported by the language. The term "semantic network" is also often used in a way that is almost synonymous with the term

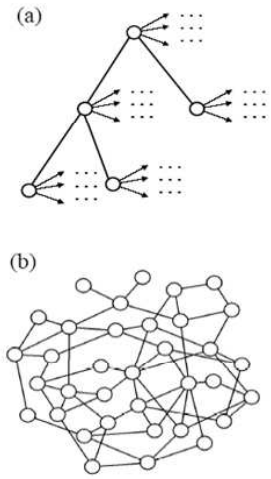


Figure 4.4: Two structures of the semantic network

Rastier (1987;1991) distinguishes three underlying paradigms: connectionism, classical computationalism and linguistics. He defines the *linguistic* approach as believing that the main interest in semantic networks lies in the ability to define the semantic proximity between concepts.[10]

Within these paradigms of semantic network models, Steyvers and Tenenbaum (2002) observed two underlying mathematical structures that have been widely used: a **tree-structured hierarchy** (e.g. Collins and Quillian, (1969)); and an **arbitrary graph** (e.g. Collins and Loftus, (1975)) (see Figure 4.4). There have been many current large-scale lexical resources developed in the form of the semantic network in a broad sense. For example, WordNet, Roget's Thesaurus, HowNet and the Chinese Concept Dictionary.[11]

conceptual graph. However, Sowa (1984) clearly distinguishes the ideas of conceptual graphs and semantic networks: each conceptual graph asserts a single proposition, while semantic networks are much larger. Sowa suggests that semantic networks are entities that embed conceptual graphs.

[10]I do not intend to give a detailed description of the numerous types of semantic networks that have been proposed for various purposes. The interested reader is referred to Rastier (1987,1991) and the book review by Corriveau (1992) in *Computational Linguistics*, Volume 18, No.4.

[11]HowNet (`http://www.keenage.com`) is an on-line common-sense knowledge base which aims to unveil inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of Chinese and their English equivalents; the Chinese Concept

4.3.2 Partial Order Relations

To our concern here, in the following, let us concentrate on three kinds of partial order relations which are quite widely used in graphic (network)-based representations of the semantic networks in a broad sense. To begin with, some definitions of basic algebraic notions are provided.

Definition 4.3.3. *(**Relation**)*
*Let A and B be sets, a **relation** $\mathcal{R}$ from A to B is a subset of* A × B*, the cartesian product of* A *and* B*. If (a,b) ∈ $\mathcal{R}$, we write aRb and say that "a is in relation $\mathcal{R}$ to b". A relation $\mathcal{R}$ on set A may have some of the following properties:*

- $\mathcal{R}$ *is* reflexive *if aRa for all a ∈ A.*
- $\mathcal{R}$ *is* symmetric *if aRb implies bRa for all a,b ∈ A.*
- $\mathcal{R}$ *is* antisymmetric *if aRb and bRa imply a = b for a,b ∈ A.*
- $\mathcal{R}$ *is* transitive *if aRb and bRc imply aRc for all a,b,c ∈ A.*

Definition 4.3.4. *(**Structure**)*
*A **structure** is a set together with one or more relations and operations defined over the set.*

Definition 4.3.5. *(**Partial Order Relation**)*
*A reflexive, antisymmetric, and transitive relation $\mathcal{R}$ on a set A is called a **partial order(relation)**. In this case, (A, $\mathcal{R}$) is called a partially ordered set or poset.*

Partial order relations describe "hierarchical" situations, and they are usually represented by the symbols ≤ or ⊆ instead of $\mathcal{R}$. Figure 4.5 shows the graphs for three kinds of partial order relations: **tree**, **lattice**, and **general**

Dictionary (CCD) (`http://www.icl.pku.edu.cn/`) is a WordNet-like semantic lexicon of contemporary Chinese.

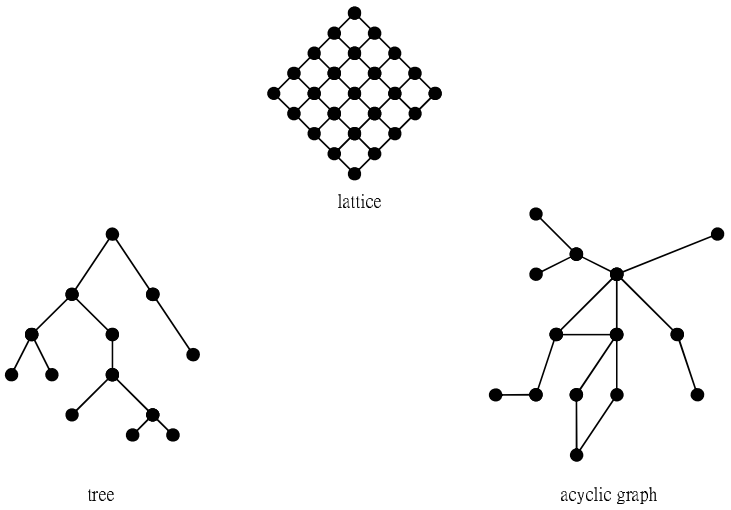


Figure 4.5: Three kinds of partial order relations (Taken from Sowa (1984))

acyclic graph.[12] In general, every tree is an acyclic graph, and every lattice is also an acyclic graph; but most lattices are not trees, and most trees are not lattices (Sowa 1983).

In practice, the *linguistic* approach to the design of semantic networks often assumes a form which falls somewhere inbetween a *tree* and a *lattice* into certain degrees of *directed acyclic graph* structures. The property of an acyclic graph will be dealt with in the following section, for the moment, we will only provide the definition.

Definition 4.3.6. *(**DAG**)*
*A **directed acyclic graph** (DAG) is a pair $\mathcal{G} = (G, <)$ such that $<$ is an acyclic relation on G. If $<$ is transitive, $\mathcal{G}$ is called a **directed transitive acyclic graph** (DTAG).*

[12]Taken from Sowa (1983:383). A *cycle* is a path that returns to its starting point - the first and last vertices are identical. A graph without any cycle is said to be *acyclic*. As Sowa noted, to simplify the drawings, for the acyclic graph in Figure 4.5, the arrows on the arcs are conventionally omitted , but to assume that the arcs are directed either from the higher node to the lower node or the other way round.

4.3.3 Tree

This subsection introduces basic notions about trees and their properties from the graph theory.[13]

Definition 4.3.7. *(**Undirected Graphs and Trees**)*
*A Graph $G = (V, E)$ is **undirected** if $(v, w) \in E$ implies $(w, v) \in E$, for all $v, w \in V$. A connected **Undirected Graph** is called a **Tree** if it has no cycles.*

Among the tree structures, some particular types (rooted and labelled trees) occur repeatedly in linguistic literatures.

Definition 4.3.8. *(**Rooted Tree**)*
*A **rooted tree** is a tree, in which one vertex is singled out as the starting point, and the branches fan out from this vertex. We refer to this particular vertex as the* root*, such that for all nodes $v \in$ V, there is a path in* G *from the root r to the node v. A **binary tree** is a* rooted tree *where every node that is not a* leaf *has exactly two* children*.*

Definition 4.3.9. *(**Labelled Tree**)*
*A **labelled tree** is a tree with labelled nodes . It can be depicted as a 5 tuple $\mathcal{T} = (N, Q, D, P, L)$ if the following conditions are satisfied:*

1. *N is a finite set of nodes.*
2. *Q is a finite set of labels.*
3. *D is a partial order relation on N, called the dominance relation.*
4. *P is a strict partial order relation on N, called the precedence relation.*
5. $(\exists x \in N)(\forall y \in N)[(x, y) \in D]$
6. $(\forall x, y \in N)[[(x, y) \in P \vee (y, x) \in P] \leftrightarrow [(x, y) \notin D \wedge (y, x) \notin D]]$

[13] These formal defnitions are modelled on Valiente (2002).

7. $(\forall x, y, z, w \in N)[[(w, x) \in P \wedge (w, y) \in D \wedge (x, z) \in D] \rightarrow (y, z) \in P]$

8. *L:* $N \rightarrow Q$ *is a label map*

Traditionally, a tree is a tuple $T = (N, D, P)$ satisfying all of the conditions in the preceding definition except (2) and (8).

Definition 4.3.10. *(**Hypernym**)*
*If the proposition "*x *is a kind of y" is true, then y is the **hypernym** of x (denoted by* $x \preceq Hz$*) or x is the **hyponym** of y. A hypernymy relation is:*

- *transitive:* $\forall x, y, z \in N, x \preceq Hy$ *and* $y \preceq Hz$
- *asymmetrical:* $\forall x, y \in N, x \preceq Hy$ *and* $y \preceq Hx \rightarrow x = y$

A few decades ago, Collins and Quillian (1969) suggested that concepts can be represented as nodes in a tree-structured hierarchy, with connections determined by the hypernym / hyponym relations (Figure 4.4). Such a hierarchy provides a compact and elegant manner for representing categorical knowledge, but as Steyvers and Tenenbaum (2002) criticized, it has severe limitations as a general model for semantic structure. The property of inheritance seems only appropriate for certain taxonomically organized concepts, such as classes of animals. Even in those cases, a strict inheritance structure seems to apply except only to the most typical members of the hierarchy.

That is, a tree-structured hierarchy is only suitable for compact representation; and if it is chosen for this task, the atypical instances of the hierarchy have to be clearly differentiated from one another in a different way.

Definition 4.3.11. *A **Huffman code tree** is an unbalanced binary tree.*

The last thing in regarding tree structure that we are concerned with is the question of an encoding scheme. For the notion of *coding* through most of this thesis, we will adopt a Huffman encoding tree-like scheme, which is widely used in the area of data and text compression, due to its efficiency.

If the Chinese characters to be encoded are assigned to a tree-like structure with binary and quaternary branches (such as Chu's concept hierarchy), then an encoding of each character can be found by following the tree from the root to the character in the leaf node: the encoding is the string of symbols on each branch followed. That is, the sequence of edges from the root to any character yields the binary code for that character.[14]

4.3.4 (Concept) Lattice

This subsection introduces another important kind of partial order relation: the *lattice structure*.

Definition 4.3.12. *(**Lattice**)*
*A **lattice** is a structure consisting of a set A, a partial order relation $\preceq$, and two binary operators $\cap$ (meet; intersection) and $\cup$ (join; union), which satisfy the following laws for all $x, y, z \in L$:*

- *(L1: communicative): $x \cap y = y \cap x$, $x \cup y = y \cup x$;*
- *(L2: associative): $x \cap (y \cap z) = (x \cap y) \cap z$, $x \cup (y \cup z) = (x \cup y) \cup z$;*
- *(L3: absorption): $x \cap (x \cup y) = x$, $x \cup (x \cap y) = x$.*

Two applications of (L3), namely, $x \cap x = x \cap (x \cup (x \cap x)) = x$, lead to the additional law:

- *(L4: idempotent): $x \cap x = x$, $x \cup x = x$.*

Definition 4.3.13. *(**Upper and Lower Bounds**)*
Let $(A, \preceq)$ be a poset and $B \subseteq A$, then

[14] In coding theory, the code uses the same number of bits to represent each symbol is called a *fixed-length code* in coding theory. The set of binary sequences is called a *code*, and the individual members of the set are called *codewords*.

(i) $a \in A$ *is called an upper bound of* B *if* $b \preceq a$ *for all* $b \in B$.

(ii) $a \in A$ *is called a lower bound of* B *if* $a \preceq b$ *for all* $b \in B$.

(iii) The greates amongst the lower bounds of B, if it exists, is called the ***greatest lower bound*** *(or infimum) of B.*

(iv) The least upper bound of B, if it exists, is called the ***least upper bound*** *(or supremum) of B.*

A ***bounded*** *lattice is one with a top* $\top$ *and a bottom* $\bot$ *, where for any element a in the lattice,* $\bot \leq a \leq \top$.

In the following, I will introduce the main ideas of *Formal Concept Analysis* method, in which *conceptual structures* are modelled as a hierarchical network in terms of a special case of lattice. The goal is to discuss whether this analysis could activate mathematical thinking for conceptual data analysis and knowledge processing of Hanzi.[15]

The Formal Concept Analysis (hereafter FCA) is a theory of data analysis in the field of applied mathematics, which is based on the mathematization of *concept* and *conceptual hierarchy*. It was introduced by a German mathematician Rudolf Wille in 1982. Since it can identify *conceptual structures* among data sets, it has been successfully applied to a broad variety of domains such as sociology, medicine, computer science and industrial engineering.

The FCA method focuses on the **Concept Lattice Structures**, also called Galois lattices, arisen from binary data tables, which have been shown to provide a theoretical framework for a number of practical problems in information retrieval, software engineering, as well as knowledge representation and

[15]The introductory part is mainly based on Wolff (1993). For a more mathematical treatment of some of the topics covered here, the reader is referred to Ganter and Wille (1999). A lot of relevant publications can be found under `http://www.mathematik.th-darmstadt.de/ags/`, in both English and German.

management. One of its best features is its capability of producing graphical visualizations of the inherent structures among data. Due to this capability, it can also be used as a fit tool in formalizing, revising and refining lexical databases, thesauri and ontologies.[16]

Priss (2003) proposes that FCA as a methodology of data analysis and knowledge representation has potential to be applied to a various of linguistic problem. For instance, we can use FCA to (1) build a lexical database, thesaurus or ontology, (2) visualize conceptual structures in a lexical database, and (3) analysis semantic relations and identify inconsistencies among semantic relations in a lexical database.

In the following, we will formally introduce FCA method and provide an example of the analysis of Chinese characters. To allow a mathematical description of extensions and intensions, FCA starts with the definition of a *formal context*.

Definition 4.3.14. *(**Formal Context**)*
*A **formal context** is a triple $\mathcal{K}$:= (G,M,I), consisting of two sets G and M, and a binary relation I between G and M. That is,* $\mathrm{I} \subseteq \mathrm{G} \times \mathrm{M}$. *The elements of G and M are called **objects (Gegenstände)** and **attributes (Merkmale)**, respectively. The relation is written as gIm or (g,m) ∈ I and is read as "the formal object g has the attribute m".*

A formal context can be represented by a *cross table* that has a row for each object *g*, a column for each attribute *m*, and a cross in the row of *g* and the column of *m* of *gIm*. For instance, Table 4.1 shows an example of a formal context for various kinds of vehicles in Chinese. It assigns the attributes 二輪 (two-tires), 四輪以上 (four-tires plus), 公用 (public), 私用 (private), 汽油引擎 (oil-burning) to the objects 車 (vehicle), 汽車 (car), 火車 (train), 脚踏車 (bicycle), 救護車 (ambulance), 機車 (motorbike), and 公車 (bus).

[16]See Priss (1998) for an analysis for WordNet and Old (2002) for Roget's Thesaurus.

Definition 4.3.15. *For $A \subseteq G$, we define*

$$A' := \{m \in M \mid \forall g \in A : (g, m) \in I\}$$

and, analogously, for $B \subseteq M$,

$$B' := \{g \in G \mid \forall m \in B : (g, m) \in I\}$$

So in Table 4.1, A' { `bus` } = { `four-tires plus, public, oil-burning` } and B' { `four-tires plus` } = { `car, train, bus` } both hold.

Definition 4.3.16. *(**Formal Concept**)*
A pair (A, B) is a formal concept $\mathcal{C}$ *of the formal context (G, M, I) if and only if*

$$A \subseteq G, B \subseteq M, A' = B, and\ A = B'.$$

For a formal concept $\mathcal{C} := (A, B)$, A is called the *extent* (denoted by *Ext(c)*) and B is called the *intent* (denoted by *Int(c)*) of the formal concept. In the example of Table 4.1, ({`car, bicycle, motorbike`}, {`private`}) is a formal concept because A' {`car, bicycle, motorbike`} = {`private`}, and B' {`private`} = {`car, bicycle, motorbike`}.

The set of all formal concepts of a context $\mathcal{K}$ with the order relation $\leq$, denoted by $\mathcal{B}$ ($\mathcal{K}$) (or $\mathcal{B}$ (G, M, I)), is called the **concept lattice** of $\mathcal{K}$. It is always a complete lattice, i.e. for each subset of concepts, there is always a unique greatest common subconcept and a unique least common superconcept. Figure 4.6 shows the concept lattice of the formal context in Table 4.1 in the form of a line diagram.

Concept lattices can be depicted as *line diagrams* as in Figure 4.6, in which a formal concept is represented by a small circle. For each formal object g, the smallest formal concept to whose extent g belongs is denoted by γg; and for each formal attribute m, the largest formal concept to whose intent m belongs is denoted by μm. The concepts γg and μm are called *object concept* and *attribute concept*, respectively. In the line diagram it

	two-tires	four-tires plus	public	private	oil-burning
vehicle					
car		√		√	√
train		√	√		
bicycle	√			√	
ambulance	√		√		√
motorbike	√			√	√
bus		√	√		√

Table 4.1: A formal context of vehicles

is not necessary to include either the full extent or intent for each concept; instead, the name (verbal form) of each formal object g is written slightly above the circle of μm.

In a line diagram, the extent of a formal concept consists of all objects whose labels are attached to subconcepts. Analogously, the intent consists of all attributes attached to superconcepts. For example, the concept labelled *oil-burning* has {car, ambulance, motorbike, bus} as extent, and {oil-burning, two-tires} as intent. Based on that, FCA method can be useful in concept learning if we add more objects and attributes. Figure 4.6 shows a more complex concept lattice of the formal context by adding more objects.

The most important structure on $\mathcal{B}$ (G, M, I) is given by the *subconcept-superconcept relation* that is defined by

$$(A_1, B_1) \leq (A_2, B_2) :\Longleftrightarrow A_1 \subseteq A_2 (\Longleftrightarrow B_2 \subseteq B_1).$$

For example, in table 4.1, ({car, bicycle, motorbike }, {private}) as a formal superconcept of ({motorbike}, {four-tires minus, private, oil-burning }), has more objects but fewer attributes than ({motorbike}, {four-tires minus, private, oil-burning }).

It follows from this definition that each formal concept is a formal subconcept of itself, in contrast to the natural language use of *subconcept*, which precludes a concept from being a subconcept of itself. The relation $\leq$ is a

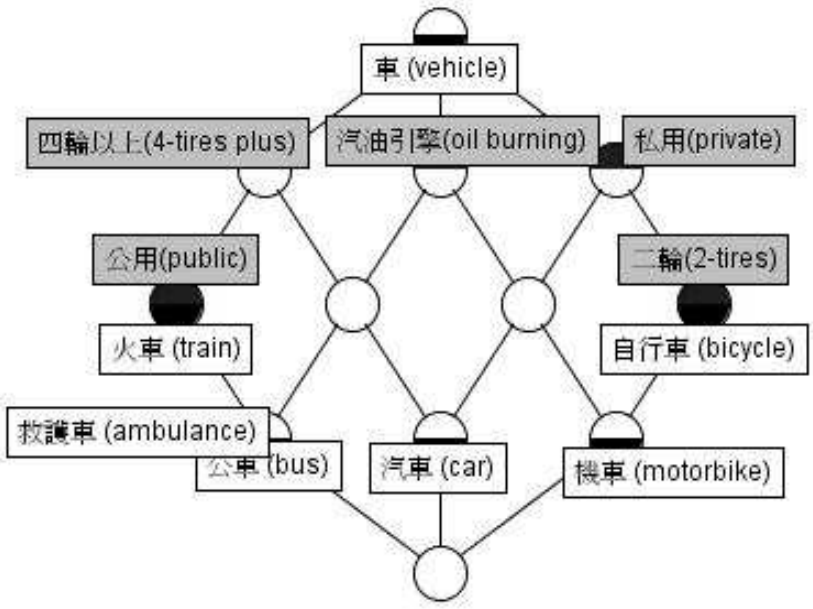


Figure 4.6: A concept lattice represented by a line diagram

mathematical order relation called *formal conceptual ordering* on $\mathcal{B}$ (G, M, I) with which the set of all formal concepts forms a mathematical lattice denoted by $\mathcal{B}$ (G, M, I).

More details of the FCA can be found in Ganter and Wille (1996).

The Concept Lattice has some advanced features over other representations, However, as Vossen (2003) observed, many lattice structures introduce a (very) large number of internal nodes for feature combinations that have no natural human interpretation (see Figure 4.7), and hence no lexical expression in many languages. Therefore he concluded that lexicalization in natural language does not obey the systematic rules of a lattice or a tree. Whereas the lattice generates all logical concepts, language tends to lexicalize only those concepts that are efficient to support communication.

Based on the comparison of EuroWordNet Top ontology with the organization of Chinese radicals, Wong and Pala (2001) concluded that, the semantic features of the component (radical) organization of Chinese charac-

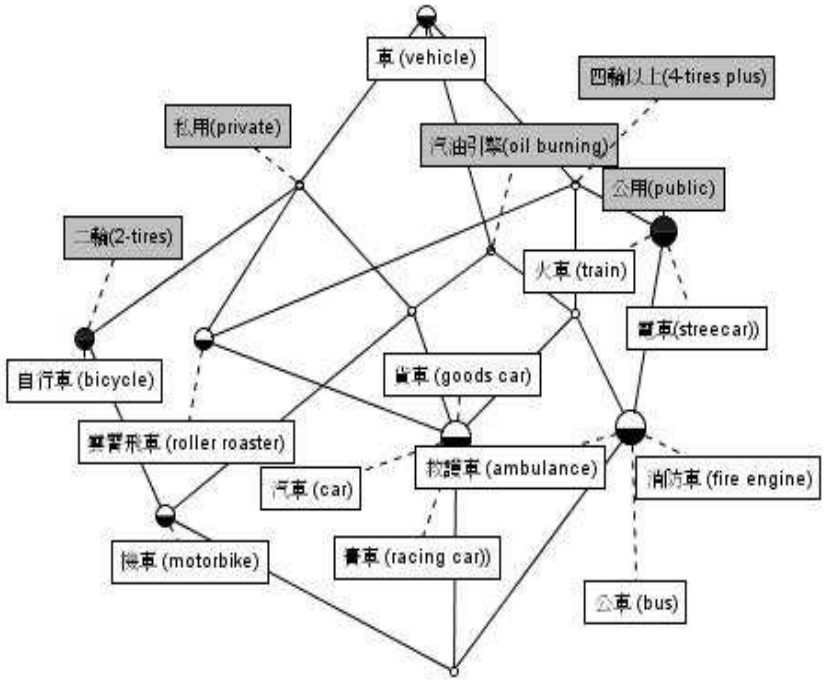


Figure 4.7: A more complex concept lattice

ters form a semantic network as opposed to tree structures found in existing ontologies. Rather, such a semantic network is richer, more complete, and more transparent in the sense that, it is a *natural* organization. It would be interesting to apply the FCA to the analysis of Hanzi components.

4.4 Statistical Models

Statistical methods have become a mainstream approache to computational linguistics and neighbouring fields. They have been successfully applied to many tasks such as word co-occurrence and similarity, lexical collocation and association.

This section will describe the statistical and probabilistic aspects of characters and their components. For clarity, it is composed of three subsetions which deal with statistical studies of three aspects of Chinese characters: the **character itself, character combination** and **character network.**

4.4.1 Character Statistics

Previous statistical studies of Chinese characters have focused mainly on *counting*. A good example of this is, the frequency table of currently used characters.[17] Though these simple counts on texts can be used in other applications, they are hardly linguistic significant (Manning and Schütze 1999).

Some linguists have turned to explore the invariable *l*aws that govern natural language. The most well-known of these is Zipf's Law: $f \propto \frac{1}{r}$, which states that the relation between the *rank* r of the word and its *absolute frequency* f, is constant. Research in the European tradition of *quantitative* and *synergetic* linguistics has made the strong assumption that language is a complex self-regulating system, within which many language laws can be detected in the quantitative dependencies, such as the relation between particular variables (e.g. frequency (F), length (L), polylexy (PL), and polytexty (PT)). Köhler (1986) derived many models of language structures that might be of some interest to the concerns of NLP. For example, $L =$

[17] Due to the lack of agreement concerning the definition of *components*, a systematic explanation of *character statistics* (e.g. the *distribution* of Character Head and Character Body, and the *correlation* between Character Head and Character Body) is not available either.

$a_1 F^{-b1}$, which means that the *Length* L is a function of the *Frequency* F; and $Pl = a_2 L^{-b_2}$, which means that *Polylexy* (the number of meanings) Pl is a function of the length L.

Basically, *Length* is measured in terms of three basic types of units (Grotjahn and Altmann 1992). Namely, graphic (letters, strokes, or radicals), phonetic (phonemes or syllables) and semantic (morphemes). They are again, *word-based*. In past studies,[18] Chinese word length has almost always been measured in terms of the number of characters. Seen from the angle of methodological consideration (Grotjahn and Altmann 1992), the choice of the unit of measurement could effect the construction of model for word length. I have proposed a character-based *Length* modelling of distribution measured in terms of the stroke numbers of *morphemes*, in order to compare with the results from the previous word-based studies.[19]

By assuming that Chinese characters also function as linguistic units, character lengths are not distributed at random, but correspond to specific laws, two small experiments investigating the relations between character length, frequency and meaning numbers were made. Figure 4.8 shows the initial results concerning the relation between character length and frequency (a), character length and polylexy (b) (The Least Square Method was used for the curve fitting.) The results demonstrate that to a certain degree, the language laws are abided by.

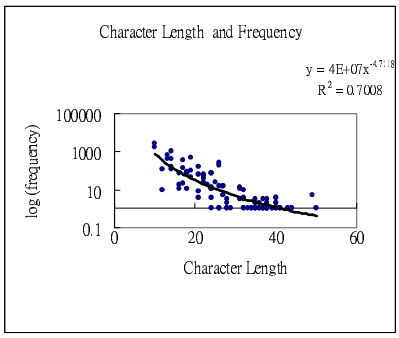
4.4.2 Statistical Measures of Productivity and Association of Characters

Chinese characters mostly do not occur in isolation, but rather as a rule combine with other characters to make (polymorphemic/polysyllabic) words. In the previous chapters, we have introduced that the ingenuity of Chinese

[18] See Zhu and Best (1998). Wortlängenhäufigkeiten in Chinesischen Kurzgeschichten. In *Asian and African Studies 7*; Hartmut Bohn (1998).

[19] Part of them were presented in Hsieh (2003). The results are promising but still need refinement.

(a)



(b)

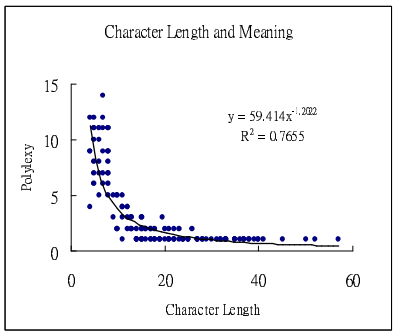


Figure 4.8: Character-based language laws testing

writing lies in word formation: A nearly unlimited number of Chinese words can be composed by combinations from a stock of around 3000 - 4000 unique characters.

The *productivity* and *association* of Chinese characters have become one of the central notions of Chinese computational (derivational) morphology. They also provide useful information for Chinese lexicography. In this section, I will introduce some measures for co-occurrence statistics widely used in the area of Lexical Knowledge Acquisition, in the hope that, by bringing statistical measures into morphological theory, we will be able to make the intuitive notion of the underlying process of conceptual combination or semantic transformation more precise.

- *Morphological Productivity: Morphemes vs.Characters*

In linguistics, the problem of vocabulary richness with respect to word formation patterns is known as the problem of **morphological productivity**.[20] Simply put, the focus of a study of *morphological productivity* is on *derivational affixation* that involves a *base* word and an *affix* (Aronoff 1976), such as the English word "*productiv + ity*". In Chinese, no matter one define the *Affix* or *Morpheme*, they are all character-based. In addition, the most

[20]The measures discussed here are based on the presentation in Sproat and Shih (2001), Baayen (2001;2003).

important word-formation process is *compounding*, i.e. the combination of two or more characters - each usually having with its own lexical meaning, (i.e. having a substantial meaning of its own) -, to produce a new unit that functions as a single word. Since Chinese characters are extremely productive in their ability to generate new words, compounding plays a major role in Chinese morphology. So the term *productivity* or *word formation capability of characters* will be preferred here.

Most research concerning morphological productivity has attempted to apprehend these aspects of this phenomenon in terms of qualitative properties of word formation patterns.[21] In fact, as Baayen (2003) noted, morphological productivity is graded or scalar in nature, with for instance, productive word formation at one extreme (*-ness, goodness*), semi-productive word formation in the middle (*-ee, employee*), and unproductive word formation at the other extreme (*-th, warmth*). It is a pre-theoretical notion with various interpretations that can each be formalized statistically.

Various measures that formalize the notion of degree of productivity have been proposed, one of these (Aronoff 1976) is defined as:

$$I = \frac{V}{S} \tag{4.1}$$

where V is the number of distinctive instances of a morphological category – e.g. the number of words in a dictionary ending in the suffix -ee –, and S is the number of *potential* types of that category. However, such numbers are difficult to even estimate, even with the aid of a dictionary or corpus.[22] Similar to the "Good-Turing" Measure (Good 1953), another measure that turns out to be more reasonable than Aronoff's was proposed by Baayen (1989):

$$\mathcal{P} = \frac{V(n,1)}{N} \tag{4.2}$$

[21]For an excellent description please refer to Baayen (2001).

[22]For a detailed discussion, please refer to Sproat and Shih (2001) or Baayen (2001).

where Productivity $\mathcal{P}$ is defined as the number of hapax legomena $V(n,1)$ divided by the number of *tokens* N, of a particular construction found in a corpus. (For instance, the number of tokens of all nouns ending in -*ness*). In the case of characters, consider a corpus where the value of N is 7468. For the human noun plural affix 們 (/mėn/), $V(n,1) = 253$, so the $\mathcal{P}$ of 們 is equal to $\frac{253}{7468} = 0.03$; for a more productive aspectual verbal affix 了 (/lė/), the value of $\mathcal{P}$ is higher ($\frac{425}{7468} = 0.05$).

Though the measurement of character productivity is meaningful for corpus-based studies in Chinese morphology, in this thesis, we are also interested in examing further whether statistics can help in explaining the semantic constraints. In the ensuing discussion, we turn to the issue of **character association**,[23] but restrict ourselves to the association between two characters.

Statistical research in character association, such as research in *collocation acquisition*, mostly took a *frequency-based metric* in measuring a certain type of collocation. So the *character association* was defined as a pair of characters that appear together more often than would be expected by chance. To estimate the correlation between character pairs, a metric called *Mutual Information* has often been adopted.[24]

- *Mutual Information*

Mutual Information $\mathrm{MI}(x;y)$ compares the probability of observing character

[23]This term is similar to the notion of *Collocation* in linguistics, which falls somewhere along a continuum between *Free-word Combination* and *Idioms*. Nonetheless, in a character-based context, careful distinction between these two terms should be made, and the term *character association* is preferred in this thesis.

[24]There are also some other statistical measures such as *t-score*, *likelihood ratio*, *chi-square* and *Yule's coefficient of colligation* Y that are often used to measure the strength of collocation. However, in his informal experiments using likelihood ratios and chi-square measures, Jun found that these two statistical methods do not provide a reliable measure of collocation as far as the two diagram lists are concerned. The problem with these two methods is that they are much less discriminative as compared with MI. For preliminary report, please take a look at this comparison page: http://www.bio.utexas.edu/staff/jun/chinese-computing/statistics/fhy-collocation.html

x and character y *together* (the joint probability) with the probabilities of observing x and y *independently* (chance).

$$\mathrm{MI}(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{4.3}$$

$$= \log_2 \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N} \cdot \frac{f(y)}{N}} \tag{4.4}$$

It provides a metric for measuring the *degree of association* between two characters. If there is a strong association between characters x and y, then the joint probability $P(x,y)$ will be much larger than chance $P(x)P(y)$, and consequently $MI(x;y) \geq 0$. If character x and character y are independent, then $P(x,y) \sim P(x)P(y)$, and thus, $MI(x;y) \sim 0$. If x and y are in complementary distribution, then $P(x,y)$ will be much less than $P(x)P(y)$, forcing $MI(x;y) \leq 0$. Character probabilities $P(x)$ and $P(y)$, are to be estimated by counting the frequency of x ($f(x)$) and y ($f(y)$), over a corpus, or frequency data, and then normalizing by N, the size of the corpus. Joint probabilities $P(x,y)$, are computed by counting the number of times that x is followed by y, $f(x,y)$, and normalizing by N.[25]

- *Information Content and Word Binding*

While it is easy to understand that an almost unlimited number of words can be composed by combinations of around 3000-4000 unique characters, it is still not clear about the *qualitative* change, i.e., the change in information content during such a progress. Lua (1990) has investigated the formation of Chinese words from characters by the application of information theory. Lua proposes that, the *binding force* in the process of word formation by combining a number of characters can be derived from the loss of information content when a word is formed.

[25]Several on-line resources are available such as Sinica Corpus `http://www.sinica.edu.tw/SinicaCorpus/`, or Jun Da's Chinese Text Computing project `http://www.bio.utexas.edu/staff/jun/chinese-computing/`.

The information content I_i of a linguistic symbol (character or word) is measured by its entropy value proposed by Shannon (1948). It is related to the probability of occurrence of the symbol, i.e., P_i, of a language system:

$$I_i = -\log_2 P_i \tag{4.5}$$

And the loss in information content (I_c) when p characters are combined to form a word can be calculated from:

$$I_c = I_w - \sum_{i=1}^{i=p} I_i \tag{4.6}$$

where I_w and I_i represent the information content of the word and its constituent characters respectively. There can be two results: $I_c < 0$ or $I_c > 0$. The first result is the *usual* case where there is information loss when the word is formed. A high loss in information content indicates strong binding between the characters. The characters are less independent and it is also easy to derive the meaning of the word from the composing characters. An example is the dissyllabic word 媽媽(/mama/; "mother"), where the meaning is almost the same as when it is in the monosyllabic form, 媽. The later result is an *unusual* case where there is information gain when the word is formed. It indicates that the meaning of a word is deviated from the meaning of its composing characters significantly. Most examples come from the foreign word such as 可可(/keke/;"coco"). Its composing characters are two 可, which means "alright". Based on the "Modern Chinese Frequency Dictionary" (1985), Lua found that most words belong to the first case.

Lua (1990) also derived a quantity called word binding ($\mathcal{B}$) from the consideration of probability of occurrence. Let us consider a word w which consists of p characters. The probabilities of occurrence for this word and each of its constituent characters are $P_w, P_1, P_2, ...P_p$. The probability of obtaining a word w by randomly selecting p characters are :

$$P_{w'} = \prod_{i=1}^{i=p} P_i \tag{4.7}$$

We now define the *word binding* B as:

$$\mathrm{B} = \frac{P_{w'}}{P_w} \tag{4.8}$$

It can easily be derived that:

$$\log_2 B = I_c \tag{4.9}$$

The derivation is shown below:

$$I_c = I_w - \sum_{i=1}^{i=p} I_i \tag{4.10}$$

$$= -\log_2 P_w + \sum_{i=1}^{i=p} \log_2 P_i \tag{4.11}$$

$$= \log_2 \left(\frac{P_{w'}}{P_w}\right) \tag{4.12}$$

$$= \log_2 B \tag{4.13}$$

Thus the change in the information content I_c when a word is formed provides a direct measuree to the *word binding* between characters.

The binding force of a word is a measure of how strongly the characters composing the word are bound together as a single unit. This force is often equated with the usage frequency of the word. It is reasonable that the change in the information content or the word binding force can serve as a guide to the degree that the original meaning of a character is extended, modified or transformed. However, in a strict sense, we are still not able to predict the meaning of a word by using these two quantities. They do, however, as will be applied in later experiment, serve as important linguistic parameters.

4.4.3 Characters in a Small World

Now let us consider the connection aspects between characters in a global, dynamic manner. The interest here is primarily focused on how Chinese characters *behave* and how this behavior is affected by their connectivity from a statistical viewpoint.

Background

The network models discussed in the previous section are be classified as *regular* and *random graphs* in traditional Graph Theory. Different network models exhibit different degrees of heterogeneity and randomness. As seen in Figure 4.5 in the previous section, among the typical network models of partial order relations, lattice-like and tree-like networks represent the highest degree of homogeneity and have no randomness, while the acyclic graph is a *random graph* such that two nodes are joined with some probability p (like Erdös-Rényi Graphs) (Solé and Valverde 2004).

From the views of Graph Theory, *regular graphs* (networks) have high clusterings and small average shortest paths, while *random graph* (networks) are found at at the opposite of the spectrum, as they have small average shortest path and low clusterings. It seemed that no interesting things existed between *regular (or deterministic) networks* and *random networks*.

By the middle of the 1990s, with astounding discoveries and the development of a vast number of *networks*- be they natural (e.g. biological networks) or artificial (e.g. the World Wide Web)-, which all have a specific architecture based on a *self-organizing, fat-tailed, non-Poisson distribution* of the number of connections of vertices that differs crucially from the "classical random graphs". The structure of networks with random connections has turned out to be an object of immense interest for researchers in the various sciences. These new trends have also been incorporated into the study of the lexical

semantic network.[26]

Based on the characteristics of real-world networks, two important quantitative features have been reported: the "small world model" by Watts and Strogatz (1998); and the "preferential attachment model of scale-free networks" by Barabási and Albert (1998). These models have reshaped the way we think of networks.[27]

- *Small World Phenomenon*

Research specific to the *small world* phenomenon in the *network* began with the idea of a "social network" employed by sociologists in the 1960s. The small world phenomenon formalises the anecdotal notion that "you are only ever six 'degrees of separation' away from anybody else on the planet." This claim infers that even when two people do *not* have a friend in common, they are separated only by a short chain of intermediaries (Watts 2004).

Since then, it has been observed that many real-world networks exhibit this so-called *small world phenomenon*, with its two distinguishing features: a *small distance* between any pair of nodes, and a *clustering effect*, which means that two nodes are more likely to be adjacent if they share a neighbor. As in the view of Graph Theory, regular networks have high clusterings and small average shortest paths, with random networks at the opposite of the spectrum, as they have small shortest paths and low clusterings. Small-world networks fall somewhere in between these two extremes. In this thesis, I use the term *small-world network* to refer to the combination of these two features: the average shortest path-length (as small as that in a random network

[26] See the pioneering work of Steyvers and Tenenbaum (2002).

[27] For more details about the revolution in network science, please refer to Ben-Naim et al (2004). One point that should be noted here is, as Dorogovtsev and Mendes (2003) remind us, the particular network we observe is only one member of a *statistical ensemble* of all possible realizations. Therefore, when speaking about *a random network*, we are actually speaking about *an ensemble of nets*. A statistical description of a random network only *suggests* the description of the corresponding statistical ensemble.

with the same parameters), and the relatively high clustering coefficient (as high as that in a regular network) (Watts 2004).

The Path Length $\mathcal{L}$ refers to the distance $d(i, j)$ between every vertex and every other vertex. "Distance" here refers to the minimum number of edges that must be traversed in order to reach vertex j from vertex i, or simply, the *shortest path length* between i and j (Watts 2004). The Clustering Coefficient $\mathcal{C}$ characterizes the "cliquishness" of the closest environment of a vertex, or, in other words, the extent of the mutual "acquaintance" of its closest vertices (Dorogovtsev and Mendes 2003).

The formal definitions in the following are taken from Watts (2004):

Definition 4.4.1. *(**Path Length**)*
The path length $\mathcal{L}$ of a graph G is the median of the means of the shortest path lengths connecting each vertex $v \in V(G)$ to all vertices. Namely, calculate $d(v, j) \forall j \in V(G)$ and find $\overline{d}_v$ for each v. Then define L as the median of $\{ \overline{d}_v \}$.

Definition 4.4.2. *(**Clustering Coefficient**)*
The clustering coefficient $\mathcal{C}_v$ depicts the extent to which vertices adjacent to any vertex v are adjacent to each other,

$$\mathcal{C}_v = \frac{|E(\Gamma_v)|}{\binom{k_v}{2}} \tag{4.14}$$

where the neighbourhood Γ_v *of a vertex v is the subgraph that consists of the vertices adjacent to v (not including v itself); $|E(\Gamma_v)|$ is the number of edges in the neighbourhood of v, and $\binom{k_v}{2}$ is the total number of possible edges in Γ_v.*[28]

For example, suppose we have an undirected network, in which one of its vertices v has four nearest neighbors, and there are two edges between

[28] In fact, such subgraphs can be regarded as small loops of length 3 in the network.

these nearest neighbours. Then the clustering coefficient $\mathcal{C}_v$ of this vertex is calculated as $\mathcal{C}_v = \frac{2}{\binom{4}{2}} = \frac{1}{3}$. If we would like to calculate the clustering coefficient of this network, $\mathcal{C}$ we simply take the average value of $\mathcal{C}_v$.

One can easily see that the clustering coefficient of a completely connected network is equal to 1, while on the other hand, the clustering coefficient of a tree is 0. For the purpose of comparison, the statistical features of the classical random graph will also be computed. Suppose that a classical random graph consists of N vertices randomly connected by M edges, with the mean degree $\bar{k}$. Each pair of vertices is connected with the same probability $\cong \frac{|E(\Gamma_v)|}{N}$. Here, $|E(\Gamma_v)| = \bar{k} = \frac{2M}{N}$. The clustering coefficient of a classical random graph is therefore $\mathcal{C}_{random} = \frac{|E(\Gamma_v)|}{N}$.

Definition 4.4.3. *(**A Small-world Network**)*
A small-world network *is a graph G with n vertices and average degree $\bar{k}$ that exhibits $\mathcal{L} \approx \mathcal{L}_{random}(n,k)$, but $\mathcal{C} \gg \mathcal{C}_{random}$.*

- *Scale-free Network*

The term *scale-free network* was first coined by the physicist Albert-Laszlo Barabási and his colleagues (Barabási 1998). This is a specific kind of network which demonstrates short-range correlations between vertices and a decrease of a local clustering coefficient with increasing degree of a vertex.

In such networks, the distribution of connectivity is extremely uneven. Some nodes act as "very connected" hubs using the *power-law* degree distribution.[29] Formally, scale-free networks are networks whose degree distribution (i.e. fractions of nodes with k degrees (connections)) behaves as:

$$P(k) \propto k^{-\lambda}, k \neq 0, m \leq k \leq K, \tag{4.15}$$

[29]In contrast to other degree distributions such as the Exponential distribution or the Poisson distribution, the Power-Law distribution has no natural scale, and hence may be called *scale-free*. Networks with such distributions are thus labelled as *scale-free* networks.

where λ is the exponent, m is the lower cutoff, and K is the upper cutoff. There is no node with a degree below m and above K.[30]

- *Do Chinese Characters Constitute an Affiliation Network?*

Extensive studies have shown that many large natural and artificial networks have both the small world and scale-free features. In the lexical network field, Steyvers and Tenenbaum (2002) investigated graph theoretic properties of the semantic networks created by WordNet, Roget's Thesaurus, and the associative word lists built by Nelson et al. All three lexical resources turned out to share distinctive statistical features of both *small-world* and *scale-free* structures. These results motivate us to speculate that this sort of property is widespread among *networks* of Chinese characters.

Do Chinese characters actually *live* in a small world? What are the most general conditions under which the world can be considered "small"? The following is devoted to tackling this question.

As discussed previously, Chinese writing system employs many thousands of characters (Hanzi), which can combine in a fairly sophiscated way. Quite unlike European writing systems, the Chinese writing system is constructed in such a fashion that it carries abundant complex conceptual, phonological and semantic information.

In order to survey the statistical properties of Chinese characters, we need a database of characters. But, how we define when two characters are connected depends on how we define the relation between them. Previous research like that of Fujiwara et al. (2002) is based on 6500 Hanzi (Kanji) used in Japan, and extracted from a character database (UTF-2000). In

[30]To illustrate the mechanism of a scale-free network, Barabási and Albert introduced an evolving network model where the number of vertices N increases linearly with time rather than remaining fixed, and a newly introduced vertex is connected to m already existing vertices with a probability linearly proportional to the degree of the selected vertex, which is called the preferential attachment rule. The degree exponent then follows the power law with the exponent $\gamma = 3$. A generalized version assigns the probability proportional to $k + m(a-1), a \geq 0$ being a tunable parameter. The the degree exponent is then $\gamma = 2 + a$.

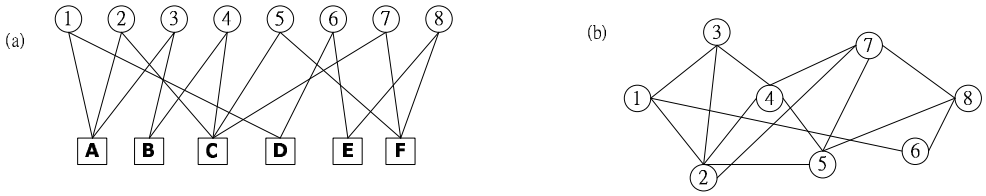


Figure 4.9: (a). Bipartite graphs of characters (the numerically indexed row) and components (the alphabetically indexed row), (b). Reduced graph from (a) containing only characters.

this experiment, characters are decomposed into components in a tree-like hierarchy, but without explicit reasons for the decomposition rules.

Instead, we will construct a bipartite network from the entries in a concept knowledge base of Chinese characters, called the Hanzi Genes Dictionary (http://www.cbflabs.com). Apart from pictographs, each character in this dictionary is decomposed into two parts: the *Character Head* and the *Character Body*, which both contribute to the meaning composition. The number of components and characters are 256 and 6493, respectively. The vertices in this bipartite network are split into two sets, S_1 (components: A,B,C,D,E,F,...), and S_2 (characters: 1,2,3,4,5,6,7,8,...) in such a way that each edge has one end in S_1 and one end in S_2. The construction of the network is depicted in Figure 4.9.

Data

The data meets the requirement $N \geq \overline{k} \geq \log(N) \geq 1$, where the total degree of network $K \geq \log(N)$ guarantees that a random graph is connected. In addition, the character network considered here is an *undirected sparse network*. Sparseness here implies that $(M) \ll \binom{k_v}{2}$, where each node is connected to only a small fraction of the network, in comparison to a "fully connected" graph. This is a necessary condition for the notion of small world to make sense.

The following main structural characteristics used in this experiment are

listed below:

- *The total number of vertices:* $\mathcal{N}$
- *The total number of edges:* $\mathcal{M}$
- *Degree of a vertex:* k
- *The total degree of a network:* $\mathcal{K}$
- *The mean degree of a network:* $\bar{k}$
- *Degree distribution:* $P(k)$
- γ *exponent of a degree distribution.* $\gamma : P(k) \sim k^{-\gamma}$
- *The undirected shortest-path length:* $\mathcal{L}$
- *The average shortest-path length:* $\bar{\mathcal{L}}$
- *Clustering coefficient of a vertex* v*:* C_v
- *Clustering coefficient of a network:* $\bar{C}$

Experiment and Results

As mentioned, two statistical quantities are presumed to describe the *static* structure properties of this network: the *Path Length* $\bar{\mathcal{L}}$, and the *Clustering Coefficient* $\bar{\mathcal{C}}$. The path length $\mathcal{L}$ measures the typical "distance" $d(i, j)$ between two vertices of the graph. Another statistic, the clustering coefficient $\mathcal{C}_v$ of a vertex, measures the "density" of connections in the environment close to a vertex. Conventionally (Watts and Strogatz (1998); Watts (1999)), $\bar{\mathcal{L}}$ can be computed as $\frac{lnN}{lnk}$; $\mathcal{C}_v$ can be calculated as the ratio between the total number of edges in $\Gamma(v)$ (the total number y of the edges connecting its

nearest neighbours), and the total number of *possible* edges in $\Gamma(v)$ (the total number of all possible edges between all of these nearest neighbours),

$$\mathcal{C}_v = \frac{\textit{the number of direct links between neighbours of v}}{\textit{the number of all such possible links}} \tag{4.16}$$

and therefore reflects the 'cliquishness' of a typical neighborhood (Watts 1998). Further, the clustering coefficient of graph G is $\bar{\mathcal{C}}$, defined as the average of $\mathcal{C}_v$ over the entire graph G.

The *scale-free property*, on the other hand, is defined by algebraic behavior in the probability distribution $\mathcal{P}(k, N)$ of k. Since the character network in this experiment is *undirected*, and vertices can be distinguished, for each vertex we can obtain degree distribution $p(k, s, N)$. This is the probability that the vertex s in the network of size N has k connections. Knowing the degree distributions of each vertex in a network, the *total degree distribution* can be calculated as:

$$P(k, N) = \frac{1}{N} \sum_{s=1}^{N} p(k, s, N) \tag{4.17}$$

The first moment of distribution, that is, the mean degree of this network is $\bar{k} = \sum_k k\mathcal{P}(k)$, and the total number M of edges in this network is equal to $\bar{k}N/2$ (Dorogovtsev and Mendes 2003).

Our first result is that this character network is highly clustered and at the same time and has a very small "length", i.e. it is a small world model in the *static* aspect. Specifically, $\mathcal{L} \gtrsim \mathcal{L}_{random}$ but $\mathcal{C} \gg \mathcal{C}_{random}$. Results for the network of characters, and a comparison with a corresponding random network with the same parameters are shown in Table 4.4.3.

Next, we consider the *dynamic* features of the character network. The distribution of the number of connections follows power laws that indicate a *scale-free* pattern of connectivity, with most nodes having relatively few connections jointed together through a small number of *hubs* with many

	$\mathcal{N}$	$\overline{k}$	$\mathcal{C}$	$\bar{L}$	ℓ_{max}
Actual configuration	6493	350	0.64	2.0	24
Random configuration	6493	350	0.06	1.5	19

Table 4.2: Statistical characteristics of the character network: $\mathcal{N}$ is the total number of nodes(characters), $\overline{k}$ is the average number of links per node, $\mathcal{C}$ is the clustering coefficient, $\bar{\ell}$ is the average shortest-path length, and ℓ_{max} is the maximum length of the shortest path between a pair of characters in the network.

connections. The degree distribution is plotted in log-log coordinates with the line showing the best fitting power law distribution. $P(k) \propto k^{-\lambda}, k \neq 0$. Here, λ is the exponent of distribution.

In conclusion, the character network we consider here shares the similar statistical features with other lexical resources both in *small-world* and *scale-free* structures: A high degree of sparsity; a single connected component containing the vast majority of nodes; very short average distances between nodes, high local clustering; and a power-law degree distribution with an exponent near 3 for undirected networks.

The real characters network - if it exists - could be more complicated than the thumbnail sketch presented here. However, the statistical regularities that we have covered in this section could be helpful in contemplating of the construction of Chinese lexical resources.

4.5 Conclusion

Summing up, in this chapter, we will restrict ourselves to formal models and their abilities of expression in relating to the conceptual/semantic structure of Chinese characters. As seen, the formal models discussed all have their own advantages and disadvantages as a explanatory framework for the conceptual knowledge representation of Hanzi. The choice of a proper representation depends on the particular problem involved.

Though the discussion of these formal models is far from composing a

theory of meaning or concept in general, it may serve as groundwork for such a theory. It is therefore necessary to recapitulate the points already made:

1 Formal language and finite-state models have been widely used in the formal analysis of Chinese characters. With their great success in many Chinese NLP tasks such as text-to-speech analysis and Chinese word segmentation, these models, which pay more attention to Chinese characters as graphic patterns instead of as meaning patterns, seem to us to be less interesting to meet our concern in this survey.

2 For the hierarchically organized semantic and conceptual information representation of Chinese characters, *Graph-based or network representations* have the advantages of generality over other more restrictive solutions, for network structures provide intuitive and useful representations for modelling semantic knowledge and inference. The semantic network model, which has been widely used in artificial intelligence for knowledge representation, is an appealing solution for semantic and conceptual information encoding.

Among the semantic network models, we focused on some widely known partial order relations such as *tree*, *lattice* and acyclic structures. The Concept Lattice, with its mathematical rigidity, has some advantages over other representations. But the degree of specification is problematic, and there has been no agreement on how many relationships between various types of concepts we should add to.

In addition, as Vossen (2003) observed, many lattice structures introduce a (very) large number of internal nodes for feature combinations that have no natural human interpretation, and hence no lexical expression in many languages. Whereas a lattice generates all logical concepts, language tends to lexicalize only those concepts that are efficient to support communication. We agree with Vossen's conclusion in that, typically, *formal ontologies* are

small, highly symmetric, and often developed top-down, whereas *large-scale ontologies* for NLP are often based on the less systematic lexicalization of concepts.

3 Empirical evidence is also necessary to motivate the construction of knowledge resource at the level of characters. In the statistical model, descriptive statistics of character components and their governing laws were introduced. At the word level, some statistical measures of the production and association of Chinese characters were introduced, which could be important parameters to use in the linguistic resource construction and some NLP tasks. The structure of a specific Hanzi-driven *semantic* network was also analyzed. It was found that this network, like many other linguistic semantic networks, such as WordNet and Roget's Thesaurus, exhibits a *small-world* property, characterized by sparse connectivity, small average shortest paths between characters, and strong local clustering. Moreover, due to its dynamic property, it appears to exhibit an asymptotic *scale-free* feature with the connectivity of power laws distribution, which is found in many other network systems as well. These results yielded the motive for the construction of a network for Chinese characters from the statistical point of view.

4 *The Candidate Solution*:
Therefore, given all of the matters discussed up to this point, a candidate solution is proposed for the formal representation of conceptual information in Hanzi: The *tree structure* provides a proper mathematical model for the Hanzi-driven conceptual hierarchy; while the *lattice structure* captures the salient features and characteristics of the semantic aspects of components in and between characters. The *Huffman code* tree-like encoding method is especially effective in the encoding of these information.

In the upcoming chapter, I will develop a Hanzi-based theoretical framework as well as software implementation based on the formal analysis in

this chapter. The validity of this selection will be examined throughout the remainder of this thesis.

Part III

Representation

Chapter 5

HanziNet: An Enriched Conceptual Network of Chinese Characters

This chapter is primarily concerned with representational issues, addressing, among others, a currently developed integrated knowledge resource concerning Chinese characters: HanziNet.

As introduced previously, while traditional Chinese philologists give emphasis to *characters* as the main focus of Chinese semantic studies, most modern Chinese linguists have acknowledged the *word* as the prime carrier of meaning. In this Chapter, it will be argued and proposed that with the integration of two perspectives, it could possibly provide a more sufficient description.

Based on this consideration, the goal of building a HanziNet are twofold: (1) to give each Hanzi in use a rigorous conceptual location, and a character conceptual network thereof and (2) to anchor HanziNet as a coupling interface between *Concept* and *WordNet* in a Chinese context, in order to facilitate lexical and knowledge processing.

The chapter is thus structured as follows: Section one introduces the motivation of the construction of HanziNet. Section two compares some recently proposed models concerning with Chinese characters. Section three describes some fundamental issues and proposes a theoretical model underlying the HanziNet. After these preliminary discussions, in section four and

five, the architecture of HanziNet, including basic design issues, components, Hanzi-grounded upper level ontology, and the coupling of WordNet will be discussed. Finally, I will close this chapter by discussing some issues in HanziNet ontology construction.

5.1 Introduction

- Why (not) Chinese WordNet?

The most widely used lexical resource for natural language processing might be WordNet,[1] which has been developed by George Miller (1995) and his colleagues (Fellbaum 1998a). Over the recent years, WordNet has grown into a large lexical database and has become a common designator for semantic networks of natural languages (Fellbaum 1998b). The success of the the Princeton WordNet seems to lie in the general framework it provided, and it has motivated the development of several other WordNet projects for numerous other languages. For example, an EC project called EuroWordNet is the building of a multilingual database with WordNets for several European languages, structured along the same lines as the Princeton WordNet (Vossen 1998). In the Chinese speaking world, some WordNet-like lexical database have been developed as well, for example, the Chinese Concept Dictionary (CDD) and HowNet.[2]

The Princeton WordNet is a lexical semantic network which contains information about nouns, verbs, adverbs and adjectives in English, and is built around the concept of a *synset*. A *synset* is a set of lexical units (e.g. words, compound nouns and collocations) with parts of speech that are synonymous, that is, these lexical units can be interchanged in a certain context. For example, {`animal, animate being, beast, brute, creature,`

[1] At the time of writing, the most updated version is WordNet 2.0, `http://www.cogsci.princeton.edu/~wn`.

[2] For other languages, See "WordNets in the World" at http://www.globalwordnet.org.

fauna} form a synset because they can be used to refer to the *same* concept, and such *sameness* can be described in a synset by a gloss: "a living organism characterized by voluntary movement". Often examples are provided as well. These *synsets* are linked with each other via *semantic relations*, such as hyperonymy, hyponym, meronym, holonym, antonym, etc., and other relations such as *entails*, *causes*, and *derivational related relations*. By means of these relations, all of the senses of words can be interconnected and thus constitute a huge semantic network.

Indeed, WordNet as a useful lexical knowledge resource has been proved, and widely applied to many NLP fields. However, in the Chinese case, an envisaged Chinese WordNet might have some deficiencies in its expression.

- **Family resemblance** of characters

  As discussed previously, due to its ideographic property, the inner structure of Chinese characters (e.g head component and body component) might possess some idiosyncratic properties (e.g. a *small world* network).

  The analogy of *family resemblance* could be used here to illustrate the case. In a family of Chinese charaters, every member shares some similar traits (i.e., components) with some other members, but there is no common components among the whole family. These relations would not be expressed in WordNet. However, this insufficiency would have an unfavorable influence on the surveys of so-called "cognate characters" proposed in the traditional study of Chinese scriptology.

  "Cognate characters" are defined as characters which share the same sound or meaning components, for instance, {喬(high)、驕 (arrogant)、橋 (bridge)、撟 (raise)} constitute a set of "cognate characters", for they share the same sound component 喬, which also share a similar *core meaning* ("top-down"; "situated at the top") carries by the compo-

shared component and meanings	cognate characters
鳥 (bird)	鳩 (dovelet); 鵬 (roc); 鶯 (parrot); 鴿 (dove)
行 (road;action)	街 (street); 衢 (thoroughfare); 衝 (dash onward)
舟 (boat)	船 (ship); 航 (navigate); 舫 (boat); 舷 (shipboard)
雨 (rain)	雪 (snow); 霜 (frost); 雲 (cloud); 雹 (hail)
保 (protect;bring up)	堡 (fort); 褓 (swaddling clothes); 媬 (baby sitter); 葆 (nurture)

Table 5.1: Cognate characters

nent.[3] Some more examples are given in Table 5.1.[4]

- **Conceptual relatedness** of characters

  As known, the most majority of Chinese *words* are composed of two *characters*. More and more Chinese morphological studies have revealed the concept formation / semantic patterns of dissyllabic words *driven* by characters (Lua 1993; Wong 2004; Xu 2004; Chu 2004). Takes an example of 取(/qǔ/,"take").[5] It can combine with some other characters to form words, in order to represent various ways, attitudes, purpose, means, objects and locations concerning with the action "take". E.g., bá-qǔ ("pull-take": eradicate), liè-qǔ("hunt-take": hunt), caí-qǔ ("pick-take": adopt), duó-qǔ ("rob-take": take by force), gōng-qǔ ("attack-take": attack and seize), huàn-qǔ ("exchange-take": exchange st. for), jì-qǔ ("record-take": bear st. in mind), boú-qǔ ("win-take":

[3]In the Western lexicological tradition, we might imagin them as phonaesthemes, for example, 'sl-' in 'slow', 'slur', 'slack', etc. suggests a kind of 'sluggishness'. However, they are only faintly suggestive of certain meaning associations.

[4]This should not be confused with the notion of "block characters". Conseder the example: The character 口 (/kǒu/,"mouth") is also a component of a number of other characters, such as {言(speech), 味 (taste), 語 (language), 裕 (plentiful), 邑 (state), 右 (right)}, to name just a few. The ability of one character to appear in multiple positions in another character makes it a block character (two dimensional). "block characters" resemble each other only in shape, so they have a beneficial use in character teaching and learning. The best explanation of "block characters" can be found in Harbaugh. (1998). Chinese characters: A genealogy and dictionary. Online version available at `http://zhongwen.com`

[5]These data are taken from Su (1995:191)

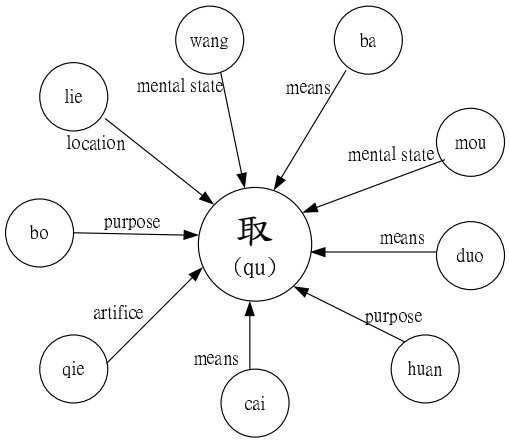


Figure 5.1: Conceptual relatedness of characters: An example of qǔ

try to gain), qiè-qǔ ("steal-take": usurp), wàng-qǔ ("unreal-take": vainly attempt)..., to name just a few. Figure 5.1 schematizes these examples.

So in constructing Chinese WordNet, it would be meaningful to utilize various levels of semantic information encoded in the Chinese writing, including, but not limited to: components, radicals, *roots* (in compounds).[6]

Such considerations might offer a starting point for building a semantic network at the character level, which aims to explore the ideosyncratic knowledge structure of Chinese ideographic writing. But it is important to note that HanziNet is *not* proposed as an alternative to the Chinese WordNet. It is rather regarded as a kind of complementary knowledge resource for NLP. In addition, We believe that the construction of such a network will shed light on the polysemy, morphological compounding, and even idioms as ways of expressing concepts in Chinese.

[6]In the 1st Workshop of *Foundational Issues in Chinese WordNet Research and Construction*, Huang (2002) shares the similar opinion.

5.2 Chinese Character Network: Some Proposed Models

This section introduces some possible design ideas proposed in relation to models of Chinese characters in recent years. These models can be classified into four groups according to the unit of analysis: morpheme, features, radicals and characters themselves.

5.2.1 Morpheme-based

Yu et al (1999) reported that a Morpheme Knowledge Base of Modern Chinese according to all Chinese characters in GB2312-80 has been constructed by the institute of Computational Linguistics of Peking University. This Morpheme Knowledge Base has been later integrated into the project called "Grammatical Knowledge Base of Contemporary Chinese".

It is noted that the "morphemes" adopted in this Database are monosyllabic "bound morphemes". "Free morphemes", that is, characters which can be independently used as words, are not included in the Knowledge Base. Figure 5.2 shows an example. The *monosyllabic character* 梳(/shu/;"comb") has two senses. For the verbal sense ("to comb"), it can be used as a *word* (also see the example sentence (a)); for the nominal sense ("a comb"), it can only be used in combining with other morphemes (sentence (b)).

(a). 妳梳過頭髮了嗎?

ni^3 shu^1 kuo^4 $'tou^2fa^3$ le^0 ma^0?

YOU COMB PERFECT-PARTICLE HAIR PARTICLE QUESTION-PARTICLE?

Have you combed your hair?

(b). 桌上有把梳子。

$zhuo^1$ $shang^0$ you^3 ba^3 shu^1zhi^0

TABLE ON HAVE MEASURE TERM COMB

There is a comb on the table.

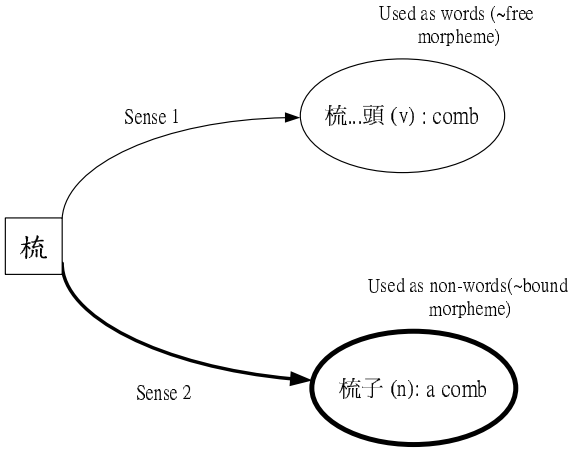


Figure 5.2: "Bound" and "free" morphemes: An example of comb

5.2.2 Feature-based

Morioka Tomohiko (2005) has proposed a "Chaon" model which is a fundamental part of the open-source CHISE (Character Information Service Environment) project conducted by Kyoto University in Japan.[7]

Chaon model is a character processing model based on character information. The main idea in this model of character representation is that, characters can be represented as *character objects*, and character objects are defined by *character features*. As known, there are various kind of information related with characters, such as shape structure, phonetic values, semantic values, code points in various character sets, etc, This model regards these various things as character features, so each character is represented by a set of the features it has. Figure 5.3 shows an example of a Venn diagram of character objects.

5.2.3 Radical Ontology-based

Nowadays, many semantic resources, such as WordNet, EuroWordNet, Cyc and HowNet, have used a hierarchical structure of language independent con-

[7]More information about the project and software development tools is available at http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/

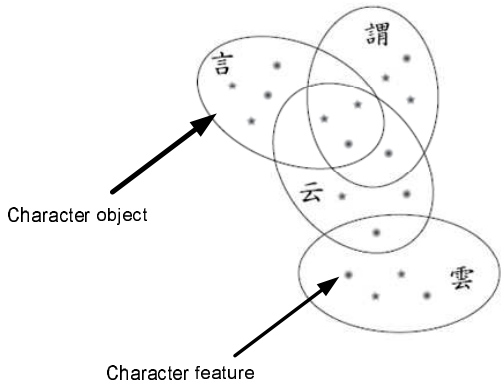


Figure 5.3: Venn diagram of characters: Chaon model

cepts to reflect the crucial semantic differences between concepts. This kind of hierarchy, with details and structure that enables computer to process its contents, is called *ontology*. Formally, an ontology consists of a set of concepts, relations, and axioms that formalize a field of interest.[8]

With the development of the next-generation World Wide Web - the Semantic Web -, ontology-based approach has become one of the main concern in the NLP. Building a linguistic resource that bridge intuitive lexical use to structured knowledge is now one of the current challenges facing computational linguists (Huang et al 2004).

In this context, Wong and Pals (2001) reported the result of comparing a selected collection of *Chinese radicals* and their meanings with *TOP-level Ontology* developed in the framework of EuroWordNet 1,2 project (EWN). In a similar way, Chou and Huang (2005) linked the Chinese radicals to the Suggested Upper Merged Ontology (SUMO) constructed by the IEEE Standard Upper Ontology Working Group and maintained at Teknowledge Corporation.[9]

[8]More details about ontological issues will be elaborated in section 5.5

[9]SUMO and its domain ontologies have formed the largest formal public ontology in existence today. It is the only formal ontology that has been mapped to all of the WordNet lexicon. More details can be seen at http://www.ontologyportal.org/

5.2.4 Hanzi Ontology-based

In comparison with the above approaches, Chu (1998) proposes a "Conceptual Hierarchy" (which will be called CBF ontology hereafter), and proposes that each Hanzi can be associated with the terminal nodes in this ontology. This approach has been introduced in Chapter 3, and will not be repeated here.

5.2.5 Remarks

The above-mentioned approaches seize diffirent aspects of Chinese characters. For the purpose of NLP, they have both advantages and disadvantages. The *Morpheme-based* approach is highly linguistically motivated, but it must risk the difficulties in clearly differentiating *free* and *bound* morphemes in Chinese.[10] In addition, due to the lack of an accompanying ontology, the semantic and conceptual relationship between "morphemes" are not easy to obtain. The *Feature-based* approach, with the aim of efficient character encoding in the information exchange, suffers the knowledge-poor disadvantage as well.

The *Radicals-Ontology based* approach has its advantage over the aboved two approaches, but the difficulties are not much less. For example, how many radicals are there? how to decide the meaningful radicals (because some of the radicals do not give a hint at meaning any more)?... and so on. As for the *Hanzi-Ontology based* approach, though with its clearness in selecting character itself as the basic unit, does not provide a sound explanation concerning with the relation between *characters*, *words*, as well as their roles in understanding the distinction between the *concept* and *meaning*, let alone the discussion of the problem of polysemy, ambiguity and homograph.

In contrast to the above-mentioned individual models, in the following, I will take a synthesis approache. In the next section, I will propose some

[10]This will be discussed in Chapter 6.

theorectical assumptions underlying the construction of HanziNet.

5.3 Theorectical Assumptions

5.3.1 Concepts, Characters and Word Meanings

Before embarking on the theorectical assumptions underlying HanziNet, it is necessary to address some certain background of the theory.

Chinese Wordhood

In the past few years, there has been a growing interest in the field of lexical semantics. Linguists and psychologists have been especially interested in the study of *word* senses in order to shed light on important aspects of human communication, such as concept formation and language use. Research in lexical semantics is rather heterogeneous as far as scope, methods, and results are concerned, but it shares the same starting point: the word-based perspective. Central to a natural language processing system is also a word-store, the lexicon. Under the influence of this theoretical trend, *wordhood* in Chinese has thus become an issue of urgency and many studies have ensued.[11]

We are not going to fall into a fixed position with regard to speculating about whether a *word* constitutes a real or an epiphenomenal construct, nor about the acquisition of the lexicon. The question we are interested in this thesis is rather, could we have an *explanatory framework* for clarifying the different ways in which *concepts* are *lexicalized* in Chinese?

An Integrated Pyramid Model

The model I want to propose here is called the integrated *Pyramid Model* for Chinese processing, whose key aspect involves a *fusion* of natural language

[11]There are, of course, contras. The famous Chinese linguist, Shuxiang Lü (1981:45) had a strong opinions, "..the reason why one cannot find a satisfactory definition for the Chinese 'word' is that there is no such thing in the first place. As a matter of fact, one does not need the notion 'word' in order to discuss Chinese grammar".

and conceptual information processing with Hanzi. In other words, it is aimed at bridging word-based natural language semantics and conceptual knowledge representation via the HanziNet. Thus, some enriched linguistic information for NLP tasks might be obtained.

The main underlying assumptions are as follows:

- *Concepts and Meanings are different*

  People *translate* the real world they perceive into a set of *Concepts*, in that there is likely to be considerable agreement over what these concepts are, even among people speaking different languages. The *Meaning* of a linguistic expression overlaps with its *Concept* to a certain extent, though not necessarily totally. The main difference lies in the fact that *Concepts* are more abstract and *prima facie*, while *Meanings* are mostly determined and used according to the pragmatic, social and cultural context.[12] Take a more extreme example, though the term "vegetable" refers to the *concept* of concrete → natural → plant, it means also "a person who is regarded as dull, passive, or unresponsive" in modern English, while in Chinese, it has a meaning for "beginner".

  As seen, due to the fuzzy boundaries between concept and meaning, though the term *conceptual view of meanings* might be psychologically sound, it still seems difficult to implement in the field of computational linguistics. In surveying the meaning representation styles adopted in the computational linguistic literature, Paola Velardi et al (1991) found that many natural language processors implemented adopt one of the following two meaning types for the lexicon: *conceptual (or deep)* and *collocative (or superficial)*, which I quote as follows:

  Conceptual meaning Conceptual meaning is the cognitive content of words;

[12] As Buitelaar (1998:17) mentioned, though the *conceptual level* is hard to identify, the assumption of a conceptual level could help to "liberate" lexical semantics from a formal semantic harness, that stresses in particular compositionality and looses sight of the wider association that most lexical items have.

it can be expressed by *features* or by *primitives*: conceptual meaning is "deep" in that it expresses phenomena that are deeply embedded in language.

Collocative meaning What is communicated through associations between words or word classes. Collocative meaning is "superficial" in that it does not seek "the real sense" of a word, but rather "describes" its uses in everyday language, or in some subworld language. It provides more than a simple analysis of co-occurrences, because it attempts an explanation of word associations in terms of *meaning relations* between a lexical item and other items or classes.

Both conceptual (defining) and collocative (syncategorematic) [13] features are formally represented in NLP literature using some subjective, human-produced set of primitives (such as *conceptual dependencies, lexical semantic relations, and conceptual categories*) about which there is no shared agreement at the present time (Paola et al. 1991). However, collocative meaning can rely on the solid evidence represented by word associations, It has been proven to be a useful knowledge resource for many NLP applications, WordNet being an obvious example.

In fact, as Velardi et al (1991) mentioned, the inferential power of collocative meaning representation is *lower* than that for conceptual meaning, because it does not account for many important aspects of human communication. However, due to the lack of trager, *"conceptual meanings"* are difficult to trace, construct and verify, and seems destined to remain in the area of subjective conjecture.

- *Characters are relatively objective cues to concepts*

  It is rarely asserted that "conceptual information" – the information about objects and events in the world – can be measured *directly* from linguistic expressions. But as introduced previously, unlike most natural languages, the Chinese language displays a considerable amount of semantic information at the character level. Based on the comparison of

[13] The better term would be "relational" here.

this distinctive feature of Chinese radicals and WordNet/EuroWordNet Top Ontology, Wong and Pala (2001;2002) also suggests that the system of Chinese characters might contain a rich but concise system of inter-related concepts.

So here it will be assumed that in Chinese, basic *conceptual information* are implicitly encoded in its characters. This means that, we can *interpret* the set of Hanzi as a set of *concept primitives*.[14] In other words, characters $\simeq$ the core and original meaning units. When a character is used as a monosyllabic *word*, or as a part of a disyllabic *word*, the meaning - though mostly overlapping with its concept -, often leads to ambiguity and polysemy while using it in different contexts after a long period of time. For example, according to the most classical etymology dictionary 說文解字 (Shuo-Wen-Jie-Zi), the character 舉 (/jǔ/) carries the core meaning of "lift up", but in modern Chinese, it has many derived meaning facets while combining with other characters, such as 舉起 (/jǔ-qǐ/, "hold up"), 舉頭 (/jǔ-tóu/,"face upwards"), 推舉 (/tūi-jǔ/,"recommend someone") and 檢舉 (/jiǎn-jǔ/,"report to the authorities").

- *Characters are interface of concepts and word senses*
  The "concept signals" are to be filtered and fused - by being processed through many layers of computation - before they can be detected and associated with word semantics. Figure 5.4 shows a schematic diagram of the proposed model. These networks have several levels of structural organization, each with a distinct scale. From the top down, the first layer structure is **concept chaos**, the second layer is a **tree-like Hanzi-driven TOP level Ontology**,[15] and the bottom layer is the long-range highly-correlated *random acyclic graphic structure* of the

[14]However, the "homograph" problem is not to be escaped. This will be discussed later.
[15]For technical reasons, the tree-like structure is not clearly depicted in Figure 5.4.

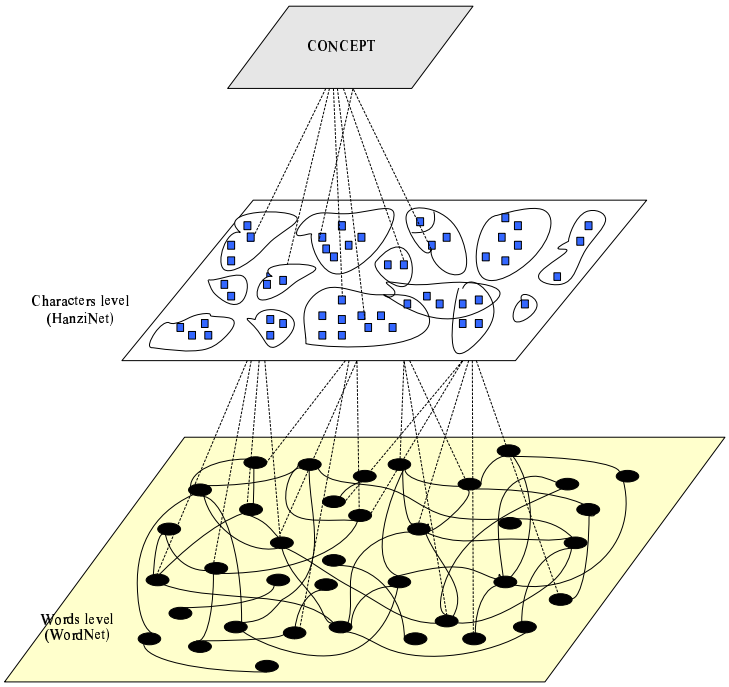


Figure 5.4: The pyramid structure model

word-based semantic network. Each of these layers is characterized by its properties of the networks.[16]

Figure 5.5 schematizes further the different representations of the middle and bottom layers. In Aitchison's (2003) terms, for the character level, we take an "atomic globule" network viewpoint, where the characters - realized as *core concept units* - which share similar conceptual information, cluster together (as do the clouds in Figure 5.5). The relationships between these *concept units* form a tree with $2^n (n = 1, 2, 3, 5, 8)$ branches. Characters are thus assigned to the leaves of the tree in terms of an assemblage of binary bits. For the word level, we take the "cobweb" viewpoint, as *words* -built up from a pool of characters- are connected to each other through semantic relations. In such case, the network does not form a tree structure but a

[16]The network model I propose here corresponds to findings from other disciplines, e.g. Dorogovtsev and Mendes (2003) has came to a similar conclusion from the view point of statistic physics.

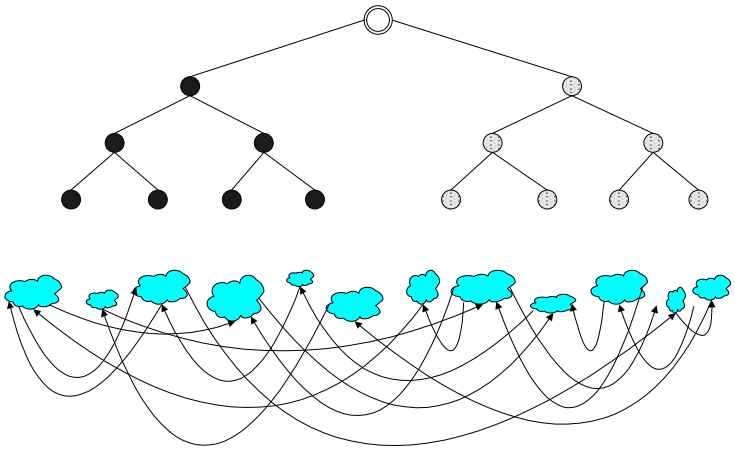


Figure 5.5: *Character-based* concept tree and *word-based* semantic clouds

more complex, acyclic structure (DAG).

5.3.2 Original Meaning, Polysemy and Homograph

In designing a conceptual network based on Chinese characters, one might hasten to doubt that even we have a list of "conceptual primitives" represented by Hanzi, it is hardly sufficient in itself, because Chinese characters, no matter they function as free or bound morphemes in Chinese *words*, can not keep away from the problems of ambiguity. And the notion of "conceptual primitives" seems to contradict the "ambiguity".

The polysemous phenomena of characters can be traced back to their historical development. In traditional Chinese scriptology, *meanings* of characters were classified into two types: *original* and *derivational* meanings. Derivational meanings of a character are systematically related centered on the original (or prototypic) meaning. Until now, some researchers argue that such framework still holds in modern Chinese. In terms of lexicology, for every character currently used, we could find its *systematic polysemy*, where senses are systematically related and therefore predictable over classes of lexical items.[17]

[17]It is fundamentally different from homonymy where senses are unrelated, non-

However, on closer inspection of many classical and modern dictionaries, I found that many modern usages of Chinese characters, though they might be provided with refined etymological intepretation, are still difficult to be linked together around a single core meaning. In order to make HanziNet more applicable rather than just a historical semantic record, the shallow syntactic tags (mainly nominal and verbal) are thus used to resolve this problem, which results in that characters with different syntactic tags will be considered as having different concepts.

Another important distinction is concerned with *homograph*. A homograph is defined here as a character that has the same shape as another character. Homographs differ from each other in meaning and pronunciation. For example, the characters 將(/jiāng/,"be going to") and 將 (/jiàng/,"general"), which have the same shape but differ in pronunciation and meaning. Homographs will be treated as different characters.

Based on these consideration, I will take a position which accords with the traditional Chinese scriptology, and meanwhile, tries to preserve harmony with Chinese WordNet of any form. Table 5.2 illustrates an example of 會. Due to the homograph, two character entries for 會 are listed. And due to the different syntactical behaviour, two concepts are defined for the first 會. The concept of character 會 with syntactical tag N is "gathering". It is the "original" meaning which is disseminated when character 會 combines with other characters.

It is noted that only the first four columns will be discussed within the framework of HanziNet.

systematic and therefore not predictable. On the theorectical development of polysemy, cf. Buitelaar (1998) and Pethő (2001).

char.	pron.	syn.	concept	(word) meaning
會$_1$	huì	N	gathering	聚會(assemble); 都會(metropolis); 開會(meeting)
	huì	V	be able to	不會(unable)
會$_2$	kuaì	N	accounting	會計(accounting)

Table 5.2: An example of hui

5.3.3 Hanzi Meaning Components as Partial Common-Sense Knowledge Indicators

As the title of this Chapter indicates, the HanziNet proposed here is an *enriched* conceptual network of Chinese characters.[18] By this we do not intend to design a system with higher numbers of *relations*, but consider also the "common-sense knowledge" within Hanzi as briefly mentioned in previous Chapters.

The notion of "common-sense knowledge" is very difficult to pinpoint exactly, and frequently means different things to different people. In principle, the term describes *beliefs* or *propositions* that seem, to most people, to be prudent and of sound judgement, without dependending on esoteric knowledge (definition taken from WIKIPEDIA). For example, we (even as children) know that a "fish" is an animal with a certain similar form, which lives in the sea.

In some cases, "concept", "commonsense knowledge", "sense" and "domain knowledge" can be roghly differentiated. Take 火 /huo/ (fire) for example, its *concept (conceptual meaning)* might be a `objective-concrete-natural things`-...; People's *commonsense knowlege* of it might be substance with properties of *hot, untouchable, bright...*; while its senses facets in words varies in diffirent context, such as 關火 ('close-fire', turn off the stove'), 火爆 ('fire-explode', fiery), 火氣 ('fire-gas', temper)、惹火 ('provoke-fire', stroke the wrong way).... its (chemical) domain knowledge could be: A form of energy

[18]In fact, it can also be called an enriched semantic network as other researches did. It just depends on the angle you see this knowledge source.

associated with the motion of atoms or molecules and capable of being transmitted through solid and fluid media by conduction, through fluid media by convection, and through empty space by radiation.

But in most cases, "common-sense knowledge" (CSK, hereafter) overlaps with "concept". They both span a huge portion of human experience, encompassing knowledge about the physical, social, temporal, spatial and psychological aspects of our everyday life. But, in a strict sense, CSK is largely *defeasible*, *context-sensitive* and more *flat* and *semi-structured* in its formal property (Liu and Singh 2004). As a result, it is more difficult to seize by computer due to the property of fuzziness. Nonetheless, much efforts have been made to show that CSK is an indispensable component in a Natural Language Understanding system.[19]

It is widely believed that Chinese characters, especially their semantic components (called the "radicals" or "character heads") carry abundant CSK information. By way of *CSK factoring* via character components, i.e. the process of semantically analyzing character components into a collection of features proposed by Chu (1998), we found that there are a number of CSK features that are implicitly expressed by the character semantic components.

In the following, some examples are shown. The first column consists of examples of the character semantic components, and the second colume lists their respective CSK feature representations and glosses. I use capital letters to represent these "CSK features", subscript numbers for the subclassification. For example, A is the category of (LUMINOUS SOURCE), B:(TEMPERATURE), C:(SHAPE), D(QUALITY), E:(COLOR), I:(ORGANIC ORGANIZATION), J:(SUBSTANCE), L:(FUNCTION), M:(SPACE) and so

[19]Currently, there are some on-going projects which attempt to construct a basis of CSK for AI systems. For example, Cyc (http://www.cyc.com). The MIT Media Lab has also embarked on an effort to give comupters "common sense", the capacity to understand and reason about the world as intimately as people do. Please refer to Commonsense Computing @ Media at http://csc.media.mit.edu/

on.[20] A_1 stands for *strong* light, A_2 for *weak* light, A_3 for *glisten*.... and so on.

日	$A_1B_1C_5J_1K_1$
月	$A_2B_2J_1K_2$
貝	$C_4D_2E_1J_5L_2$
門	C_1C_2
骨	$C_3D_2I_5M_1$
目	$C_4C_5E_2I_6M_2$

Formally, conjunctions of these CSK features with Character semantic components (mainly CH) generate a *lattice* hierarchy, which can be formalized using the FCA method discussed earlier (Figure 5.6). In many cases, characters in the same "cloud" in Figure 5.5 could be further differentiated by the CSK features carried by the character semantic components. Table 5.3 shows a set of characters with similar conceptual information, which can be further classified using their CSK codes. That is, conceptual code and semantic code of a Hanzi might be regarded as *necessary* and *sufficient* conditions of the understanding of it. The former are used to identify a type of concept and the latter are used to distinguish the instance from other instances in the same set of concept type.

The caveat at this point should be carefully formulated, due to the restricted (or partial) expressive power of HanziNet in representing CSK. The CSK which Hanzi indicates is *primitives-based* rather than *relation-based*. Therefore, CSK such as "Birds can fly" or "Magpies are birds" can be indicated via Hanzi, but for cases like "If you drop an object, it will fall", or "If you forget someone's birthday, they may be unhappy with you", Hanzi can not say anything of it.[21]

[20] A more detailed list will be given in the Appendix.

[21] This deficiency could be remedied by integrating relation-based common-sense knowledge base such as ConceptNet (http://web.media.mit.edu/~hugo/conceptnet/).

Characters with similar concepts	Concept code	CSK code
說 (speak)	11101000	(言) $C_8H_1H_5L_1$
謂 (be called)	11101000	(言) $C_8H_1H_5L_1$
講 (explain)	11101000	(言) $C_8H_1H_5L_1$
訴 (inform)	11101000	(言) $C_8H_1H_5L_1$
告 (tell)	11101000	(口) $H_1H_5I_7$
述 (state)	11101000	(辶) F_3L_1
道 (chatter)	11101000	(辶) F_3L_1

Table 5.3: Concept differentiation via CSK

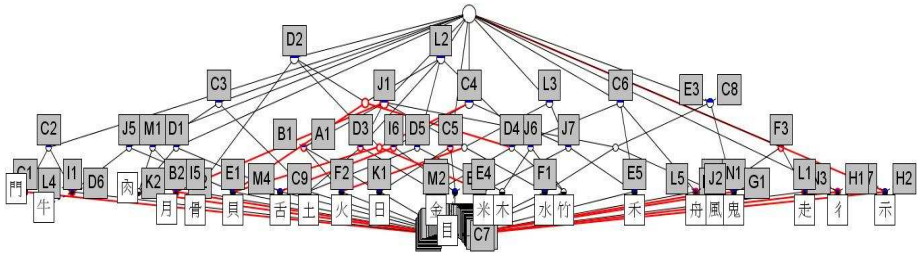


Figure 5.6: A common-sense knowledge lattice

To sum up, the fundamental suggestion of the present proposal is to treat the set of characters as a *coupling interface*, a *conceptual mediator* and a *commonsense knowledge indicator* as well. The coupling of these layers constitutes a picture of Chinese natural language understanding. To make this proposal a bit more tangible, the remainder of this chapter is devoted to the architecture and its actual implementation.

5.4 Architecture

5.4.1 Basic Design Issues: Comparing Different Large-Scale Lexical Semantic Resources

The HanziNet we have proposed here is an *enriched conceptual network*. In the expression *conceptual network*, the adjectives "enriched" and "concep-

tual" are to be understood in the sense proposed in previous discussions. It is developed as a system combining the design of a **character data dictionary** and a **character-driven Top-Level-Ontology**.[22]

In order to elaborate on the proposal, this subsection spells out the design ideas of HanziNet, by comparing two main large-scale lexical resources: WordNet and Roget's Thesaurus (Roget's International Thesaurus, 4rd Edition, 1977).

WordNet has already been introduced, so we will skip to the second resource. A Thesaurus is a hierarchical representation of a lexicon, in which words are arranged according to their semantic similarity. There have been attempts to construct semantic thesaurus, among which the Roget's Thesaurus is a representative one. Roget's Thesaurus has been described as a synonym dictionary, but it is different from other alphabetically organized synonym dictionaries in that other dictionaries lack the hierarchical conceptual organization of Roget's Thesaurus. Paul Mark Roget intended his thesaurus as a classification of words by the ideas they express, and as such, it has been described as a conceptual hierarchy (Old 2003:15).

Though in contrast with WordNet and Roger's Thesaurus, the construction of lexical knowledge resource in terms of Chinese characters does not depend on a vast computational infrastructure, the design of the HanziNet raised a whole new raft of issues in the early stages of development. In the following, I will focus on three issues:

- How should the HanziNet be constructed? By hand or automatically?

The experience of WordNet or some other lexical knowledge resources has shown that, the construction of a generic semantic resource is not a trivial task. In particular, the construction of an ontology is also a very expensive

[22] "Character-driven" here means that this ontology is constructed based on the induction and introspection of the cocnceptual information carried by Chinese characters. More details on the ontology design will be discussed in section 5.5, and all data will be made available in the Weblog http://www.hanzinet.org.

and time-consuming undertaking. Considering the limited manpower and time, a few freely available dictionaries such as Cedit,[23] Unicode UniHan fiels,[24] Free Chinese dictionary[25] and CBF Hanzi dictionary[26] have been taken over as the prototype character dictionary in the HanziNet, so we don't have to build it from scratch. However, many specific changes have been made to the design of the database. For the moment, the database contains about 5600 characters, covering the most used characters in modern Chinese. As for the ontology construction, I take a bootstrapping strategy, which will be discussed later.

- What kind of information should the HanziNet contain?

Chinese characters have a history of over 5,000 years. Each character constitutes a synchronic small network with diachronic cultural-historical background.

In practice, what should be encoded in a *character-stored lexicon* depends largely upon the aim of applications. For the purpose of ontology-based NLP, the set of characters is seen in HanziNet mainly as the embodiment of a tree-like structure configuration of conceptual information.

In comparison, a crucial difference between WordNet and HanziNet rests with the information content represented by the nodes of network. A **synset**, or synonym set in WordNet contains a group of words,[27] and each of which is synonymous with the other words in the same synset. In WordNet's design, each synset can be viewed as a *concept* in a taxonomy, While in HanziNet, we are seeking to align Hanzi which share a given putatively primitive meaning, so a new term **conset** (concept set) is proposed. A *conset* contains a group

[23] `http://www.mandarintools.com/cedit`
[24] `http://www.unicode.org`
[25] `http://home.swipnet.se/~w-123680/`
[26] `http://www.cbflabs.com/book/dic/hanzijiyin2/a0.htm`
[27] To put it exactly, it contains a group of lexical units, which can be words or collocations.

of *Chinese characters similar in concept*,[28] and each of which shares with similar conceptual information with the other characters in the same conset.

As discussed earlier, we would tend to think of a concept as information, measurable in bits. So "*characters similar in concept*" in this context means that characters that are in the same *conset*, i.e., characters that have the same *conceptual code* based on the conceptual hierarchy (ontology). For instance, 說 (speak), 道 (chatter), 曰 (say), 云 (say), 告 (tell), 訴 (inform), 講 (explain), 敘 (narrate), 謂 (be called), 述 (state), these characters are assigned to the same conset with the binary code `11101000`. For every *conset*, short definitions are also provided.

In WordNet, words typically participate in several synsets. In HanziNet, characters can also participate in several *consets* (please refer to section 5.3.2); WordNet distinguishes syntactic categories between nouns, verbs, adjectives and adverbs on the assumption that these are stored differently in the human brain. For HanziNet, only shallow syntactical information are provided. In addition, WordNet does not include information about *etymology*, while HanziNet does.

- How should the contents be organized and made accessible?

At this point, a discussion of the differences between 字典 (character dictionary, glyphography) and 詞典 (conventional dictionary, lexicography).[29] within

[28]In fact, in addition to individual character, it also contains and a few dissyllabic morphemes collected by Sproat (2000:149-150). In general, these dissyllabic morphemes are not listed as a entry in normal dictionaries. I quote Sproat's explanation : " ..., the reason for the relative neglect of dissyllabic morphemes comes from the fact that traditional Chinese dictionaries are organized around monosyllabic characters, not around words or morphemes. Since meanings are traditionally listed in the dictionary as entries for these characters, this obscures the fact that many characters are in fact meaningless unless combined with a specific second character".

[29]Due to the mixed structure of its writing system, contemporary Japanese culture favors the separation between character dictionaries and conventional dictionaries. Conventional Japanese dictionaries are structured the same way as those of European languages, and deal with lexical meanings, maximizing the quality of semantic analyses. While in character dictionaries (usually called *Kanji dictionaries*,) Kanji entries contain mainly the meanings of the morpheme.

the context of Chinese lexicography would seem to be helpful.

In both kinds of dictionaries, the collection of information associated with an item, that is, the item itself and its linguistic specification, is referred to as a *lexical entry*. But the arrangement of the lexical entry differs. Conventional dictionary contains an alphabetical list of words, with orthographic and linguistic information given for each word, usually including its meaning, pronunciation, modern usage, etymology and so on. A character dictionary, in contrast, contains lists of characters in a non-alphabetical order (usually arranged by radical) , with the *similar* information as provided in the conventional dictionary.

For instance, in the great KāngXī character dictionary, which was complied in 1716, there were around 48,000 characters alloted to one of the 214 radicals. The characters assigned to a radical are listed under it in ascending order according to the number of residual strokes. This method, with some minor modifications, is still used in many traditional Chinese character dictionary and in Japanese Kanji Dictionaries as well.[30] Since HanziNet is designed to be a conceptual network, the characters are ordered according to the *conset* they belong, instead of ordered according to the KāngXī Radicals.

In order to be able to efficiently **access** the lexical information, WordNet provides an interface which is able to deduce the root form of a word from the user's input, for only the root form is stored in the database. In HanziNet, only characters, which are not necessarily words, are stored, and all of them are equipped with CāngJié code, so the CāngJié input method is the most efficient access to the HanziNet.[31]

However, within the field of NLP, lexical entries have evolved from simple pairings of phonological forms with grammatical categories into elaborate

[30]However, there is a trend in modern Chinese lexicographical practice, which indicates that the distinction between these two kinds of dictionary compilation is no longer rigid. A conventional dictionary contains a hybrid structure of a character dictionary and an ordinary dictionary.

[31]For the reason of this consideration, please consult Chapter 3.

Table 5.4: A comparison of explicit structure of different lexical resources

Word association	Ord. Dictionary	Thesaurus	Char. Dictionary
Word A	Word A	Sense 1	Character 1
- Word B	- Sense 1	- Word A	- Word A
- Word C	- Sense 2	- Word B	- Word B
- Word D	- Sense 3	- Word C	- Word C
Word B	Word B	Sense 2	Character 2
- Word D	- Sense 2	- Word A	- Word D
- Word C	- Sense 5	- Word D	- Word E
- Word E	- Sense 6	- Word G	- Word F
.....	.....	.....	.....

information structures, usually formatted in some knowledge representation formalisms in favor of being manipulated by a computer. However, if we look at the representation techniques involved and the content of the representations, numerous dissimilarities emerge (Handke 1995:50). For instance, the **organizational structures** of the various current lexical resources differ enormously. Based on Old (2003), I will inspect these organizational structures both from *explicit* and *implicit* perspectives. Table 5.4 shows the comparison of explicit structures between *word association, ordinary dictionary, Roget's Thesaurus*, and *character dictionary*.

Figure 5.7 shows the explicit structure of HanziNet. From the perspective of explicit organization structure, the main difference between *traditional character dictionary* and HanziNet lies in that the latter re-organizes the characters by inserting them in the level between conceptual level and word level, which is in accordance with the pyramid model proposed before.

The organization principles from *implicit* perspective might not be clear at a glance. Over the years, though several lexical databases have been developed, these databases differ in their detailed organisation of conceptual and lexical knowledge.

Roget's Thesaurus's (hereafter, RT) organizational structure is a classifi-

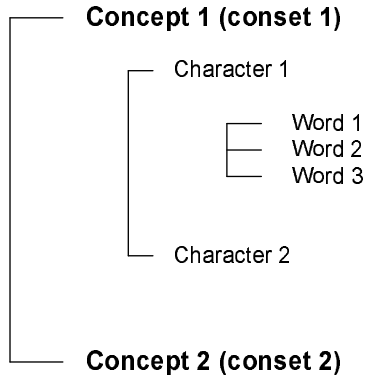


Figure 5.7: The explicit structure of HanziNet

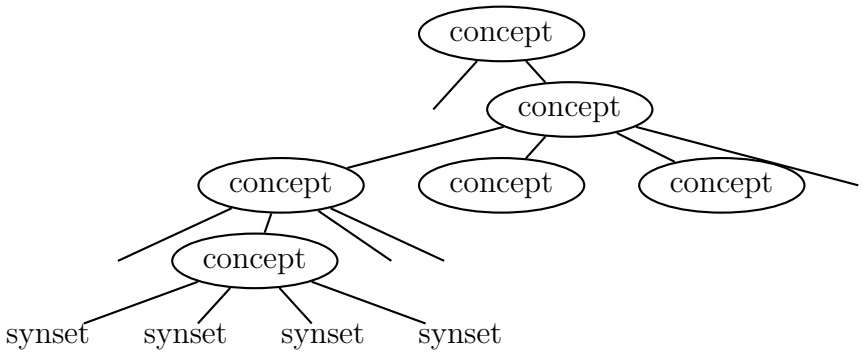
cation tree, or conceptual hierarchy. It contains a six-level classification, and at the lowest level, words are grouped that are either synonymous with each other or closely related according to some semantic field, such as animals or food items. Because of polysemy, many words occur multiple times in the thesaurus.

In contrast to RT, the synonym sets in WordNet occur at all levels of the hierarchy, not just at the bottom. It should be noted that the hyponymy hierarchy is only one relational structure in WordNet. Not all of the other relations in WordNet are hierarchical.

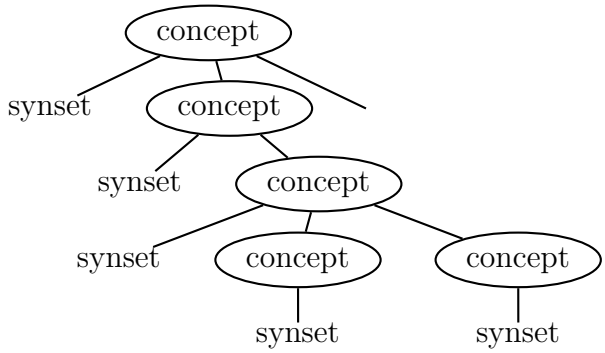
Similar to RT, HanziNet presumes a tree-like conceptual hierarchy as the organizing structure, in which each leaf node stands for a concept cluster (called conset), while each internal node represents a concept class, and domination stands for set inclusion. It is noted that, in HanziNet, the conceptual hierarchy is a **rooted branching tree-like** structure, where the top node is different from that (those) of WordNet and RT. Traditionally, top nodes have been entities, properties and relations. However, in some cases, the number of top nodes may increase and thus differ. For example, WordNet uses 11 top nodes and does not include relations among them. In HanziNet, the top node is presumed to be the universal type (drawn as $\top$), which has no differentiae. The following figure shows a comparison of the implicit organizing structure

of Roget's Thesaurus, WordNet and HanziNet.

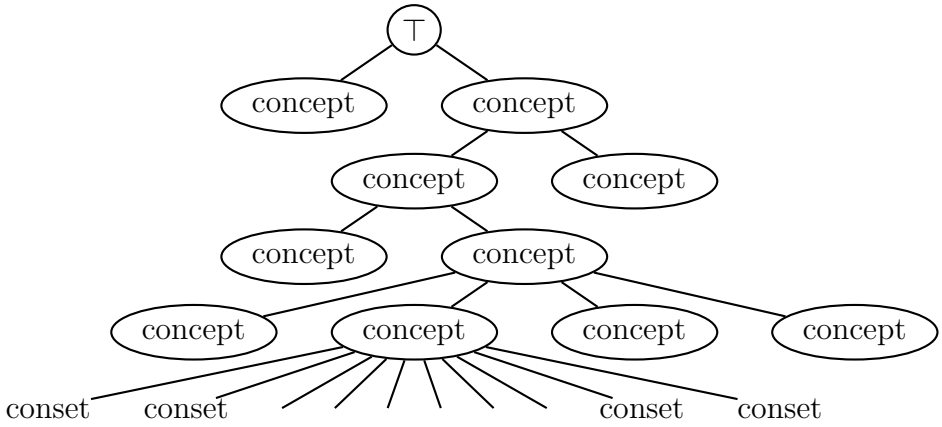
==========Roget's Thesaurus==========



==========WordNet==========



==========HanziNet==========



5.4.2 Components

In light of the previous discussion, this subsection attempts to further clarify the building blocks of the HanziNet system, with the goal to arrive at a working model which will serve as a framework for experiments in the following chapter. Breifly, HanziNet is consisted of two main parts: a character-stored machine-readable lexicon and a top-level character ontology.

Character-stored Lexicon

One of the core component of HanziNet is a *character-stored machine readable lexicon*, which provides essential information on the characters of Chinese. The current lexicon contains over 5000 characters, and 30,000 derived words in total. Since this lexicon aims at establishing an knowledge resource for modern Chinese NLP, characters and words are mostly extracted from the Academia Sinica Balanced Corpus of Modern Chinese,[32] those characters and words which have probably only appeared in classical literary works, (considered *ghost words* in the lexicography), will be discarded.

In the following, the building of the lexical specification of the entries in HanziNet will be introduced:

1. Conceptual code(s)
2. Common-sense knowledge code
3. Shallow parts of speech
4. Gloss of original meaning
5. List of combined words
6. Further aspects such as character types and cognates

(1) Conceptual Code(s)

The conceptual code is the core part of the MRD lexicon in HanziNet. Concepts in HanziNet are indicated by means of a label (conset name) with a code form. In Chapter 4, we concluded that in order to increase the efficiency, an ideal strategy is to adopt the Huffmann-coding-like method, by encoding the conceptual structure of Hanzi as a pattern of bits set within a bit string. The *coding* thus refers to the assignment of code sequences to an character. The sequence of edges from the root to any character yields the code for

[32] http://www.sinica.edu.tw/SinicaCorpus/

that character, and the number of bits varies from one character to another. Currently, for each conset (309 in total) there are 10 characters assigned on the average.

(2) Common-Sense Knowledge Code (of CSH, CSM)

As discussed in subsection 5.3.3, CSK Code denotes the common sense knowledge expressed by the **semantic components** of Hanzi. It has to be admitted that choosing the most basic semantic components is a rather subjective task. Traditionally, the set of head components is approximatly a maximum of 214 items.[33] In this study, I propose a new notion of the semantic component of characters based on the following assumptions:

- Every individual character is composed of two components (as proposed by Chu), but unlike Chu's decomposition scheme, in HanziNet, the component which is more representative in indicating *common-sense knowledge* than other component, is called the **Chararcter Semantic Head** (CSH), while the component which is less representative in indicating CSK, but often contributes to the meaning construction of the given character, is called **Character Semantic Modifier**.[34] The decision of CSH and CSM is based on their *semantic functions* in a given character, so a character component can be regarded as CSH in one character, but as CSM in another character. For example, the component "水" 氵 (water) is a very commonly used CSH and appears in 游、漂、湖、江 ..., yet functions as CSM in 溫.

- The notion of CSH is different from the traditional notion of "radicals". These two sets are not the same but overlap. For example, the CSH 虎

[33]The English theologist Joshua Marshman, 1700 years after Xu Shen died, extracted 1689 "primitive" components among 43,496 characters; Chalmers (1911) gave an account of the structure of Chinese characters under 300 primacy forms; Wieger (1940) proposed that there are 224 primitives; Wang (1983:77) even asserted that, "in any event, there cannot be more than 350 radicals". See Stalph (1989:34,46).

[34]The only exception are the so-called *pictographs* (e.g. 又), which are themselves both CSH and CSM. But the number of CSH is very limited.

is not a "radical", and the "radical" 冖 is not CSH.

The determination of CSH for every character in HanziNet is based on cross-references from many different resources: Gu et al (2003), Chu (2003), Ann (1982) and other Chinese etymology dictionaries such as Shuō Wén Jiě Zì. For every CSH, a set of CSK features are assigned. In addition, the location (TOPO) information of CSH and CSM within a character is also given (see Figure 2.1).

(3) Shallow parts of speech (mainly Nominal(N) and Verbal(V) tags)
After the heated debate in the fifties and the subsequent inconclusive discussions, there is still no prevalent consensus about the criteria for assigning syntactic information to a given character in Chinese. HanziNet provides only shallow syntactic information associated with an entry. The reason was discussed in 5.3.2.

(4) Gloss of original meaning
For every character in HanziNet, a short gloss of the *original meaning* of character is given.

(5) Cognates
The term "cognate" here is defined as characters that share the same CSH or CSM.

(6) Character types
According to ancient study, characters can be compartmentalized into six groups based on the six classical principles of character construction. Character type here means which group the character belongs to.

- Constraint-based Formalism

$$\begin{bmatrix} \textsf{PHON} & \begin{bmatrix} \textsf{SYL} & \begin{bmatrix} \textsf{SEG} & \left\langle \begin{bmatrix} \textsf{ONS}\ ch \end{bmatrix} \begin{bmatrix} \textsf{RIME}\ an \end{bmatrix} \right\rangle \\ \textsf{TONE}\ 2 \end{bmatrix} \end{bmatrix}_1 \\ \textsf{SYNSEM} & \begin{bmatrix} \textsf{CAT} & noun \\ \textsf{SEM} & cicada\ _2 \end{bmatrix} \\ \textsf{ORTH} & \left\{ 虫\ _2\ ,\ 單\ _1 \right\} \end{bmatrix}$$

Table 5.5: An AVM specification of character "chan" proposed by Sproat

In order to define more precisely what we mean by an orthographic object (characters) *representing* both linguistic and knowledge information, let's have a closer look at an example entry within constraint-based formalism.[35]

In Sproat's (Sproat 2000:9-12) design, orthography fits into an AVM by simply assuming another attribute ORTH, with an unordered list of objects as its values, which are indicated by the standard curly-brace notation for sets. In addition, he presents licensing using numerical coindexation, where the index of the licenser is marked with an asterisk . In what he considered partly logographic writing systems such as Chinese, he proposes that part of a complex glyph may be licensed by a portion of the SYNSEM part of the representation. Table 5.5 shows an AVM for the character「蟬」(*chán*, cicada), where the INSECT component 虫 (the left-hand portion of the character) is the so-called semantic radical, and the right-hand component, 單 *chán*, cues the pronunciation. In this AVM, the INSECT portion is licensed by the SEM entry, and the phonological portion is licensed by the syllable.

Modeled on Sproat's specification for the representation of character infor-

[35]This formalism became popular in the last decade in the area of grammar formalism, and is probably the most common form of knowledge representation used in NLP systems. It uses the kind of data structure termed feature structure for modelling linguistic entities such as words, phrases and sentences, and allows for structure sharing via co-references and a uniform representation of different levels of linguistic knowledge. Since feature structures are technically difficult to display, linguists usually opt for a kind of feature structure called an AVM (Attribute Value Matrix) instead.

休

CONSET	10111F	
CSH	亻 [CSK H_1H_4; TOPO 000]	1
CSM	木 [CSK $D_5J_6L_2$; TOPO 001]	2
SYNSEM	[SYN V; SEM 1 + 2]	
COGNATE	⟨CSH ⟨僅, 仁, 任 ...⟩; CSM ⟨沐, 炓, 蚞 ...⟩⟩	
GlOSS	人依傍於樹木, 停止歇息也。(rest)	
CHARACTER TYPES	形聲字 (picto-phonetic principle)	

Table 5.6: An example entry for the character "休" (/xīu/, rest).

mation, Table 5.6 shows an example from HanziNet.

Top-level Ontology

It has been widely recognised that effective lexicons cannot contain only flat lists of words. They must contain conceptual hierarchies that facilitate induction and generalization. Most current large-scale lexical resources use some basic language-independent basic *Top Concepts* to reflect fundamental semantic distinctions, and a set of hierarchically related Top Concepts called *Top-level Ontology* (or Upper ontology). For example, in WordNet, the hypernym / hypomy relationships among its noun synsets can be used as an *ontology*; EuroWordNet 1.2 is also enriched with the *Top Ontology* and the set of *Base Concepts* (Vossen 1998).

Hence for HanziNet, we have also developed a top-level ontology. We took

a bootstrapping approach by adopting the CBF Ontology as backbone (to the level 4), and refer to other ontologies (RT, SUMO, Cyc etc)[36] It contains 309 basic Top concept types nodes with labels described in Chinese. Every basic Top concept type is defined as a set of consets with the same conceptual code in binary form. They are related mainly through an IS-A relation. All Chinese characters, including the *dissyllabic morphemes* and *borrowed polysyllabic morphemes* discussed in Section 2.2.2, are assigned to *at least* one of these basic consets as well.

- A Set of Binary Relations

In designing a lexical database, there are some widely used *paradigmatic sense relations* that hold between lexical units, such as *hyponymy*, *synonymy* and *antonymy*.[37]

Hyponymy, also called *inclusion*, defines a relationship between a more and a less general term, where the meaning of the more general term is totally included within the meaning of the less general term. Complex hierarchies can be established on the basis of hyponymy relations. In contrast to hyponymy, the relationship of meaning inclusion is called *synonymy*, and defines a relationship of sameness (or betterr: similarity) of meaning. The third sense relation, *antonymy*, is normally defined as "oppositeness of meaning", which is often referred to as the opposite of synonymy.

It is now recognized that theoretically, and for practical applications as well, that no sense relation can be said to be totally without significance. For instance, in WordNet, every synset is connected to other synsets via a number of semantic relations, and these relations vary based on the type of word. WordNet has 15 semantic relations, the most important of which is synonymy.

[36] A more detailed discussion of Hanzi-derived "ontology" will be given in the next section. The complete HanziNet ontology is listed in appexdix.

[37] Another well-known disputed sense relation is "meronymy", which establishes so-called part-whole relationships.

Virtually any number and type of conceptual relation can be singled out and declared as required for any particular need. Accordingly, any complex relation may be introduced, for example, *cause, property, quality, states*, etc. But, as will be explained in the next section, enlarging the number of relations may enrich the ontology but it also makes it difficult to maintain consistency. So, in selecting relationships for HanziNet ontology, I use "**simple monotonic inheritance**", which means that each node inherits properties only from a single ancestor, and the inherited value cannot be overwritten at any point of the ontology. To put it simply, the basic conceptual relations allowed in the ontology is mainly the "**IS-A**" relation. The decision to keep the relations to one single parent was made in order to guarantee that the structure would be able to grow indefinitely and still be manageable, i.e. that the transitive quality of the relations between the nodes would not degenerate with size. Moreover, though the meaning of a *word* is reflected in its contextual relations,[38] it is noted that we are not dealing with *the meaning of a word*, but *the concept information of a character*. Therefore, in HanziNet, we restrict the links to other characters to simple relations. Explicitly, HanziNet includes only the taxonomical relation (hyponymy-hypernymy) between characters, "synonymy", one relation between consets, and the hypernym relation. That is,

- "synonyms": conset
- hypernyms: Y is a hypernym of X if every X is a (kind of) Y.
- hyponyms: Y is a hyponym of X if every Y is a (kind of) X.
- coordinate terms: Y is a coordinate term of X if X and Y share a common hypernym.

[38]Dong, the creator of HowNet, has put it metaphorically: 'relation is the soul of meaning'.

To sum up, Figure 5.8 schematizes the whole architecture of proposed HanziNet.

5.5 Issues in Hanzi Ontology Development

Ontologies, which are now commonly accepted as an *explicit specification of a conceptualization* (Gruber 1995), have been recognized as important components of linguistic and knowledge processing systems during the last few years. As proposed, Hanzi ontology constitutes a core part in the design of HanziNet. This section discusses some issues faced in designing Hanzi ontology.

5.5.1 What is an Ontology : A General Introduction from Different Perspectives

Over the years, ontologies have been used in the modelling of problems and domains. Sowa (2000:292) defines the word "ontology" as follows:

> "The subject of *ontology* is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the types of things that are assumed to exist in a domain of interest $\mathcal{D}$ from the perspective of a person who uses a language $\mathcal{L}$ for the purpose of talking about $\mathcal{D}$ ".

In the usage of information processing, we may take ontology to be a shared understanding of terminology within some domain of interest, based on *conceptualization* of domain entities and relations between them. And what is then conceptualization? Genesereth and Nilsson (1987) define conceptualization as:

> "A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to

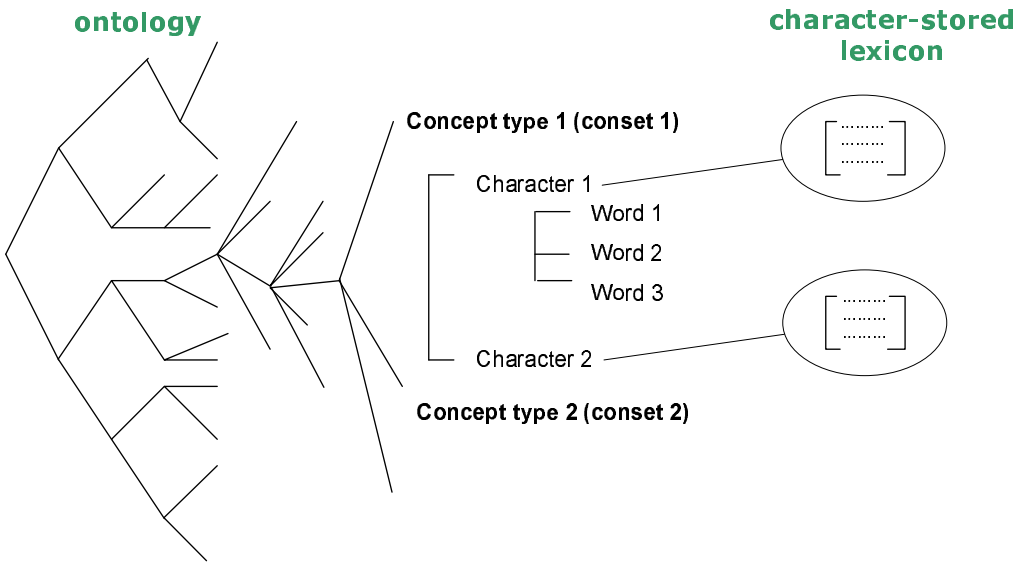


Figure 5.8: The complete architecture of HanziNet

exist in some area of interest and the relationships that hold among them".

Gruber (1993) extends this definition and gives a pragmatic definition of an ontology:

> "A conceptualization is an abstract, simplified view of the world that *we wish to represent* for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. Ontology is an explicit specification of a conceptualization."

This quotation points out the widely accepted definition of "ontology" in the NLP and AI-relevant field. In fact, studies of ontologies have been made from a highly interdisciplinary perspective. There are different traditions concerning the studies of ontologies.[39]

Historically, the term "ontology" originates from Greek philosophy, and was used to refer to a systematic account of existence. Ontology in the Aristotelian philosophical sense, is referred to as a particular system of categories accounting for a certain vision of the world, such as a classification of species on the basis of their *genus* and *differentiae*. The *genus* is the category to which something belongs, and the *differentiae* are the properties that uniquely distinguish the category members from their parent and from one another. Such methods of definition can derive a *taxonomic* hierarchy of classes.[40]

In the AI tradition, ontology has been understood as a way of differentiating between knowledge representation schemes and the content of knowledge

[39]The following discussion follows the line of Vossen (2003).

[40]In contrast to traditional taxonomy, recent formal concept hierarchies have become more complex. The main difference lies in the fact that while they both describe a structure of knowledge using concepts and relations, taxonomies represent only one perspective, namely, the subsumption relation is *transitive*. In the design of modern ontology, a tangled network model (e.g. lattice) instead of tree structure, has been proposed.

representation. In general, we might say that *knowledge representation* addresses *how* something is described, while an *ontology* is used to express *what* exactly is being described. Knowledge is defined with a focus on functionality and reasoning of machines (comparable to the discipline of epistemology in philosophy, which refers to the study of nature and sources of knowledge), while ontology represents the modelling view of knowledge representation.

In the field of NLP, there is also a tendency towards *ontology-based* surveys. Recently, a number of works in the area of language engineering have aimed at developing systems of basic semantic categories(often called *upper-level ontology* or *TOP ontology*),[41] to be used as main organizational backbones, suitable for imposing a structure on large lexical repositories.

In *lexical semantics*, ontologies have been adopted as the categorization of words in a lexicon.[42] Vossen (2003) distinguishes between two main approaches in the linguistic tradition:

- Semantic features or meaning components: In this feature approach, words are associated with a limited number of semantic features, which are then used to describe certain linguistic structures or behavior (e.g. CORELEX, DELIS, MikroKosmos, etc.)

- Lexical semantic network: In this approach, words meanings are defined in terms of relations to each other, without any reference to our cogni-

[41] According to Guarino (1998a), it is more convenient to agree on a single *TOP ontology* instead of relying on agreements based on the intersection of different ontologies. The same considerations suggest the opportunity to develop different kinds of ontologies according to their generality, and to define new ontologies in terms of existing, higher-level ontologies. In addition to the *TOP-ontology*, which describes generic concepts that are independent of a particular language, he also mentioned the type of *domain or task-specific ontologies*, which describes concepts related to a generic domain or task, by means of terms specified in the *TOP-ontology*; and the type of *application ontologies*, which describes concepts depending both on particular domain and task, which are often specializations of the corresponding ontologies.

[42] Vossen (2003) made a distinction between lexicon and ontology: Whenever we store information to make common-sense-like inferences, we tend to speak of an ontology or knowledge base. Whenever the stored information is more of a linguistic nature, such as part of speech, we tend to speak of a lexicon as part of a linguistic theory.

tive understanding and reasoning (e.g. WordNet, EuroWordNet, etc).

In the linguistic tradition, ontologies are regarded as artificial constructs built with the primary purpose of serving as lexical databases for knowledge representation systems.

To sum up, in the context of language and information processing we are concerned with here, ontology can be described as an inventory of concepts, their definitions, as well as a specification of the relationships between them, e.g. inheritance. But it has also become apparent that, there is no clear-cut definition for what an Ontology is.[43] As Fernández et al.(1997) stated that, ontological engineering is a craft, rather than a science. In the following, we will discuss some problems with respect to the design of an ontology, and how a HanziNet Ontology can circumvent them.

5.5.2 Designing a Hanzi-grounded Ontology

ISA overloading

Though there is no common, *a priori* agreement on how to build an ontology, in principle, all ontologies are centered on a classification scheme, which is based on a partial ordering relation named in various ways, such as the IS-A relation (ISA), subsumption, or hyperonymy / hyponymy. Such a taxonomy is the main *backbone* of the ontology, which can be "fleshed" with the addition of attributes and other relations among nodes (like meronymy or antonymy). In the design of modern ontologies, they thus provide a richer representation of concepts and relations that allow *multiple inheritance* and *multiple classifications*.

As is usual, we shall generically call ISA the main taxonomic relation. The problem with ISA when considering linguistic ontologies like WordNet

[43]For interested readers, Sowa (1999) gives a complete overview of Ontologies in the philosophical tradition; and Guarino (1998) gives a full overview of the role of ontologies in information systems, and a proposal for *ontology-driven* information systems.

is that: it is intended as a lexical relation between words, which not always reflects an ontological relation between classes of entities of the world.

Although this fact is well known, the tendency to confuse the two aspects (*conceptual and linguistic*) is quite common, especially when linguistic ontologies are used for non-linguistic applications. For example, in linguistic ontologies, it is quite common to rely on multiple inheritance to represent regular polysemy. This results in an *overloading* of the role of ISA links, which may cause serious semantic problems (Guarino 1998). Guarino reports five kinds of examples of what he considers ISA overloading and makes proposals to eliminate these problems.

To avoid the ISA overloading, we propose that the conceptual taxonomy, which reflects the basic top-level ontological structure, should reveal a clear and bright semantics, while the extra information can be represented by means of specialized links and attributes at the word sense level. Following the basic line of *OntoClear* methodology (Guarino and Welty (2002)), we use *simple monotonic inheritance*, which means that each node inherits properties only from a single ancestor, and the inherited value cannot be overwritten at any point of the ontology. The decision to keep the relations to one single parent was made in order to guarantee that the structure would be able to grow indefinitely and still be manageable, i.e. that the transitive quality of the relations between the nodes would not degenerate with size. Figure 5.9 shows a snapshot of the character ontology.

Tangled models: two levels in one

Vossen (1997;2003) also pointed out that many systems do not make a distinction between the *conceptual level* and the *lexical level*, or have direct mapping between the lexicon and the ontology.

He claimed that, if the system's purpose is to manage the substitution of words in text (for example, information retrieval or language generation), then we need a *linguistic ontology* (such as the EWN Ontology) which pre-

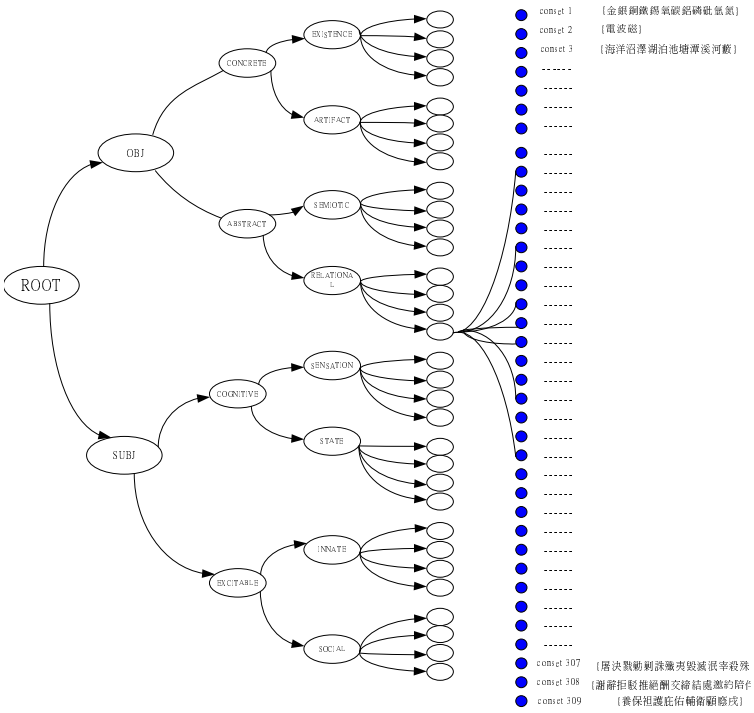


Figure 5.9: The HanziNet ontology: A snapshot

cisely reflects the lexicalization and the relations between the words in a language, and thus predicts how the same content can be paraphrased differently in a language. The lexicalization in a language has to be the starting point for ontologies for paraphrasing content. Such a design idea is, in fact, a "wordnet" in the true sense of the word and therefore captures valuable information about conceptualizations that are *lexicalized* in a language.[44]

If, on the other hand, the purpose of the ontology, more like an ontology in AI (like Cyc Ontology), is to manage semantic properties used for inference only, then it may be the case that a particular level or structuring is required to achieve better control or performance, or more compact and coherent structures. In addition, many words in a language may not be relevant for

[44]That's why in EuroWordNet, we get a much flatter hierarchy in which particular properties cannot be derived from the hyponymy relations. A detailed analysis can be found in Vossen (1998).

storing the relevant inferences, and many concepts may be needed that are not lexicalized at all. For this purpose, it may be necessary to introduce artificial levels for concepts which are *not* lexicalized in a language (e.g. natural objects), or it may be necessary to neglect levels that are lexicalized but not relevant for the purpose of the ontology (Vossen 1998).

Our concern here is, could it be possible that something exists between the *conceptual* and *lexical* level? Do Chinese characters provide an alternative answer to this question? Let us first look at an experiment.

Wong and Pala (2001) compared a selected collection of Chinese radicals, which they called a *natural semantic situation* in a natural language, with the corresponding top concepts listed, mainly, in the first and second order entities developed in the EuroWordNet (EWN). The results show that, though Chinese radicals do not form a well-defined hierarchical system as the EWN TOP Ontology does, many of the important counterparts of TOP Ontology entities can be appropriately found among Chinese radicals. Based on this finding, the authors concluded that, by using them, a construction similar to EWN TO could then be created. In their view, the result implies that we do not need to be afraid so much of the arbitrariness which is an inevitable property of any ontology of EWN TOP type.

Surely, one might asl that since the similar meaning elements can be found in any culturally developed natural language, why Chinese radicals should be considered as being so special? I regard the answer that Wong and Pala provided as a very convincing one: Their exclusiveness consists of the fact that they represent a natural language collection of basic meaning elements which do not exist *in such a compact form* in any of the known natural languages – there are only 214 radicals.

Based on the three-layer Pyramid Model proposed previously, a Hanzi-grounded ontology (a non-lexicalized but "characterized" ontology) employed in HanziNet seems to be a promising solution to this problem in a natural way. That is, Hanzi could be used as an *ontological mediator* which facilitates

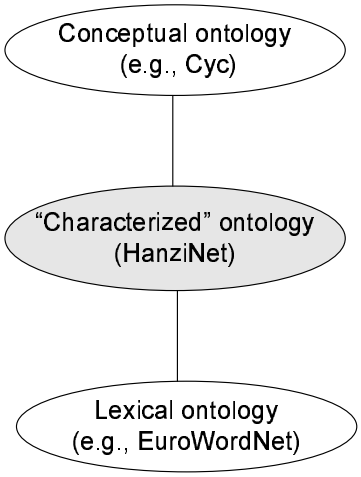


Figure 5.10: A proposed "characterized" Ontology design

"communication" between *concepts* and *words*. Figure 5.10 schematizes the idea from the view of ontology design.

I argue that such a combination of "Hanzi-grounded" ontology and HanziNet will give each character a rigorous conceptual location. With the integration of other lexical resources,[45] the relationship between conceptual classification and its linguistic instantiation would be clarified.

A Bootstrapping Strategy

Another important issues in designing HanziNet ontology involves the *ontology mapping*. In the field of NLP, more and more ontologies are being made publicly available. Unfortunately, since it is not easy for everyone to agree on one ontology, a problem appears when different ontologies are used for the same domain. At this point, ontology mapping comes into question.

The HanziNet Ontology proposed here relies heavily on the backbone of CBF Ontology [46]. This is mainly because in the methodological consideration of knowledge engineering, it is almost always worth considering what some-

[45]Such as the on-going project Chinese WordNet, being developed in Academia Sinica, Taiwan.

[46]Please refer to section 5.2.4

one else has done and checking if we can refine and extend existing sources for our particular domain and task. CBF ontology has the advantage of compactness (with only 256 concept types) over other ontologies,[47] but its validity needs more experimental testings. Since HanziNet is intended to develop a flexible framework which future studies in Hanzi ontology could draw upon and reassess, we take a *bootstrapping* strategy in the Hanzi ontology development. That is, we start with some modifications of CBF ontology, and then fine-tune with the concept types found in other resources such as Roget's Thesaurus.

In the interim, the mapping between HanziNet Ontology and SUMO Ontology [48]is in progress in parallel. The reason for choosing SUMO is twofold: (1). Currently, SUMO is linked with many other main ontologies (e.g. WordNet synsets, OpenCyc Upper Ontology, etc). (2). The on-going Chinese WordNet project (e.g. Sinica BOW).[49] is heavily based on SUMO.

Figure 5.11 shows a snapshot of the HanziNet Ontology environment.

At this juncture, HanziNet Ontology is still in its initial stages of crystallization. In terms of axiomization, it is still far away from being a formal ontology. In time, it is hoped to become a useful knowledge resource with wide ranging applications. In the next chapter, I will present a case study which takes a HanziNet-based approach to perform the NLP tasks.

[47]For example, a total of 109,377 synsets are defined in WordNet; a total of 16,788 word concepts can be found in HowNet. In the Chinese Concept Dictionary, the goal includes at least 60,000 concepts.

[48]SUMO (Suggested Upper Merged Ontology) is a shared upper ontology developed and sanctioned by the IEEE Standard Upper Ontology Working Group. It is a theory in first-order logic that consists of approximately one thousand concepts and 4000 axioms. See http://www.ontologyportal.org.

[49]`http://bow.sinica.edu.tw`

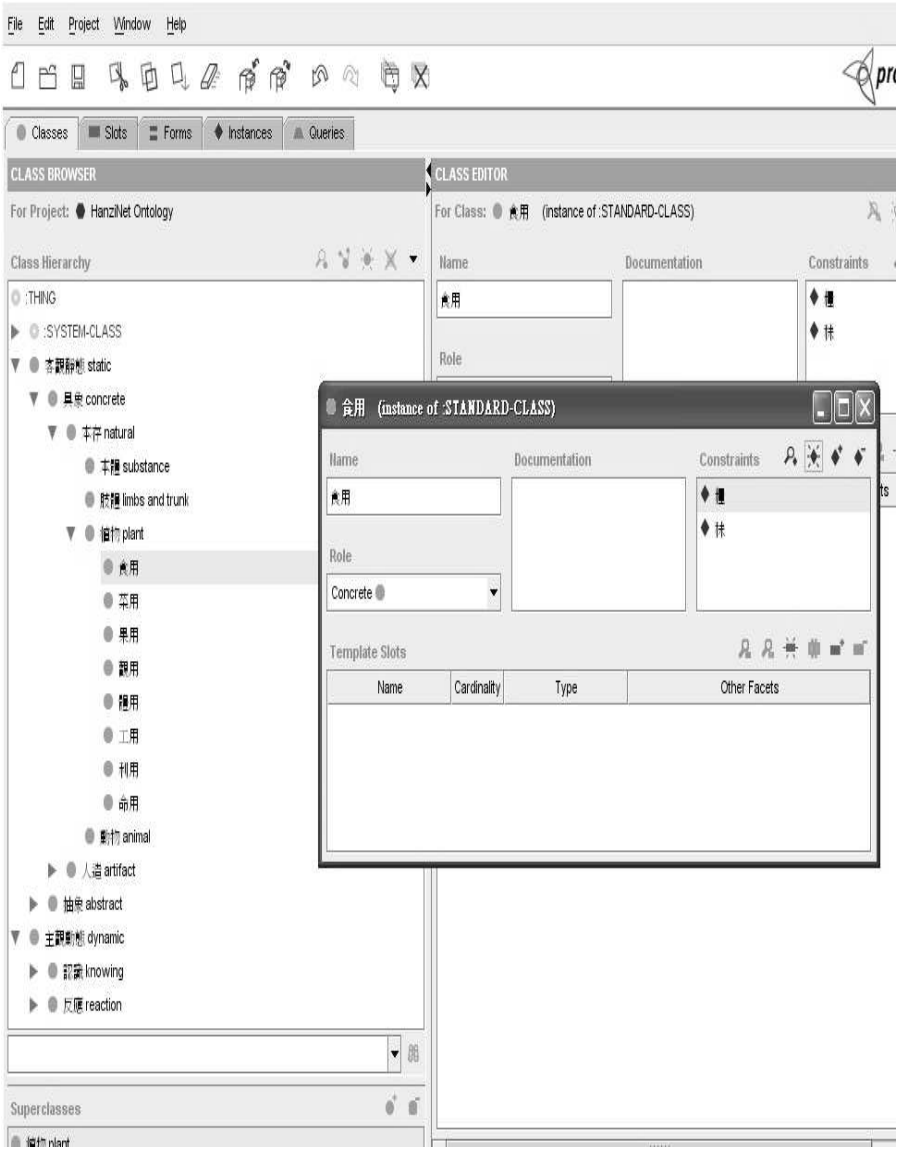


Figure 5.11: A snapshot of the HanziNet Ontology environment

Part IV

Case Study

Chapter 6

Semantic Prediction of Chinese Two-Character Words

In previous Chapter, I have described an on-going work called HanziNet, which is set to be a knowledge resource based on Chinese characters. In this Chapter, we turn our attention to a case study of the feasibility of using Hanzi-encoded conceptual knowledge in certain NLP tasks, which forms the main motive of this work. Drawing lessons from the previous studies, this knowledge resource could be used to conduct experiments in relation to meaning processing tasks, such as the *prediction* (or classification) and *disambiguation* of (unknown) word senses. This chapter will focus on the task of semantic class prediction of (unknown) words.

The chapter starts with a quick overview of some background of Chinese morphology, then an account of the morpho-semantic analysis of Chinese words will be given. Based on these discussions, I propose a simple scheme of sense prediction of unknown words. The experiment yields satisfactory results which turn out to be that the task of semantic class prediction of Chinese words could be greatly facilitated using Chinese characters as a knowledge resource.

6.1 Introduction

This chapter describes the theoretical consideration concerning with the interaction of character (morphemic component) semantics morphology, and an NLP experiment is performed to do semantic class prediction of unknown two-character words based on HanziNet - a character-grounded ontological and lexical knowledge resource of Chinese. The task that the semantic predictor performs is to automatically assign the (predefined) semantic thesaurus classes to the unknown two-character words of Chinese.

Before proceeding, it would be helpful to take a look at some background knowledge of Chinese morphology. In the theoretical setting of (Western) morphology, a single word can be viewed as made up of one or more basic units, called *morphemes*. Generally, *morphemes* are classified into two kinds - *free morphemes* are able to act as words in isolation (e.g., `cry, clear`); and *bound morphemes* can operate only as parts of other words (e.g., `ing`). In English, the latter usually take the form of *affixes*. Affixes are seen as bound morphemes that are productive in forming words. They are "grammatical" (i.e., functional) rather than "lexical" (i.e., content) in nature, and are usually classified as either inflectional (if they mark grammatical relations or agreement and do not change the form class of a word) or derivational (if they derive words with a new form class) (Packard 2000).

Chinese has often been described as a "morphologically impoverished" language in the sense that there are no *true* inflectional affixes as defined above. That is, Chinese has no word components that vary as members of a paradigm and mark grammatical values such as case, number, tense, and gender. This often leads to some general remarks (Trost 2003) asserting that *isolating language* such as Mandarin Chinese has no bound forms, e.g. no affixes, and the only morphological operation is composition.

Although there are no unequivocal opinions concerning this issue, many researchers (Li and Thompson (1981); Packard (2000); Sproat and Shih

(2001)) have shown that there exist various kinds of morphological phenomenon in Chinese, including *reduplication,*[1] *affixation, compounding and acronyms*. As Packard (1997;2000) carefully shows that, "... Chinese has a different morphological system because it selects different 'settings' on parameters shared by all language." Unlike "typical" affixing languages, Chinese has a large class of *morphemes* - which he calls "bound roots" - that possess certain *affixal* properties (namely, they are bound and productive in forming words), but encode **lexical** rather than grammatical information. These may occur as either the left- or right-hand component of a word.[2] For example, the *morpheme* 力 (/lì/; "strength, power") can be used as either the first *morpheme* (e.g., 力量 (/lì-liàng/;power-capacity "physical strength"), or the second *morpheme* (e.g., 權力 /quán-lì/; authority-strength "power") of a dissyllabic word, but cannot occur in isolation. According to Packard (2000), the class of word formation with these characteristics, in, e.g., English, is virtually nonexistent.[3]

The existence of "bound roots" has made it vague in defining Chinese *compounding*, which comprises the largest category of morphological phenomena in Chinese. Generally, the term "compound" in linguistics is restricted to words formed from two free lexical morphemes, e.g., *blackbird* (Lipka 2002). But in Chinese, there is a large number of what the Chinese linguists call *root compounding* (Sproat and Shih 1996) or *bound root words* (Packard 2000). Such *compounds* are composed of a free morpheme and

[1] A monosyllabic or dissyllabic Chinese word can reduplicate in various forms, e.g., AA form, ABAB form or AABB form, etc

[2] Chao (1968:145) calls them "root words", and Dai (1992:40) calls them "bound stems". Examples of bound roots from English might be the -ceive (receive, conceive).

[3] Similar examples would be the so-called 'latinate' stems in English(*anti-, -itis, -osis, etc*) that are also bound and productive but lexical rather than grammatical. But unlike these examples, "bound roots" in Chinese may, and in fact usually do, form words by combining with other bound roots. In addition, Chinese "bound roots" are less positionally restricted, i.e., they may in general occur as either the first or second constituent of a word, whereas in English, a given bound root generally is restricted to occurring as either a left- or right-hand word constituent, but not both.

a "bound root" morpheme which we mentioned above. An example from Sproat and Shih (2001) is the word for "白蟻"(/báiyǐ/), literally "white ant", where we underline the portion meaning "ant". The normal Mandarin Chinese word for"ant" is 螞蟻 (/máyǐ/), that is, 蟻(/yǐ/) cannot be used as a free word. It is bound and productive in forming words, but encode lexical instead of grammatical information (that's why it is not an affix).

The difficulty in deciding what constitutes a compound is exacerbated by the fact that, the boundary between *bound* and *free morphemes* is fuzzy. This is due to the fact that instances of *bound morpheme* often occur as *free morphemes* in classical Chinese texts, in some proverb-type expressions and in modern Chinese texts mixed with classical style and register. For example, the "normal" Mandarin word for "mushroom" is 蘑菇 (/múo-gū/), and 菇 (/gū/) alone cannot be used as a separate word, it is thus regarded as a *bound morpheme* in Sproat (2001), but in the following sentence, it is used as a *free morpheme*.

(6a). 這是什麼樣的菇啊
What kinds of "mushroom" (in general) is it?

Together with the discussion in Chapter 2, we have seen that in Chinese morphology, the notions of *word*, *morpheme* and *compounding* are not exactly in accord with the definition common in western linguistics. As Chao (1968) put it, the term "compound" as used by Sinologists represents a rather broader concept. Practically any word written with two or more characters is a *compound* in this sense. To avoid unnecessary misunderstanding, the pre-theoretical term *two-character* words will be mostly used instead of *compound* words in the following work.[4] This contains then all the four word

[4]Li and Thompson (1981) take a very similar view of compounding by saying that there is "a great deal of disagreement over the definition of compound. The reason is that, no matter what criteria one picks, there is no clear demarcation between compounds and non-compounds. ... we may consider as compounds all polysyllabic units that have certain properties of single words and that can be analyzed into two or more meaningful

Table 6.1: Chinese word types (adopted from Packard (2000:81))

combine what?	= word types	examples
two root words	compound word	冰山(/bīng-shān/, 'ice-mountain', iceberg),馬路(/mǎ-lù/, 'horse-road', street)
root word plus bound root, or two bound roots	bound root word	電腦(/diàn-nǎu/, 'electric-brain', computer), 橡皮(/xiàng-pí/, 'rubber-skin', rubber)
bound root or root word plus word-forming affix	derived word	房子(/fáng-zi/, 'house-AFFIX', house), 插頭(/chātoú/, 'insert-AFFIX', plug)
word plus grammatical affix	grammatical word	走了(/zǒu-le/, 'go-ASPECT', went), 我們(/wǒ-men/, 'me-PLURAL', us)

types of Chinese proposed by Packard (2000). Table 6.1 shows these types and examples.

6.2 Word Meaning Inducing via Character Meaning

Having briefly presented the background knowledge of Chinese morphology, we now move on to the morpho-semantic analysis, with the main concern in this Chapter, i.e., to what degree can/can not the meaning of words be *induced* by the meanings of their individual morphemic components (i.e., characters).

elements, or morphemes, even if these morphemes cannot occur independently [i.e. as words] in modern Mandarin." (Li and Thompson 1981: 45-46)

6.2.1 Morpho-Semantic Description

Even though the interaction between syntax and lexical semantics has been a fertile ground for research in both theoretical and computational linguistics, there has been limited work on the interaction on semantics and morphology (Hong et al 2004).

From the view of inflectional languages like Indo-European languages, semantic interpretation of derivation as well as compounding might be difficult. In the field of NLP, Goldsmith (2001) and Klimova and Pala (2000) also shows that, with the presence of allomorphs and irregular morphology in words, to retrieve the composite meaning of a word by analyzing its morpheme structure is not an easy task.

In Chinese, "**bound roots**" are the largest class of morpheme type, and as introduced, they are very productive and represent lexical rather than grammatical information. This morphological phenomena leads many Chinese linguists[5] to view the morphemic components (i.e., characters) as building blocks in the *semantic composition* process of di- or trisyllabic words. In many empirical studies (Tseng and Chen (2002); Tseng (2003); Lua (1993); Chen (2004)), this view has confirmed repeatedly.

In the semantic studies of Chinese word formation, many descriptive and cognitive semantic approaches have been proposed, such as argument structure analysis (Chang 1998) and the frame-based semantic analysis (Chu-Iang 2004). However, among these qualitative explanation theoretical models, problems often appear in the *lack of predictability* on the one end of spectrum, or *overgeneration* on the other.[6] Empirical data have also shown that in many cases, – e.g., the abundance of phrasal lexical units in any natural language, – the principle of *compositionality* in a strict sense, that is, "the

[5]For a detailed updated review in this field, please refer to Chu (2004).

[6]For example, in applying Lieber's (1992) analysis of argument structure and theta-grid in Chinese V-V compounds, Chang (1998) found some examples which may satisfy the semantic and syntactic constraints, but they may not be acceptable to native speakers.

meaning of a complex expression can be *fully* derivable from the meanings of its component parts, and from the schemas which sanction their combination", which is taken to be a fundamental proposition in some of *morpho-semantically motivated* analysis, is highly questionable.

In the field of quantitative approach, to my knowledge, Lua's relevant studies (1993;1995;2002) might be the first and most comprehensive one in researching the semantic construct and transformation of Chinese words that constructed from individual characters.

Based on data derived from a Chinese Thesaurus entitled 同義詞詞林 (CILIN, henceforth) (Mei et al. 1998),[7] Lua (1993a, 1993b) observed that the vast majority of Chinese *compounds* are constructed using 16 types of *semantic transformation patterns*.

However, as the author admits, one weak point in this approach is that it is unable to separate *conceptual* and *semantic* levels of a character or a word. His experimental result indicates that, due to the homographic effect of loading multiple meanings into a single character, characters which are conceptually correlated are not necessarily semantically correlated, for example, 父/fù/ and 子/zǐ/. Characters which are conceptually not correlated are semantically correlated, such as 單/dān/ and 白/bái/. In addition, the measure of conceptual relatedness does not tell us any more about the constraints of conceptual combination.

This has given to the consideration of the embeddedness of linguistic meanings within broader *conceptual structures* (Taylor 2002). In the later experiment, I will argue that an *ontology-based* approach would provide an interesting and efficient prospective toward the *character-triggered* morpho-semantic analysis of Chinese words. In what follows, we will argue that an

[7]CILIN classifies Chinese words using three-level semantic tree structures with 12 major, 95 medium and 1428 minor semantic classes. With a total number of about 70,000 Chinese words, it is one of the most comprehensive semantic resources in Chinese NLP so far. The "semantic class" mentioned in Lua's studies is predefined in CILIN. A section of semantic classification tree of CILIN is listed in Appendix.

ontology-based approach would provide an interesting and efficient prospective toward the *character-triggered* morpho-semantic analysis of Chinese words.

6.2.2 Conceptual Aggregate in *Compounding*: A Shift Toward Character Ontology

In prior studies, it is widely presumed that the category (be it syntactical or semantic) of a word, is somehow strongly associated with that of its composing characters. The *semantic compositionality* underlying two-character words appears in different terms in the literature.[8]

Word semantic similarity calculation techniques have been commonly used to retrieve the similar compositional patterns based on semantic taxonomic thesaurus. However, one weak point in these studies is that they are unable to separate *conceptual* and *semantic* levels. Problem raises when words in question are *conceptually* correlated are not necessarily *semantically* correlated, viz, they might or might not be physically close in the CILIN thesaurus (Mei et al 1998). On closer observations, we found that most synonymic words (i.e., with the same CILIN semantic class) have characters which carry similar conceptual information. This could be best illustrated by examples. Table 6.2 shows the *conceptual distribution* of the *modifiers* of an example of VV compound by presuming the second character 取(/qǔ/, get) as a *head*. The first column is the semantic class of CILIN (middle level), the second column lists the instances with lower level classification number, and the third column lists their conceptual types adopted from the HanziNet ontology. As we can see, though there are 12 resulting semantic classes for the * 取 compounds, the modifier components of these compounds involve only 4 concept types as follows:

11000 (SUBJECTIVE → EXCITABILITY → ABILITY → ORGANIC FUNCTION) 吸、攝,
11010 (SUBJECTIVE → EXCITABILITY → ABILITY → SKILLS) 摘、榨、拾、拔、提、攝、選,

[8] As mentioned, Lua (1993) called it as *semantic transformation patterns*, while in Chen (2004), the combination pattern is referred to as *compounding semantic template*.

Semantic class	VV compounds	Concept types of modifier component
Ee (virtue)	37 進取 ('move forward-get', be enterprising)	11110
Fa (act of upper limbs)	05 榨取 ('squeeze-get', extort) 08 摘取 ('cull-get', pick) 15 拾取 ('pick-get', collect)	11010
Fc (act of head)	05 聽取 ('hear-get', listen to)	11011
Gb (psychological activity)	07 記取 ('record-get', bear in mind)	11011
Ha (political activity)	06 奪取 ('deprive-get', seize)	11110
Hb (military activity)	08 襲取 ('attack-get', take over) 12 攻取 ('attack-get', capture) 12 奪取 ('deprive-get', seize) 12 襲取 ('attack-get', take over)	11110
Hc (administration)	07 收取 ('receive-get', collect) 23 拔取 ('pull-get', promote) 25 錄取 ('employ-get', enroll)	{11110;11011}
Hi (sociality)	27 領取 ('receive-get', get) 27 提取 ('lift', distill)	{11010;11110}
Hj (living)	25 選取 ('choose-get', choose) 25 摘取 ('cull-get', pick)	{11010;11110}
Hn (ferocity)	03 掠取 ('plunder-get', plunder) 10 剽取 ('rob-get', plagiarize) 12 榨取 ('squeeze-get', extort)	11110
If (circumstances)	09 考取 ('examine-get', pass an entrance examination)	11011
Je (influence)	12 爭取 ('strive-get', strive for) 12 詐取 12 吸取 12 攝取 12 竊取 12 牟取 12 謀取 12 掠取 12 獵取 12 截取 12 獲取 12 換取 12 奪取	{11000;11110;11011;11110}

Table 6.2: Conceptual aggregate patterns in two-character VV (compound) words: An example of * 取 (get)

11011 (SUBJECTIVE → EXCITABILITY → ABILITY → INTELLECT) 牟、謀、考、選、錄、記、聽,

11110 (SUBJECTIVE → EXCITABILITY → SOCIAL EXPERIENCE → DEAL WITH THINGS) 收、獲、領、換、奪、竊、詐、獵、掠、爭、剽、襲、攻、進

We defined these patterns as *conceptual aggregate pattern* in compounding. Unlike statistical measure of the co-occurrence restrictions or association strength, a *concept aggregate pattern* provides a more knowledge-rich scenario to represent a specific manner in which concepts are aggregated in the ontological background, and how they affect the compounding words. We will propose that the semantic class prediction of Chinese two-character words could be improved by making use of their *conceptual aggregate pattern* of head/modifier component.

6.3 Semantic Prediction of Unknown two-character Words

6.3.1 Background

This section describes an NLP experiment on semantic prediction (or classification) of unknown two-character words based on HanziNet. The practical task intended to be experimented here involves the automatic classification of Chinese two-character words into a predetermined number of semantic classes.

Before embarking on this, some background knowledge is introduced as follows. In section 2.2.2, we briefly introduced one of the most complex problems in computer processing of Chinese language: word segmentation. As already known, due to the lack of blanks to mark word boundaries, it is a difficult task for a computer (even for native speakers), to identify the words in an input Chinese sentence. In addition to segmentation ambiguities, occurrences of *out-of-vocabulary* words (i.e. unknown words) constitute the main difficulty. While the number of newly coined words grows daily,

it would be tremendously time-consuming to manually create a dictionary that contains all of these words previously unseen. That is the reason why *recognizer and classifier* needed to be developed to do the task automatically. Many works (Chang et al (1997); Chen and Lee (1994); Chen and Bai (1997)) have been done in the area of identification and syntactic tagging of Chinese unknown words. Until recently, more and more researches have focused on the semantic classification of Chinese unknown words.

- *Types of Unknown Words*

According to a reported statistical analysis (Chen et al.(1997)), upon examining a 3.5 million word data from a Chinese corpus, five frequently occurring types of unknown words were found. These are briefly explained as follows:

(Suōxiě) Suōxiě can be better understood as acronym in English, i.e., shortened forms of long names. E.g., 台灣大學 (/tai-wan-da-xue/, "Taiwan University") is commonly shortened as 台大 (/tai-da/) by choosing the first and the third characters to yield a new form.

(Proper names) These involve person names, place names, organization names and so on.

(Derived words) These types can be realized as *derivational suffixation*. There are very limited (and controversial !) numbers of "suffixes" in Mandarin Chinese, but they are very productive. For example, 工業化(/gōng-yè-huà/,"industrial-ize").

(Numeric type compounds) These types include dates, numbers, time, etc. E.g., 一九七零年 ("1970-year")

(Compound words) E.g., huò-yǔn ("receive-permission"; obtain permission)

Since each type has its own morphological structure and must be treated separately, there has not been one satisfactory unified solution for unknown word detection. Among these types of unknown words, Chen and Chen (2000) pointed out that *compound words* constitute the most productive type of unknown words in Chinese texts. In contrast with other types, such as numeric type compounds, which can be well explained by regular expression, the set of general compounds words is the most difficult type to predicate, for these compounds are constantly newly coined - though under certain idiosyncratic co-occurrence restrictions - by combining two or more characters from a large set of characters.

- *Shallow vs Deep Semantic Classification*

The *semantic classification* task discussed in this chapter can be regarded as *deep* semantic classification work,[9] which aims to assign more specific semantic category (or *sense tags*) from the bottom-level semantic classes in the taxonomy of a certain thesaurus. This differs from the *shallow* semantic classification in that the latter aims at assigning broader *sense tags* from the top-level semantic classes.

6.3.2 Resources

The following resources are used in the experiments:

- HanziNet

  The ontology and its Hanzi instances of HanziNet, as introduced in Chapter 5, will be used as a knowledge resource in performing the task. The dissyllabic words in the character-stored lexicon of HanziNet will also be adopted as a supplemental training data.

- TongYiCi CiLin (CILIN Thesaurus)

  Word semantic classification needs a predefined set of word senses to

[9] This distinction was proposed by Chen (2004).

disambiguate between. Although there is no infallible criteria for identifying "words", data based on lexicographical considerations might be *reliable* enough to be our experimental data. In most recent works, the set has been taken from a general-purpose thesaurus resource called TongYiCi CiLin with the assumption that the lexical resource describes properly the word senses of modern Chinese.

CILIN has been widely accepted as a semantic categorization standard of Chinese word in Chinese NLP, it is used for training and evaluating semantic classifier systems. In CILIN, words are categorized in hierarchies of which organization is similar to Roget's Thesaurus. CILIN is a collection of about 52,206 Chinese words, classified into a **3-level** hierarchy: 12 major (level-1), 95 middle (level-2) and 1428 minor (level-3) semantic classes. Among the 56,830 words in CILIN's index, 7866 (15.05 %) words belong to more than one semantic category. The categorization hierarchy of this thesaurus is described in a table in the Appendix.

- Sinica Corpus

  The Sinica Corpus (CKIP, 1995) is the only publicly available, fully tagged (with part-of-speech) *traditional* Chinese corpus of its size.[10] The Sinica Corpus aims to someday acquire five million word entries. The latest version, the Sinica 1.0 Balance Corpus, includes slightly more than two million word entries. This corpus is balanced over five of its source attributes, namely: topic; genre; medium; style; and mode (Hsu and Huang, 1995). Originally, the Sinica Corpus was in BIG-5 encoded. Word segmentation within the Sinica Corpus was done according to the standard proposed by Computational Linguistic Society of Taiwan. Tagging was done with a set of 46 POS tags. Sinica Corpus is used to

[10]The term *traditional* here means the characters currently used in Taiwan and Hong Kong rather than the simplified one used in VR China.

assign the POS tags (mainly V and N) of training and testing character data in the experiment.

6.3.3 Previous Research

The task involves here is the automatic classification of Chinese words into a predetermined number of semantic categories. Previous researches can be summarized as two models as follows.

Connectionist Model

Lua (1995, 2002) explored semantic category prediction of Chinese bi-character words using a three-layer back propagation neural network in a pioneering experiment. By inputting (1) the semantic classes, which are directly derived from CILIN, and (2) a parameter called *semantic strength*,[11] the system is reported to give good results, yielding an accuracy of 81% in predicting the semantic classes of Chinese dissyllabic words.

This theoretical proposal comes from some of his previous papers (Lua 1993a and 1993b). He asserted that each Chinese compound word is a result of *semantic aggregates*, and that it derives its meaning through certain semantic transformations from the meanings of its constituent characters. He also proposed a scheme to quantify various types of these semantic transformations.

In principle, in presuming the properties of Chinese compounds, we share a similar position with Lua, however, the existing connectionist model, though praised for its high rate of accuracy, encounters some crucial difficulties. First, it cannot deal with any "incompleteness" in characters lexicon, for this system depends heavily on CILIN, a semantic thesaurus containing only about 4,133 characters. As a result, if unknown words contain characters that are not listed in CILIN, then the prediction task cannot be performed. Sec-

[11] We have introduced this notion in Chapter 5.

ond, the ambiguity problem is shunned by pre-selection of character meaning in the training step.

The Example-Based Model

Chen and Chen (2000) propose an *example-based* learning method (also called *similarity* or *instance*-based method) to perform the task of automatic semantic classification for Chinese unknown *compound nouns*. According their inspection on the Sinica Corpus, *compound nouns* are the most frequently occurred unknown words in Chinese text. The unknown *compound nouns* extracted from the Sinica Corpus were classified according to the morphological representation by the similarity-based algorithm.

For each input, the classifier does the morphological analysis at first, that is, determine the *head* and *modifier morpheme*, and get the syntactic and semantic categories of the modifiers. The semantic category of the input unknown word will be assigned with the semantic category of the example word with the most similar morpho-semantic structures calculated by a similarity measure. The proposed semantic classifier uses a measure of semantic similarity very much like that described in Resnik (1995). For evaluation, 200 samples from the output (the total number is not given!) are picked out randomly for the performance evaluation by examining the semantic classification manually. The accuracy rate of 81% is reported.

Under the example-based paradigm, Tseng (2003) presents a semantic classifying algorithm for unknown Chinese words using the $\mathcal{K}$ *nearest-neighbor method*. This approach computes the distance between an unknown word and examples from the CILIN thesaurus, based upon the similar metric of morphological similarity of words whose semantic category is known. This yields the results of 70.84% of nouns, 47.19% of verbs and 53.50% of adjectives, respectively.

Problems

Difficulties encountered in previous researches could be summarized as follows:

First, many models (Chen and Chen 1998;2000) cannot deal with the *incompleteness* of characters in the lexicon, for these models depend heavily on CILIN, a Chinese Thesaurus containing only about 4,133 monosyllabic morphemic components (characters).[12] As a result, if unknown words contain characters that are not listed in CILIN, then the prediction task cannot be performed automatically.

Second, the ambiguity of characters is often shunned by manual preselection of character meaning in the training step, which causes great difficulty for an automatic work.

Third, it has been widely assumed (Lua (1995;1997;2002); Chen and Chen (2000)) that the overwhelming majority of Chinese words have *semantic head* in their morphological structures. That is, Chinese compound words are more or less *endocentric*, where the compounds denote a hyponym of the *head* component in the compound. So for example, 電郵 ("electric-mail"; e-mail) IS-A a kind of mail, 郵 is the *head* of the word 電郵. The process of identifying semantic class of a *compound* thus boils down to find and to determine the semantic class of its *head* morpheme.

Though the head-oriented presumption works well in NN compound words, where the head is the rightmost character by default (Chen and Chen 2000), there is also an amount of exocentric and appositional compounds[13] where no straightforward criteria can be made to determine the head component. For example, in a case of VV compound 訓斥 ("denounce-scold", drop-on), it is difficult (and subjective) to say which character is the head that can assign

[12]The reason for this lies in that, since many characters in modern Chinese texts have not been used as a *free* morphemes, they will not be included in thesaurus or dictionary for modern Chinese. This seems to demand a character-based (including free and bound morpheme) knowledge resources.

[13]Lua reports a result of 14.14% (Z3 type).

a semantic class to the compound. In addition, generally speaking, the *head* discussed here is understood in terms of semantic value, instead of syntactic or structural description.[14] But in (Chen and Chen (2000)), they assume that the syntactic and semantic categorization of a head are closely related for coarse-grained analysis. So in their model, the syntactic categories of an unknown word have to be predicted first, and the possible semantic categories will be identified according to its top-ranked syntactic tags. The problems multiplied due to the ambiguity of syntactic categories. Lua (2002)'s proposal of "semantic nuclear/modifier" using the distance counting method based on CILIN has advantages of objectivity and syntax-independent, but, again, faces the problems due to the limited coverage of characters in CILIN.

At the time of writing, Chen (2004) is the most updated example-based approach to this topic. To solve above-mentioned problems, he proposed a non head-oriented character-sense association model to retrieve the *latent* senses of characters and the *latent* synonymous compounds among characters by measuring similarity of semantic template in compounding by using a MRD. However, as the author remarked in the final discussion of classification errors, the performance of this model relies much on the productivity of compounding semantic templates of the target compounds. To correctly predict the semantic category of a compound with an unproductive semantic template is no doubt very difficult due to a sparse existence of the template-similar compounds. In addition, pure statistical measures of sense association do not tell us more about the constraints and knowledge of conceptual combination.

In the following, we will propose that a knowledge resource at the morpheme (character) level could be a straightforward remedy to the first and second problems. By treating characters as instances of conceptual prim-

[14] Packard (2000) distinguishes "semantic head" and "structural head" in Chinese words. The latter is a head defined by reference to syntactic rather than semantic value. These two kinds of head match in some cases, but are totally different in other cases.

itives, a character ontology might provide an interpretation of *conceptual grounding* of word senses. At a coarse grain, the character ontological model does have advantages in efficiently defining the *conceptual space* within which hanzi - intepreted as instances of concept primitives -, and their relations, are implicitly located.

6.3.4 A Proposed HanziNet-based Approach

With HanziNet at hand, we approach the task in a hybrid way that combines the strengths of *ontology-based* and *example-based* model to arrive at better result for this task. In general, the approach proposed here differs in some ways from previous research based on the following considerations:

- *Context-freeness*:

Roughly, the semantic classes of unknown words can be predicted by their (local) content and (global) contextual information. In carrying out the word sense prediction task, we first presume the *context-freeness hypothesis*, i.e., without resorting to any contextual information.[15] The consideration is twofold. First, we observe that native speaker seems to reconstruct their new conceptual structure *locally* in the processing of unknown compound words. Second, the *context-freeness hypothesis* has the advantage especially for those unknown words that occur only once and hence have limited context.

- *HanziNet as a Knowledge Resource*

As stated, one of the most intractable problems of automatically assigning semantic classes to Chinese unknown words lies in the *incompleteness* of morphemic components (i.e., characters) in the CILIN Thesaurus. This problem causes great difficulty, especially for example-based models, to perform this task (Chen and Chen 1998).

[15]This differs from the task of word sense disambiguation, in which context might play an important role.

HanziNet could be a remedy to this *out-of-coverage* problem. The new model presented here relies on a coarsely grained upper level character grounded ontology, which is one of the core components of HanziNet. As introduced in Section 5.4.2, this character ontology is a tree-structured conceptual taxonomy in terms of which only two kinds of relations are allowed: the INSTANCE-OF (i.e., certain characters are instances of certain concept types) and IS-A relations (i.e., certain concept type is a kind of certain concept type). For the review of the HanziNet ontology, readers can refer to Figure 5.9 and Appendix.

- *Character-triggered Latent Near-synonyms*

The rationale behind this approach is that similar conceptual primitives - in terms of characters - probably participate in similar context or have similar *meaning-inducing* functions. This can be rephrased as the following presumptions: (1). Near-synonymic words often overlap in senses, i.e., they have same or close semantic classes. (2). Words with characters which share similar conceptual information tend to form a *latent* cluster of synonyms. (2). These similar conceptual information can be formalized as *conceptual aggregate patterns* extracted from a character ontology. (3). Identifying such *conceptual aggregate patterns* might thus greatly benefit the automatically acquired near-synonyms, which give a set of good candidates in predicting the semantic class of previously unknown ones.

The proposed semantic prediction (SC) system retrieves at first a set of near-synonym candidates using *conceptual aggregation patterns*. Considerations from the view of lexicography can winnow the *overgenerated* candidates, that is, a final decision of a list of near-synonym candidates is formed on the basis of the CILIN's verdict as to what latent near-synonyms should be. Thus the semantic class of the target unknown two-character words will be assigned with the semantic class of the top-ranked near-synonym calculated by the similarity measurement between them. This method has advantage

of avoiding the snag of apparent multiplicity of semantic usages (ambiguity) of a character as well.

Take for an example. Suppose that the semantic class (Hi37) of a two-character word 保護 (/bǎo-hù/, 'protect') is unknown. By presuming the left-most character 保 as the head of the word, and the rightmost character 護 as the modifier of the word, the system first identifies the conset which 護 belongs to. Other instances in this conset are 保, 袒, 戍, 衛, 庇, 佑, 顧, 輔, 佐, 守, 養, etc. So the system retrieves a set of possible *near-synonym* candidates by replacing 護 with other character instances in the same conset, namely, NS_1: {保保, 保袒, 保戍, 保衛, 保庇, 保佑, 保顧, 保輔, 保佐, 保守, 保養}; in the same way, by presuming the rightmost character 護 as the head, and the leftmost character 保 as the modifier of the word, we have a second set of possible *near-synonym* candidates, NS_2: {護護, 袒護, 戍護, 衛護, 庇護, 佑護, 顧護, 輔護, 佐護, 守護, 養護}[16]. Aligned with CILIN, those candidates which are also listed in the CILIN are adopted as the final two list of the near-synonym candidates for the unknown word 保護: NS'_1: {袒護 (Hi41"be partial to"), 衛護 (Hb04;Hi37 "guard"), 庇護 (Hi47 "shelter"), 守護 (Hi37 "shield"), 養護 (Hd01 "bring up")}, and NS'_2: {保佑 (Hl33 "bless"), 保養 (Hj33 "maintain"), 保守 (Ee39 "conserve")}. Thus the semantic class of the target unknown two-character word 保護 will be assigned with the semantic class of the top-ranked near-synonym calculated by the similarity measurement between 保護 and its candidates of near-synonyms.

- *Semantic Similarity Measure of Unknown Word and its Near-Synonyms*

Given two sets of character-triggered near-synonyms candidates, the next step is to calculate the semantic similarity between the unknown word (UW) and these near-synonyms.

CILIN Thesaurus is a tree-structured taxonomic semantic structure of Chinese words, which can be seen as a special case of semantic network

[16]Note that in this case, 保 and 護 are happened to be in the same conset.

(Figure 4.4 in Chapter 5). To calculate semantic similarity between nodes in the network can thus make use of the structural information represented in the network.

Previous works of taxonomy-based semantic similarity measurement can be categorized as two approaches: *path length-based* approaches (Leacock-Chodorow (1998); Wu and Palmer (1994)) and *information content-based* (Resnik 1995; Lin 1998; Jiang and Conrath 1997). The *path length-based*, also called *edge-based* approaches, use an intuitive idea of evaluating semantic similarity by counting the number of nodes or relation links between nodes in a taxonomy: the lower the distance between two items, the higher their similarity. As Resnik (1995) reports, this approach has well-known problem in that it relies on the notion that links in the taxonomy represent uniform distances. In a more realistic cases, the distances between any two adjacent nodes are not necessarily equal, most latter approaches have to be determine the weight of path length by incorporating the considerations of other structural features of the network.

So we used the information content-based approach to perform the task, which uses the notion that the more information content two *semantic class* share, the more similar they are. In Resnik (1995), the semantic similarity score of two semantic classes sc_1, sc_2 in a IS-A taxonomy, equals the *information content* (IC) value of their lowermost common subsumer (LCS) (i.e., the lowest node subsuming them both). Following the notation in information theory, the *information content* (IC) value of a semantic class sc is defined as: $IC(sc) = \log^{-1} P(sc)$, where $P(sc)$ is the probability of encountering an instance of semantic class sc. So the semantic similarity of two semantic classes is formalized as:

$$\begin{aligned} sim_{Resnik}(sc_1, sc_2) &= \overset{max}{sc \in LCS(sc_1, sc_2)} [IC(sc)] \\ &= \overset{max}{sc \in LCS(sc_1, sc_2)} [\log^{-1} P(sc)], \end{aligned} \tag{6.1}$$

where $LCS(sc_1, sc_2)$ is the set of semantic classes that dominate both sc_1 and sc_2. A semantic class which achieves the maximal value among the LCS in the equation, is called the *most informative subsumer*.

One is often interested in measuring the similarity of *words*, rather than the similarity of *semantic classes*. In the case of similarity measure of words, where words may have more than one semantic class, and hence they might have more than one direct superordinate semantic classes, the similarity measure of words can be calculated by the best similarity value among all the semantic classes pairs which their various senses belong to:

$$sim_{Resnik}(w_1, w_2) = \max_{sc_1 \in sense(w_1), sc_2 \in sense(w_2)} [sim_{Resnik}(sc_1, sc_2)], \qquad (6.2)$$

where $sense(w)$ represents the set of possible senses for word w.

Following this information content-based model, in measuring the semantic similarity between unknown word and its candidate near-synonymic words, we propose a measure metric modified from those of Chen and Chen (2000), which is a simplification of the Resnik algorithm by assuming that the occurrence probability of each leaf node is equal. Given two sets (NS'_1, NS'_2) of candidate near synonyms, each with m and n near synonyms respectively, the similarity is calculated as:

$$sim_{\mu}(UW, NS'_1) = \arg^{max}_{i=1,m} \frac{IL(LCS(sc_{uwc1}, sc_i)) * f_i}{\sum_{i=1}^{m} IL(LCS(sc_{uwc1}, sc_i)) * f_i}(\beta) \qquad (6.3)$$

$$sim_{\nu}(UW, NS'_2) = \arg^{max}_{j=1,n} \frac{IL(LCS(sc_{uwc2}, sc_j)) * f_j}{\sum_{j=1}^{n} IL(LCS(sc_{uwc2}, sc_j)) * f_j}(1-\beta) \qquad (6.4)$$

where sc_{uwc1} and sc_{uwc2} are the semantic class(es) of the first and second morphemic component (i.e., character) of a given unknown word, respectively. sc_i and sc_j are the semantic classes of the first and second morphemic components on the list of candidate near-synonyms NS'_1 and NS'_2. f is the frequency of the semantic classes, and the denominator is the total value of numerator for the purpose of normalization. β and $1-\beta$ are the weights

which will be discussed later. The *Information Load* (IL) of a semantic class sc is defined as:

$$\begin{aligned} IL(sc) &= Entropy(system) - Entropy(sc) \\ &\simeq (-\frac{1}{n}\sum \log_2 \frac{1}{n}) - (-\frac{1}{m}\sum \log_2 \frac{1}{m}) \\ &= \log_2 n - \log_2 m \\ &= -\log_2(\frac{m}{n}), \end{aligned} \tag{6.5}$$

if there is n the number of the minimal semantic classes in the system,[17] m is the number of the semantic classes subordinate sc.

To take an example, consider how the semantic similarity between a given unknown word UW and one set of near-synonyms candidates $sim_\mu(UW, NS'_1)$ would be computed, using the Equation (6.3), (6.4) and CILIN taxonomic information. Suppose the unknown two-character word in question is 保護(/bǎo-hù/, protect) again, whose semantic class to be guessed is `Hi37`. The system retrieves two sets of candidate near synonyms, namely, NS'_1: {袒護 (/tǎn-hù/, `Hi41`"be partial to"), 衛護 (/wèi-hù/,`Hb04;Hi37`, "guard"), 庇護 (/pì-hù/, `Hi47`, "shelter"), 守護 (/shǒ-hù/, `Hi37`, "shield"), 養護 (/iǎng-hù/, `Hd01`, "bring up")}, and NS'_2: {保佑 (/bǎo-ioù/, `Hl33`, "bless"), 保養 (/bǎo-iǎng/, `Hj33`, "maintain"), 保守 (/bǎo-shǒ/, `Ee39`,"conserve")}. For the first part of calculation, Table 6.3 shows the first character (保)(/bǎo/) of the unknown word, and the first character of the candidate near synonyms in NS'_1 (i.e., 袒 (/tǎn/), 衛 (/wèi/), 庇 (/pì/), 守(/shǒ/), 養(/iǎng/)), together with their semantic classes. From Table 6.3 we get Table 6.4. In this case, the first character of the test unknown word has five semantic classes, the system first starts with $sc_{uwc1} = Hb04$, By Equation 6.3, we have

IL(LCS(Hb04 ∩ Ah15)) * f = IL(Root) * 1 = $-\log_2(\frac{3915}{3915})$ * 1 = 0;
IL(LCS(Hb04 ∩ Ea11)) * f = IL(Root) * 1 = $-\log_2(\frac{3915}{3915})$ * 1 = 0;
IL(LCS(Hb04 ∩ Hi39)) * f = IL(H) * 1 = $-\log_2(\frac{834}{3915})$ * 1 = 2.231;

[17] In CILIN, $n = 3915$.

1st characters	semantic classes
[gray].90 保 (/bǎo/)	Hb04; Aj14; Di02; Jd01; Ka15
袒 (/tǎn/)	Ah15; Ea11
衛 (/wèi/)	Hb04
庇 (/pì/)	(not found in CILIN)
守 (/shǒ/)	Hb04; Hi39; Id21
養 (/iǎng/)	Ed59; Hd01; Hd27; Hg07; Hi38; Hj33; Ib01; Jd01

Table 6.3: The first characters and their semantic classes

sc_i	frequency (f)
Ah15	1
Ea11	1
Ed59	1
Hb04	2
Hd01	1
Hd27	1
Hg07	1
Hi38	1
Hi39	1
Hj33	1
Ib01	1
Jd01	1

Table 6.4: The semantic classes and their distribution of the first characters

IL(LCS(Hb04 $\cap$ Hb04)) * f = IL(Hb) * 2 = $-\log_2(\frac{4}{3915})$ * 2 = 19.870 and so on. The denominator is then calculated as: $\sum_{i=1}^{k} IL(Hb04 \cap sc_i) * f = 0 + 0 + 2.231 + 19.870 + \cdots + 0 = 33.256$. The next steps, - when $sc_{uwc1} =$

`Aj14, Di02, Jd01, Ka15` and so on -, work in the same way. Finally,

$$\begin{aligned}\sum_{i=1}^{m} IL(LCS(sc_{uwc}1, sc_i)) * f_i &= \sum_{i=1}^{m} IL(LCS(Hb04, sc_i)) * f_i \\ &+ \sum_{i=1}^{m} IL(LCS(Aj14, sc_i)) * f_i \\ &+ \sum_{i=1}^{m} IL(LCS(Di02, sc_i)) * f_i \\ &+ \sum_{i=1}^{m} IL(LCS(Jd01, sc_i)) * f_i \\ &+ \sum_{i=1}^{m} IL(LCS(Ka15, sc_i)) * f_i \\ &= 33.256 + 3.654 + 0 + 10.350 + 0 = 47.26\end{aligned}$$

So, the value of ${}_{max}sim\mu(UW, NS'_1)$ equals to $\frac{19.870}{47.26} = 0.421$. For the sim_{nu}, namely, to calculate the semantic similarity between the second character 護 (/hù/) of unknown word and the second characters of candidate near synonyms in NS'_2 (佑 (/iù/), 養(/iǎng/), 守 (/shǒ/)) by Equation (6.4), the calculation process is the same. The value of ${}_{max}sim_{nu}(UW, NS'_2) = \frac{9.935}{46.59} = 0.213$. Without taking weights into consideration, we now have a ranked list of near synonyms (6.5): {衛護 (0.421)、守護 (0.421)、養護 (0.219)、...}, among which 衛護 gets the maximal value of MAX(sim_{mu}, sim_{nu}), its semantic class (`Hb04`) is then assigned to the unknown word. It is also a correct answer in this case.

- *Circumventing the strict "Headedness" Presupposition*

As remarked in Chen (2004), the previous research concerning the automatic semantic classification of Chinese compounds (Lua 1997; Chen and Chen 2000) presupposes the *endocentric* feature of compounds. That is, by supposing that compounds are composed of a *head* and a *modifier*, determining the semantic category of the *target* therefore boils down to determine the semantic category of the *head* compound.

rank	synonyms	sim.score	sem.cat	near-syn set
[gray].90 1	衛護 (衛:Hb04)	$\frac{19.870}{47.260} = 0.421$	Hb04	NS'_1
1	守護 (守:Hb04)	$\frac{19.870}{47.260} = 0.421$	Hb04	NS'_1
2	養護 (養:Jd01)	$\frac{10.350}{47.260} = 0.219$	Hd01	NS'_1
3	保養(養:Hd01)	$\frac{9.935}{46.590} = 0.213$	Hj33	NS'_2
4	保養(養:Hd27)	$\frac{5.631}{46.590} = 0.121$	Hj33	NS'_2
5	袒護 (袒:Ah15)	$\frac{3.654}{47.26} = 0.077$	Hi41	NS'_1
6	保守(守:Hb04)	$\frac{2.231}{46.59} = 0.048$	Ee39	NS'_2

Table 6.5: The final result: A ranking list

In order to circumventing the strict "headednes" presumption, which might suffer problems in some borderline cases of V-V compounds, the weight value (β and $1-\beta$) is proposed. The idea of weighting comes from the discussion of *morphological productivity* in Section 5.4.2. I presume that, within a given two-character words, the more productive, that is, the more numbers of characters a character can combine with, the more possible it is a *head*, and the more weight should be given to it. The weight is defined as $\beta = \frac{C(n,1)}{N}$, viz, the number of candidate *morphemic components* divided by the total number of N. For instance, in the above-mentioned example, NS_1 should gain more weights than NS_2, for 護 can combine with more characters (5 near-synonyms candidates) in NS_1 than 保 does in NS_2 (3 near-synonyms candidates). In this case, $\beta = \frac{5}{8} = 0.625$. It is noted that the weight assignment should be character and position independent.

6.3.5 Experimental Settings

Data

As introduced in previous chapters, dissyllabic words, that is, words consisting of two characters, are the most widely used types in Chinese. In the following, I will focus specifically on these two-character words.

The goal of this experiment, is set to implement a classifier that assigns semantic categories to Chinese unknown words. We conducted an open test

experiment, which meant that the training data was different from the testing data. In order to compare with previous studies, 200 N-N, 100 V-N and 500 V-V two-character words were chosen at random from CILIN to serve as test data, and all the words in the test data set were assumed to be unknown. The syntactic categories were assigned based on the POS tagset proposed by the Academia Sinica Corpus. Some examples are: N-N:火災 (fire-disaster, 'fire'), 茅屋 (thatch grass-cottage,'hut'), V-N:犯罪 (commit-crime, 'commit a crime'), 看病 (see-sickness, 'see a doctor'), V-V:估計 (appraise-calculate, 'estimate'), 推想 (reason-think, 'suppose'), 演講 (perform-address, 'lecture'), 報到 (report-come, 'register') and so on.[18]

Baseline

The baseline method assigns the semantic class of the randomly picked head component to the semantic class of the unknown word in question. It is noted that most of the morphemic components (characters) are ambiguous, in such cases, semantic class is chosen at random as well.

Outline of the Algorithm

Table 6.6 illustrates a step-by-step explanation of the algorithm. In summary, the strategy to predict the semantic class of a unknown two-character word is, to measure the semantic similarity of unknown words and their candidate near-synonyms which are retrieved based on the HanziNet ontology. For any unknown word UW, which is the character sequence of C_1C_2, the $RANK(sim_\mu(\beta), sim_\nu(1-\beta))$ is computed. The semantic category sc of the candidate synonym which has the value of $MAX(sim_\mu(\beta), sim_\nu(1-\beta))$, will be the top-ranked guess for the target unknown word.

[18] All the test data are listed in Appendix.

Steps	Instructions	Results
1)	For each input unknown two-character word,	訓斥 (/shiùn-cì/, VV, "drop-on",),
2)	Bisect this unknown word and look up instances in the character ontology which share the same conceptual meaning (i.e., in the same conset) of the first and second morphemic components with the unknown word. Thus, two sets of words (NS_1 and NS_2), sharing one character (in the first and the last position) with the unknown word, are constructed.	In this case, two sets of words are retrieved. NS_1:{責斥、譴斥、眥斥、罵斥、叱斥、譏斥、諷斥、貶斥 }, and NS_2:{訓責、訓譴、訓眥、訓罵、訓叱、訓譏、訓諷、訓誨、訓誡、訓貶}.
3)	Compare these two sets with CILIN, only those words which are also entries in the CILIN are adopted as the near-synonym candidates for the unknown word. (This step resolves the possible ambiguities of morphemic components and over-generated examples from the practical consideration of lexicography.) If no guiding examples are found in CILIN, then the system falls back to step (2) to retrieve these two set using neighbor conset.	NS'_1:{貶斥 (Hc25) } and NS'_2: {訓責 (Hi21)、訓誡 (Hg04)、訓誨 (Hg01)}.
4)	Applying the Equation (6.3) and (6.4) to calculate the semantic similarity between the target word 訓斥 and the candidate words, respectively.	$max(sim_\mu, sim_\nu) = 0.617$, which is the similarity of 訓責(Hi21).
5)	The semantic class of the near-synonym with the maximal semantic similarity value will be the final guess.	In this case, (Hi21) is the final guess, and a correct answer.

Table 6.6: Outline of algorithm with examples

Compound types	Baseline	Our algorithm
V-V	12.20%	42.00%
V-N	14.00%	37.00%
N-N	11.00%	72.50%

Table 6.7: Accuracy in the test set (level 3)

6.3.6 Results and Error Analysis

The correctly predicted semantic category is the sematic category listed in CILIN. In the case of ambiguity, when the unknown word in question belongs to more than one semantic category, the system chooses only one possible category. In evaluation, any one of the categories of an ambiguous word is considered correct.

Primary Results

The *SC* prediction algorithm was performed on the test data, and achieved accuracy of 42.00%, 37.00% and 72.50% in V-V, V-N and N-N two-characters compounds for a task whose baseline was 12.20%, 14.00% and 11.00% evaluated at the 3-level of CILIN Thesaurus, respectively. The resulting accuracy is shown in Table 6.7. For the more shallow semantic classification (the 2-level in CILIN), it worked even better (46.20%, 45.00% and 80.50%), which are shown in Table 6.8. Table 6.10 shows further the *near miss* of the SC system performance, where n stands for the first n ranked semantic classes predicted. The accuracy $\mathcal{A}$ here is defined as follows:

$$\mathcal{A} = \frac{\textit{number of correct predictions}}{\textit{total number of unknown words in the testing data}} \tag{6.6}$$

Error Analysis

Generally, without contextual information, the classifier is able to predict the meaning of Chinese two-character words with satisfactory accuracy against

Compound types	Baseline	Our algorithm
V-V	13.20%	46.20%
V-N	16.00%	45.00%
N-N	12.50%	80.50%

Table 6.8: Accuracy in the test set (level 2)

n	N-N (200)	V-N (100)	V-V (500)
1	72.50%	37.00%	42.00%
2	79.00%	47.00%	54.80%
3	81.50%	51.00%	60.40%

Table 6.9: Performance for the first n ranked semantic class prediction (level 3)

the baseline. A further examination of the bad cases indicates that error can be grouped into the following sources:

(Words with no semantic transparency) Like "proper names", these types have no semantic transparency property, i.e., the word meanings can not be derived from their morphemic components. Loan words such as N-N 摩托 (/múo-tūo/; "motor"), N-N 嗎啡 (/mǎ-fēi/;"morphine"), V-N 可汗 (/kě-hǎn/;"cham") and some dissyllabic morphems such as N-N 蟾蜍 (/chán-chú/;"hoptoad") are examples.

(Words with weak semantic transparency) These can be further classified into four types:

- Appositional compounds:
  Words whose two characters stand in a coordinate relationship, e.g. N-N 東西 ('east-west', "thing").
- Lexicalized idiomatic usage:
  For such usage, each word is an indivisible construct and each has its meaning which can hardly be computed by adding up the separate meaning of the components of the word. The sources of

these idiomatic words might lie in the *etymological past* and are at best meaningless to the modern native speaker. e.g, N-N 薪水 ('salary-water',"salary"), V-V 燒賣 ('burn-sell',"steamed dumpling"), V-N 完蛋 ('over-egg',"be finished")

- Metaphorical usage:
  The meaning of such words are therefore different from the literal meaning. Some testing data are not semantically transparent due to their metaphorical uses, For instance, the system assigned N-N 喉舌 ('throat-tongue', "spokesman") the SC of the N-N 喉頭('throat-head',"larynx")(Bk04), the *correct* SC provided by CILIN is, however, based on its metaphorical use (Aj13). The same is with instaces such as V-N 開口 ('open-mouth', "to say it") vs. V-N 關口 ('close-mouth', "col") and so on.

(Derived words) Such as V-V 進去 ('enter-*directional suffix*', "enter"), V-N 玩兒 ('play-particle', "play") and N-N 鼻子 ('nose-suffix',"nose"). These could be filter out using syntactical information.

(The coverage and quality of CILIN and character ontology) Since our SC system's test and training data are gleaned from CILIN and the character ontology, the coverage and quality of these resources thus play a crucial role. For example, for the unknown compound word 慅慅 ('disturb-disturb',"be in tumult"), there not even an example which has 慅 as the first character or as the second character. The same problem of falling short on coverage and data sparseness goes to the character ontology, too. For instance, there are some dissyllabic morphemes which are not listed in ontology, such as 覬覦 (/jìyú/;"covet").

From the perspective of quality, there are some semantic categories predicted by the system which may sound reasonable to the native speakers, but happen not to be the *correct* answer provided by CILIN.

Compound types	level 2	level 3
V-V	47.80%	43.60%
V-N	45.00%	39.00%
N-N	78.50%	74.50%

Table 6.10: Accuracy in the test set (level 3) after syntactic filtering

For example, V-V 痛打 ('pain-hit', 'lash out') is assigned to the category (Hb08) of 痛擊 ('pain-beat', "bitterly hit"), whose CILIN SC is Fa01; V-N 助手 ('help-hand', "helper") is assigned to the category (Hi36) of 援手 ('aide-hand', "assistant"), whose CILIN SC is Aj09, etc.

Incorporating Syntactic Knowledge

As mentioned previously, we adopt the traditional Chinese philological definition which supposes a *compound word* to be a word made up of two characters. The data that are ramdomly chosen from CILIN therefore include four frequent types of words listed in Table 6.1. From the modern linguistic point of view, some instances can be processed by the separate syntactic module. By filtering out derived words (*bound root* or *root word* plus word-forming *affix*) and grammatical words (word plus grammatical *affix*), the new result is shown in Table 6.10.

6.3.7 Evaluation

So far as we know, no evaluation in the previous works was done. This might be due to many reasons: (1) the different scale of experiment (how many words are in the test data?), (2) the selection of syntactic category (VV, VN or NN?) of morphemic components, and (3) the number of morphemic components involved (two or three-character words?).. etc. Hence it is difficult to compare our results to other models. As set up at the beginning, in order to make the results more comparable to a certain degree, we have chosen the same number of test data with other similar works, which report the best

Compound types	Our model	Current best model
V-V	43.60%	39.80% (Chen 2004)
N-N	74.50%	81.00% (Chen and Chen 2000)

Table 6.11: Level-3 performance in the outside test: a comparison

result on this task to my knowledge.

Among the current similar works, Table 6.11 shows that our system outperforms Chen(2004) in VV compounds, and approximates the Chen and Chen(2000) in NN compounds. However, in comparison with Chen and Chen's model in NN compounds, some critical points should be noted: (1). they evaluate the system performance by examining the semantic classification *manually*, (2). the level accuracy is *not* clearly stated, and (3) the test data do not constrain to the two-characters words. Three-characters words, such as 照相機 ('photography-machine',"camera") are also included. In Chinese compounds, most of the last characters of these three-characters words function as *suffix*, and they often provide a very strong hint for the syntactic and semantic information of the whole words. It is therefore definitely easier to be correctly guessed. The comparison with Chen and Chen's model should be only taken for reference.

6.4 Conclusion

The approach proposed in this chapter stemmed from our desire to answer questions such as:

1. How does Hanzi-concept set participate in the interpretation of the semantic classification of compounds?

2. Would this model/system also work best for all kinds of other compounds? and

3. Which parameters govern the models?

In this Chapter, we propose a SC system that aims to gain the possible semantic classes of unknown words via similarity computation based on character ontology and CILIN thesaurus. The simple scheme we use for automatic semantic class prediction takes advantage of the presumptions that the conceptual information *wired* in Chinese characters can help retrieve the near-synonyms, and the near-synonyms constitute a key indicator for the semantic class guess of unknown words in question.

The results obtained show that, our SC prediction algorithm can achieve fairly high level of performance. While the work presented here in still in progress, a first attempt to analyze a test set of 800 examples has already shown a 43.60% correctness for VV compounds, 39.00% for VN compounds, and 74.50% for NN compounds at the level-3 of CILIN. If shallow semantics is taken into consideration, the results are even better.

Working in this framework, however, one point as suggested by other ontology-based approach is that, human language processing is not limited to an abstract ontology alone (Hong et al. 2004). In practical applications, ontologies are seldom used as the only knowledge resources. For those unknown words with very weak semantic transparency, it would be interesting to show that an ontology-based system can be greatly boosted when other information sources such as metaphor and etymological information integrated. Future work is aimed at improving this accuracy by adding other linguistic knowledge sources and extending the technique to WSD (Word Sense Disambiguation).

Part V

Gaining Perspectives

Chapter 7

Conclusion

This chapter sketches the conclusion, a summary of contributions and some envisaged future research.

7.1 Contributions

The goal of this research is set to survey the unique characteristics of Chinese Ideographs. It is still widely believed today that the general trend in the development of human writing systems is an evolutionary process that began with the *pictogram*, evolved to the *ideogram* and ended with the *phonogram*. The rationale for this viewpoint is mainly the assmption that Chinese characters lack precision. Some (Hannas 2003) even claims that Chinese characters curb Asian creativity. A unifying aim of this thesis is therefore to re-estimate of the role of Chinese characters in scientific theories of meaning/concept in a formal way, especially with respect to concept-based information processing.

Though it has been well understood and agreed upon in cognitive linguistics that concepts can be represented in many ways, using various constructions at different syntactical levels, conceptual representation at the script level has been unfortunately both undervalued and underrepresented in computational linguistics. Therefore, the Hanzi-driven conceptual approach in this thesis might require that we consider the Chinese writing system from

a perspective that is not normally found in canonical treatments of writing systems in contemporary linguistics.

The HanziNet, which we constructed and introduced in this work, is the first lexical knowledge resource based on Chinese characters in the field of linguistic as well as in the NLP. In addition to the contribution in the area of Hanzi teaching and learning, we believe that, introducing conceptual knowledge information encoded in Chinese characters to conceptual modelling is a viable process. An experiment concerning with *sense prediction* yields satisfactory results as well. This said, it has to be conceded that HanziNet, as a general knowledge resource, should not claim to be a sufficient knowledge resource in and of itself, but instead seek to provide a groundwork for the incremental integration of other knowledge resources for language processing tasks. In order to augment HanziNet, additional information will needed to be incorporated and mapped into HanziNet. This leads us to several avenues of future research.

7.2 Future Researches

Although still in its infancy, I believe that the HanziNet proposed in this thesis could eventually provide important insight into the problems of understanding the complexities of Chinese writing systems and its interaction with Chinese natural language. In the long term, I would like to extend the current research to cover other linguistic levels and writings. The following are some suggestions.

7.2.1 Multilevel Extensions

In this thesis, HanziNet was used as a general knowledge resource for building conceptual models. In future, the architecture of this archetypal implementation of HanziNet could be used as a base for a much larger, yet more specific, *reusable knowledge library*. This *reusable knowledge library* could

be viewed as a *multilevel* repository based on the specificity, or refinement, of the concepts found at the most coarse-grained level of description, e.g, as a set of static knowledge sources, vis-à-vis other lexical and domain knowledge resources. This might involve the combination of a lexical knowledge resource (such as Chinese WordNet), and an onomasticon or lexicon, in which the knowledge is typical of a specific specialized domain but whose specific meaning cannot be found at the lexical knowledge level.

7.2.2 Multilingual extensions

Upper level ontology is language-neutral. In this thesis, HanziNet has been proposed as an important knowledge resource for Chinese NLP. With more experimental testing, it could also have the potential to make significant contributions to multilingual studies. These might include: *Mapping CSH to Indo-European Word-roots*;[1] *Hanzi sense disambiguation among CJKV*, and so on.

7.3 Concluding Remarks

In conclusion, a tentative computational theory of Chinese characters was developed, taking advantage of the abundant information contained within the characters themselves, which has not yet been widely recognized by most Chinese computational linguists. In doing so, an enriched semantic network of Chinese characters (HanziNet) is proposed, and was proven as useful in accomplishing specific NLP tasks. Due to the approach employed, with its considerations of borderline aspects in computational linguistics and cognitive science, this approach marks a route for a future Hanzi-triggered concept-based inquiry to complement a statistical approach to NLP.

However, the construction HanziNet, both from theoretical and engineering viewpoints, is still in the early stages of development. There are many

[1]See the American Heritage Dictionary of the English Language, 2000.

details yet to be discovered and discussed. This work has only begun to scratch the surface of such spadework. Though not completely developed, I hope that the primarily representation of this work can at least shed new light on new findings of old problems, and will thus serve to stimulate more research in this new scientific field.

Bibliography

[1] Academia Sinica. (1998). Academia Sinica balanced corpus (Version 3) [Electronic database]. Taipei, Taiwan.

[2] Agirre, E. and G. Rigau. (1996). Word sense disambiguation using conceptual Density. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96).*

[3] Aitchison, Jean. (1985). Cognitive clouds and semantic shadows. In *Language and Communication* 5, 69-93.

[4] Aitchison, Jean. (2003). Words in the mind: an introduction to the mental lexicon. Blackwell publishing.

[5] Aldous, J.M. and Robin J. Wilson. (2000). Graphs and applications. Springer.

[6] Ando, R. Kubota and Lillian Lee. (2000). Mostly-unsupervised statistical segmentation of Japanese: applications to Kanji. *ANLP-NAACL* 2000.

[7] Ann, T.K. (1982). Cracking Chinese puzzles. Vol 1-5. Stockflows Co. Hong Kong.

[8] Baayen, Harald. (2001). Word frequency distributions. Kluwer Academic Publishers.

[9] Baayen, Harald. (2003). Probabilistic approaches to morphology. In *Probabilistic linguistics*, Rens Bod et al (eds). MIT.

[10] Barabasi, Albert-Laszlo and Reka Albert. (1999). Emergence of scaling in random networks. *Science*, 286:509-512.

[11] Ben-Naim, El, Hans Frauenfelder and Zoltan Toroczkai. (2004). Complex Networks. *Lecture Notes in Physics*. Springer Verlag.

[12] Bohn, H. (1998). Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift. Verlag Dr.Kovac.

[13] Buitelaar, Paul. (1998). CoreLex: Systematic polysemy and underspecification. PhD thesis, Brandeis University.

[14] Burg, J.F.M and R.P. van de Riet. (1996). COLOR-X : Using knowledge form WordNet for conceptual modeling. in *Fellbaum, Christiane (ed). (1998). WordNet.*

[15] Bussmann, Hadumod (ed). (1990). Lexikon der Sprachwissenschaft. Kroner: Stuttgart.

[16] Chao, Yuen-Ren. (1968). A grammar of spoken Chinese. London: Cambridge University Press.

[17] Chang, Han-Liang. (1988). Hallucinating the other: Derridean fantasies of Chinese script. *Center for Twentieth Century Studies*. Working paper No.4.

[18] Chen, Chao-Ren. (2004). Character-Sense association and compounding template similarity: Automatic semantic classification of Chinese compounds. ACL SIGHAN Workshop 2004.

[19] Chen, Chao-Ren, Ming-Hong Bai and Keh-Jiann Chen. (1997). Category guessing for Chinese unknown words. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, Thailand.

[20] Chen, Keh-Jiann and Chao-Jan Chen. (2000). Automatic semantic classification for Chinese unknown compound nouns. COLING 2000, Saarbrücken, Germany.

[21] Chou, Yia-Min and Chu-Ren Huang. (2005). 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.

[22] Chu, Bang-Foo. (1998). Nine discourses of wisdom study: Concepts, common sense and systems. (Trans.) Walter, J. van Patten.

[23] Chu, Bong-Foo Lab. http://www.cbflabs.com/

[24] Chu, Iang. (2004). 漢語複合詞語意構詞法研究. (Semantic word formation of Chinese compound words). Peking University Press.

[25] Coulmas, Florian. (1996). The Blackwell encyclopedia of writing systems. Oxford: Blackwell.

[26] Coulmas, Florian. (2003). Writing systems: An introduction to their linguistic analysis. Cambridge University Press.

[27] Croft, William and Alan Cruse. (2004). Cognitive linguistics. Cambridge University.

[28] Daniels, P.T. and W. Bright (eds). (1996). The world's writing systems, New York: Oxford university press.

[29] DeFrancis, John. (1984). The Chinese language: Fact and Fanatasy. Honolulu: University of Hawai'i Press.

[30] DeFrancis, John. (1989). Visible speech: The diverse oneness of writing systems. Honolulu:University of Hawai'i Press.

[31] DeFrancis, John and J. Marshall Unger. (1994). Rejoinder to Geoffrey Sampson, 'Chinese script and the diversity of writing systems.' *Linguistics* 32(3):549-554.

[32] Derrida, Jacques. (1967). De la grammatologie. Paris: Minuit. [Rheinberger and Zischler trans. (1970). Grammatologie, Suhrkamp].

[33] Dorogovtsev S. N. and J. F. F. Mendes. (2003). Evolution of Networks: From Biological Nets to the Internet and WWW. New York: Oxford University Press.

[34] Fellbaum, Christiane (ed). (1998). WordNet: An electronic lexical database. Cambridge, MA:MIT Press.

[35] Fellbaum, Christiane. (1998). A semantic network of English: The mothers of all WordNets. In *EuroWordNet*. P209.

[36] Fernandez, Gomez-Perez, Juristo. (1997). METHONTOLOGY: From ontological arts towards ontological engineering. In *Proc. AAAI97 Symposium Ontological Engineering*.

[37] Findler, N.V. (ed.) (1979). Associative Networks: Representation and use of knowledge by computers. New York: Academic Press.

[38] Fischer, Steven Roger. (2001). A history of writing. Reaktion Books Ltd.

[39] Fukumori, Takahiro and Jun Ikeda. (2002). A classification of writing systems: A step toward general graphology. (in Japanese). In *Journal of General Linguistics*, Vol 4/5.

[40] Ganter, Bernhard and Rudolf Wille. Applied lattice theory: formal concept analysis. In : *A Formal Concept Analysis Homepage*: http://www.upriss.org.uk/fca/fca.html

[41] Ganter, Bernhard and Rudolf Wille. (1999). Formal concept analysis: mathematical foundations. Springer.

[42] Gelb, I.J. (1963). A study of writing: The foundations of Grammatology, Rev.ed., Chicago: University of Chicago press.

[43] Giammarresi, D. and A. Restivo. Two-dimensional languages. in *G. Rozenberg and A. Salomaa, (eds), Handbook of Formal Languages*, Volume III, Springer Verlag, New York, 1996, pp. 215-267.

[44] Goddard, Cliff. (1998). Bad arguments against semantic primitives. In: *Theorectical Linguistics*, Vol.24, No.2-3.

[45] Goddard, Cliff and Anna Wierzbicka. (2002). Semantic primes and universal grammar. In: Cliff Goddard and Anna Wierzbicka (eds), *Meaning and Universal Grammar: Theory and empirical findings.* Vol. I. Amsterdam: John Benjamins, 41-85.

[46] Goldsmith. (2001). Unsupervised learning of the morphology of a natural language. In:*Computational Linguistics (27).*

[47] Grotjahn, R and G. Altmann. (1992). Modelling the distribution of word length: Some methodological problems. In: *Köhler and Rieger (eds). Contributions to Quantitative Linguistics, 141-153.*

[48] Gu, Iang-Kui et al. (2003). Hanzi Etymology Dictionary. China: Hua-Xia press.

[49] Guarino, Nicola. (1998). Some ontological principles for designing upper level lexical resources. *First International Conference on Language Resources and Evaluation.* Granada, Spain,

[50] Guarino, Nicola and Chris Welty. (2002). Evaluating ontological decisions with *OntoClean.* In: *Communications of the ACM 45(2):61-65.*

[51] Guder-Manitius, Andreas. (1998). Sinographemdidaktik: Aspekte einer systematischen Vermittlung der chinesischen Schrift im Unterricht Chinesisch als Fremdsprache. Heidelberg: Julius Groos Verlag.

[52] Halpern Jack. (2002). Lexicon-based orthographic disambiguation in CJK intelligent information retrieval. In: *COLING 2002*. Taipei.

[53] Handke, Jürgen. (1995). The structure of the lexicon: Human versus machine. Mouton de Gruyter.

[54] Hannas, William C. (2003). The writing on the wall: How Asian orthography curbs creativity. Philadelphia: University of Pennsylvania Press.

[55] Harbaugh, R. Zhongwen.com - Chinese characters and culture. [Online:access time 2004]. Available at: http:
www.zhongwen.com.

[56] Harbaugh, R. (1998). Chinese characters: A genealogy and dictionary. Han-Lu Publishing: Taipei.

[57] Harris, Roy. (2000). Rethinking writing. The Athlone Press, London.

[58] Hasan, Maruf and Yuji Matsumoto. (2000). Chinese-Japanese cross-Language information retrieval: a Han character based approach. In: *Proceedings of the SIGLEX Workshop on Word Senses and Multi-Linguality*, pp19-26. Hong Kong.

[59] Hinrichs, Erhard W. (1999). Welchen Beitrag kann die Linguistik zu technologischen Innovationen leisten? In: Meyer-Krahmer und Lange. (Hrsg.). *Geisteswissenschaften und Innovationen*, Physika Verlag, Heidelberg.

[60] Hinrichs, Erhard W and Julia Trushkina (2002). Getting a grip on morphological disambiguation. In: *Tagungsband der 6. Konferenz zur Verarbeitung natuerlicher Sprache* (KONVENS 2002). Saarbruecken, 59-66.

[61] Hinrichs, Erhard W and Julia Trushkina (2004). Rule-based and statistical approaches to morpho-syntactic tagging of German. In:*Proceedings of the Conference on Intelligent Information Systems*. Zakopane, Polen.

[62] Hjelmslev, L. (1961). Prolegomena to a theory of language. Trans. F.J. Whitfield (rev.ed). Madison: University of Wisconsin press.

[63] Hong, Jia-Fei, Xiang-Bing Li and Chu-Ren Huang. (2004). Ontology-based Prediction of Compound Relations - A study based on SUMO. *PACLIC 18*, Japan.

[64] Hoosain, R. (1991). Psycholinguistic implications for linguistic relativity: A case study of Chinese. Hillsdale, NJ: Lawrence Erlbaum Associates.

[65] Hoosain, R. (1992). Psychological reality of the word in Chinese. In H.-C. Chen and O. J.-L. (eds.), *Language processing in Chinese* (pp. 111-130). Amsterdam, Netherlands: North-Holland.

[66] HowNet Knowledge Database. http://www.keenage.com/

[67] Hsieh(Xie), Shu-Kai. (2003a). Revisiting the word length problems of Chinese. In *International Conference of Quantitative Linguistics*. Athens, USA.

[68] Hsieh(Xie), Shu-Kai. (2003b). Do Chinese characters really carry meaning? A critical review. In *3rd Conference of the European Association of Chinese Linguistics*, Ghent, Belgium.

[69] Hsieh, Shu-Kai. (2005a). HanziNet: An enriched conceptual network of Chinese characters. In: *The 6th Chinese Lexical Semantics Workshop*. Xia-Men: China.

[70] Hsieh, Shu-Kai. (2005b). A character-driven three-layers network model of meaning and conceptual representation. *AsiaLex 2005*. Singapore.

[71] Hsieh, Shu-Kai. (2005c). Word meaning inducing via character ontology: A Survey on the Semantic Prediction of Chinese Two-Character. In: *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. Jeju: Korea.

[72] 胡繼明, (2003). 廣雅疏證同源詞研究. 巴蜀書社: China.

[73] Huang, Chu-Ren, Keh-Jiann Chen, and Lili Chang. (1996). Segmentation Standard for Chinese Natural Language Processing. In:*Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

[74] Huang, Chu-Ren, Kathleen Ahrens, and Keh-Jiann Chen. (1998). A data-driven approach to the mental lexicon: Two studies on Chinese corpus linguistics. In:*Bulletin of the institute of history and philology*. Vol.69.

[75] Huang, Chu-Ren, Ru-Yng Chang, and Shiang-bin Li. (2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. *LREC 2004*. Lisbon.

[76] Humboldt, von Wilhelm. (1979). Brief an M. Abel-Rémusat: Über die Natur grammatischer Formen im allgemeinen und über den Geist der chinesischen Sprache im besonderen. [trans. by C.Harbsmeier].

[77] Jiang, Jay and David Conrath. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In:*Proceedings of International Research on Computational Linguistics (ROCLING X)*. Taiwan.

[78] Kauppi, Raili. (1967). Einführung in die Theorie der Begriffssysteme. *Acta Universitatis Tamperensis, Vol.15. Tampere.*

[79] Kess, Joseph F. and Miyamoto Tadao. (1999). The Japanese mental lexicon: psycholinguistic studies of kana and kanji processing. Amsterdam: John Benjamins.

[80] Köhler, R. (1986). Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Bochum: Brockmeyer.

[81] Kuo, W.J, et al. (2004). Orthographic and phonological processing of Chinese characters: an fMRI study. In: *NeuroImage* 21.

[82] Langacker, Ronald. (1987). Foundations of cognitive grammar: Theorectical prerequisites. Standford University Press.

[83] Leacock, C. and M. Chodorow. (1998). Combining local context and WordNet similarity for word sense identification. In C.Fellbaum(ed). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.:MIT Press, 265-84.

[84] Lenat, Douglas. (2002). Artificial intelligence as common sense knowledge. In: *Truth Journal*, online available: http://www.leaderu.com/truth/2truth07.html

[85] Leibniz, Gottfried Wilhelm. (1971). Neue Abhandlungen über den menschlichen Verstand. [trans: E.Cassirer]. Unveränderter Nachdruck, Felix Meiner Verlag.

[86] Li, Charles N. and Sandra A. Thompson. (1981). Mandarin Chinese: A Functional Reference Grammar. Berkeley: U. of California Press.

[87] Lin, Dekang. (1998). Automatic retrieval and clustering of similar Words. *COLING - ACL 98*. Montreal Canada.

[88] Liu and Singh. (2004). ConceptNet - a practical commonsense reasoning toolkit. *BT Technology Journal. Vol 22 No 4.*

[89] Lua, K. T. (1990). From character to word : An application of information theory. *Journal of Computer Processing of Chinese and Oriental Languages*. Vol 4, No 4.

[90] Lua, K. T. (1993a). A study of Chinese word semantics. In *Computer Processing of Chinese and Oriental Languages*, Vol 7, No 1.

[91] Lua, K. T. (1993b). A study of Chinese word semantics and its prediction. In *Computer Processing of Chinese and Oriental Languages*, Vol 7, No 2.

[92] Lua, K. T. (1995). Deriving Proximit Data From The Construct of Chinese Compound Words, Unpublished manuscript.

[93] Lua, K. T. (2002). The Semantic Transformation of Chinese Compound Words (漢語造詞過程的語意轉換). *The 3rd workshop on Chinese lexical semantics*, Taipei.

[94] Luo, Xiaoqiang. (2003). A maximum entropy Chinese character-based parser. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

[95] Luo, Xiaoqiaug and Salim Roukos. (1996). An Iterative Algorithm to Build Chinese Language Models. In: *Proceedings of ACL-96*, pages 139-145.

[96] Lyre, Holger. (2002). Informationstheorie: Eine philosophisch-naturwissenschaftliche Einführung. W. Fink Verlag.

[97] Mei et al (1998). 同義詞詞林. Dong-Hua Bookstore: Taipei.

[98] Meye, Dirk. (1999). Unihan disambiguation through font Technology. In: *15th International Unicode Conference*. San Jose, CA.

[99] Miller,Georges A. (1993). The Science of Words (Wörter: Sreifzüge durch die Psycholinguistik.) Hrsg.und übers. von Joachim Grabowski und Christiane Fellbaum. Spektrum Akademischer Verlag, Heidelberg.

[100] Miller, G.A. Nouns in WordNet. In Fellbaum (ed.) WordNet: An Electronic Lexical Database.

[101] Mohri, Mehryar. (1997). Finite-state transducers in language and speech processing. *Computational linguistics* 23:2.

[102] Morioka, Tomohiko. (2005). Character processing based on character ontology. *Kyoto University 21st Century COE Program.* In: http://www.kanji.zinbun.kyoto-u.ac.jp/ projects/chise/papers/beijing-2005/b2005-chise.pdf

[103] Murphy, Gregory. (2002). The big book of concepts. The MIT Press.

[104] 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.

[105] Old, John. (2002). Information cartography applied to the semantics of Roget's Thesaurus. *Proceedings of 13th Midwest Artificial Intelligence and Cognitive Science Conference.* Chicago,Illinois.

[106] Old, John. (2003). The semantic structure of roget's, A Whole-Language Thesaurus. Ph.D thesis. Indiana University.

[107] Packard, Jerome. (1996). Chinese evidence against inflection-derivation as a universal distinction. In: Cheng and Zhang (eds.) *Proceedings of ICCL-4/NACCL-7, Vol.2.* Los Angeles: GSIL Publications, University of Southern California.

[108] Packard, Jerome. (ed). (1997). New approaches to Chinese word formation. Berlin: Mouton de Gruyter.

[109] Packard, Jerome. (1999). Lexical access in Chinese speech comprehension and production.*Brain and Language, 68*, 89-94.

[110] Packard, Jerome. (2000). The morphology of Chinese. Cambridge, UK: Cambridge University Press.

[111] Paola Velardi, Maria Teresa Pazienza and Michela Fasolo (1991). How to encode semantic knowledge: A method for meaning representation and computer-aided acquisition. *computational linguistics* Vol.17. No.2.

[112] Perfetti, Charles. (1999). The cognitive science of word reading: What has been learned from comparisons across writing systems? In: *The 2nd International conference on cognitive science*, Tokyo: Japan.

[113] Pethő, Gergely. (2001). What is polysemy? - A survey of current research and results. In: Enikő Németh T and Károly Bibok (eds). *Pragmatics and the flexibility of word meaning*. ELSEVIER:UK.

[114] Pollack, Robert. (1994). Signs of life: The language and meanings of DNA. Viking: England.

[115] Porter, David. (2001). Ideographia: The Chinese cipher in early modern Europe. Stanford: Stanford University Press.

[116] Priss, Uta. (1998). Relational concept analysis: Semantic structures in dictionaries and lexical databases (Ph.D Thesis). Verlag Shaker, Aachen: Germany.

[117] Priss, Uta. (2003). Linguistic applications of formal concept analysis. In:*Proceedings of ICFCA 2003*.

[118] Quillian, M. R. (1968). Semantic memory. In *Minsky, M. Semantic information processing*. Cambridge.

[119] Rastier, François. (1987). Sémantique et intelligence artificielle.

[120] Rastier, François. (1991). Sémantique et recherches cognitives. Paris: PUF.

[121] Ren, Manling and David Al-Dabass. (2001). Simulation of fuzzy possibilistic algorithms for recognising Chinese characters. In *International Journal of Simulation* Vol.2 No1.

[122] Resnik,P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of IJCAI. http://xxx.lanl.gov/abs/cmp-lg/9511007.*

[123] Ross, Scheldon. (2003). Introduction to probability Models. AP.

[124] Sampson, Geoffrey. (1985). Writing systems: a linguistic introduction. Stanford, CA: Stanford university press.

[125] Sampson, Geoffrey. (1994). Chinese scripts and the diversity of writing systems. In *Linguistics* 32(1).

[126] Saussure, Ferdinand de. (1916; 1965). Cours de linguistique générale, 3ième ed., publié par Ch. Bally, A. Sechehayer & A. Riedlinger, Paris: Payot.

[127] Schank, Roger C. (1975). Conceptual information processing. North-Holland Publishing Company.

[128] Schank, Roger C. and Abelson, R. (1977). Scripts, plans, goals and understanding. Lawrence Erlbaum Associates.

[129] Schank, Roger C. (1972). Conceptual dependency: a theory of natural language understanding. In Schank and Colby (eds.), Computer models of thought and language, pp. 152-186.

[130] Solé, Ricard V. and Sergi Valverde. (2004). Information theory of complex networks: On evolution and architectural constraints. In: *Ben-Naim, Eli et al (eds). Complex Networks, Lecture Notes in physics 650. Springer.*

[131] Siertsema, B. (1965). A study of Glossematics. The Hague: Matinus Nijhoff.

[132] Sierra,G. and J. McNaught (2000). Design of an onomasiological search system: A concept-oriented tool for terminology. *Terminology* 6(1), 1-34.

[133] Sinica BOW : The Academia Sinica Bilingual Ontological Wordnet. http://BOW.sinica.edu.tw

[134] SUMO : Suggested Upper Merged Ontology. http://ontology.teknowledge.com/

[135] Sowa, John F. (1984). Conceptual structures: information processing in mind and machine. Addison-Wesley.

[136] Sowa, John F. (1991). *Toward the expressive power of natural language,* In J.F. Sowa (ed). Principles of semantic networks: explorations in the representation of knowledge, Morgan Kaufmann, San Mateo.

[137] Sowa, John F. (2000). Knowledge representation: logical, philosophical, and computational foundations, Brooks Cole Publishing Co.

[138] Sproat, Richard. (1992). Morphology and computation. MIT Press, Cambridge, MA.

[139] Sproat, Richard. (1996). A corpus-based analysis of Mandarin nominal root compounds. In *Journal of East Asian Linguistics* 5, 49-71.

[140] Sproat, Richard. (2000). A computational theory of writing systems. Cambridge: Cambridge university press.

[141] Sproat Richard and Chilin Shih. (2001). Corpus-based methods in Chinese morphology and phonology. Lecture notes for the 2001 Summer Institute of the Linguistic Society of America. University of California.

[142] Stalph, Jürgen. (1989). Grundlagen einer Grammatik der sinojapanischen Schrift. Wiesbaden: Harrassowitz Verlag.

[143] Steyvers, M. and Tenenbaum, J.B. (2002). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*.

[144] Su, Xin-Cun. (1995). Contemporary Chinese lexicology. [in Chinese]. China: Guang-Dong Education Publisher.

[145] Takahiro Fukumori and Jun Ikeda. (2002). A classification of writing systems: A step toward general graphology. (in Japanese) In *Journal of General Linguistics*. Vol. 4/5.

[146] Takaki, Ryuij. (2001). Towards a reformation of Chinese ideographs. *Forma*, 16.

[147] Taylor, John. (2002). Cognitive grammar. Oxford University Press.

[148] Trost, Harald. (2003). Morphology. In M. Ruslan (ed). *The Oxford Handbook of Computational Linguistics.*

[149] Tsai, Chih-Hao. (2001). Word identification and eye movements in reading Chinese: a modeling approach. Ph.D. thesis. University of Illinois at Urbana-Champaign.

[150] Tseng, Huihsin and Keh-Jiann Chen. (2002). Design of Chinese morphological analyzer. *First SIGHAN Workshop*.

[151] Tseng, Huihsin. (2003). Semantic classification of chinese unknown words. *Proceedings of the ACL-2003 Student Research Workshop.*

[152] Tzeng, O. J.-L., Hung, D. L. and Wang, W. S.-Y. (1977). Speech recoding in reading Chinese characters. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 621-630.

[153] Tzeng, Ovid.et al. (1992). Auto activation of linguistic information in Chinese character recognition. In:*Advances in Psychology. Vol.94*, p119-130.

[154] Unger, J. Marshall. (2004). Ideogram: Chinese characters and the myth of disembodied meaning. Honolulu: University of Hawai'i Press.

[155] Valiente, Gabriel. (2002). Algorithms on trees and graphs. Springer.

[156] Vandermeersch, Léon. (2001). Writing in China. In Christin Anne-Marie (ed), *Histoire de l'écriture: De l'idéogramme au multimedia*. Flammarion.

[157] Vossen, Piek. (1998). EuroWordNet: a multilingual database with lexical semantic networks. Dordrecht: Kluwer Academic Publishers.

[158] Vossen, Piek. (2003). Ontologies. In M. Ruslan (ed). *The Oxford Handbook of Computational Linguistics.*

[159] Wang, Hui and Shiwen Yu. (2003). The semantic knowledge-base of contemporary Chinese and its applications in WSD. *Proceedings of the 41st ACL.* Japan.

[160] Wang, Niang-Suen.(1744-1832). 廣雅疏證. 上海古籍出版社影印, reprinted in 1983.

[161] Wang, Jason. (1983). Toward a generative grammar of Chinese character structure and stroke order. Ph.D thesis. University of Wisconsin.

[162] Wang, J., Inhoff, A. W., and Chen H.-C. (eds.). (1999). Reading Chinese script: a cognitive analysis. Mahwah, NJ: Lawrence Erlbaum Associates.

[163] Wang, Patrick. (1987). The intelligent Chinese characters. In *International Conference on Chinese and Oriental Language Computing*.

[164] Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393:440-42.

[165] Watts, D. J. (2004). Small worlds: The dynamics of networks between order and randomness. Princeton University Press.

[166] Widmaier, Rita. (1983). Die Rolle der chinesischen Schrift in Leibniz's Zeichentheorie. Wiesbaden: Steiner.

[167] Wierzbicka, Anna. (1996). Semantics, Primes and Universals. Oxford: OUP.

[168] Wong, Shun-Ha. S. and Karel Pala. (2001). Chinese radicals and Top Ontology in WordNet. In:*Text, Speech and Dialogue: Proceedings of the Fourth International Workshop, TSD 2001*, Lecture Notes in Artificial Intelligence. Springer.

[169] 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.*

[170] Wong, Shun-Ha. S. (2004). Fighting arbitrariness in WordNet-like lexical databases- A natural language motivated remedy. In: *The Second Global WordNet Conference, Brno: Czech Republic.*

[171] Wille, Rudolf. (1992). Concept lattices and conceptual knowledge systems. *Computers and Mathematics with Applications 23*.

[172] Wille, Rudolf. (1992). Begriffliche Datensysteme als Werkzeug der Wissenskommunikation. In *Mensch und Maschine - Informationelle Schnittstellen der Kommunikation*.

[173] Wille, Rudolf. (1995). Begriffsdenken : Von der griechischen Philosophie bis zur Künstlichen Intelligenz heute.

[174] Wille, Rudolf. (2000). Begriffliche Wissensverarbeitung.

[175] Hannas, William. (1997). Asia's orthographic dilemma. Hawaii: Uni.of Hawaii Press.

[176] Wisniewski, E. (2000). Similarity, alignment, and conceptual combination: Comments on Estes and Glucksberg. *Memory and Cognition* 28, pp.35-38.

[177] Wolff, Karl Erich. (1993). A first course in formal concept analysis. In: Faulbaum, F (ed). *SoftStat'93: Advances in Statistical Software*.

[178] Wong, S.H. Sylvia and Karel Pals. Chinese radicals and top ontology in EuroWordNet. In V. Matoušek et al(eds). (2001). *Text, speech and dialogue*. Lecture Notes in Computer Science. Springer.

[179] Wong, S.H. Sylvia. (2004). Fighting arbitrariness in WordNet-like lexical databases - A natural language motivated remedy. *The second global WordNet conference*. Brno, Czech Republic.

[180] Wu, Dekai and Pascale Fung. (1994). Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In: *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 180-181, Stuttgart.

[181] Xia, Fei. (1999). Extracting tree adjoining grammars from bracketed corpora. In: *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99).*

[182] Xu, Tong-Qiang. (2004). 漢語研究方法論初探. (A primary exploration of methodology of Chinese linguistics). Peking: Shang-Wu press.

[183] Yiu, Candy and Wai Wong. (2001). Chinese character synthesis using METAPOST. In:*TUGboat*, Vol.24, No.1.

[184] Yip, Po-Ching. (2000). The Chinese lexicon: a comprehensive survey. Routledge.

[185] 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.

[186] Zhou, X. and D. Marslen-Wilson. (1996). The nature of sub-lexical processing in Chinese. *Experimental Psychology Society*, Bristol.

Appendix A

Test Data

```
% ************************************************************
%  NN 200
% (level-3 accuracy)
%  -> wrong: 55/ right: 145 ... 145/200= 72.50%
% (level-3 with syn filtering)
%  -> wrong: 51/ right: 149 ...149/200=74.50%
% (level-2 accuracy)
%  -> wrong: 39/ right: 161  ... 161/200= 80.50%
% ************************************************************
Bk04 門齒            An03 盜匪            Bh12 Dg03 胚芽
AAc03 美人           Ec02 湖色            An01 匪幫
DDn10 毫升           Ae13 男生            Bi07 狼狗
BBh11 花苞           Dc03 姿勢            Dh01 Ak03 魔鬼
AAd03 里閭           Da21 Be06 泥坑       Bl05 雞蛋
AAe10 軍長           Ba08 廢物            Da09 火災
Dm07 報館            Af04 店主            Ae10 士兵
BBa04 貨物           Dm04 酒館            Di25 法紀
AAe15 名醫           Bi06 牡馬            Ga01 悲哀
AAe17 名伶           Bi21 螟蟲            Di14 民風
```

EEb25 片時
AAf02 傭人
BBf01 冰塊
AAh05 伯母
AAh08 夫妻
AAh09 胞兄
DDg05 屏障
DDd05 範疇
Dc03 面容
DDf01 Df02 心神
BBq05 草帽
DDi02 江山
BBi08 家鼠
DDk29 昆曲
BBk02 顱腦
BBk13 脊柱
BBl03 耳屎
DDl03 脚癬
AAf10 廳長
BBk09 Bn19 鹿角
BBp33 雞心
BBk04 犬牙
BBp19 船票
AAg04 財主
BBp13 木魚
DDj07 軍餉
Bm02 鐵窗
AAb04 寶寶
BBo21 貨車
BBr13 蜜糖
Bk02 腦顱
Db09 Dg07 門路
Be05 河道
Bg02 波谷
Bh12 禾苗
Bk04 Hi42 口角
Da21 家道
Db06 Db07 理路
Bi06 母馬
Dj07 稿費
Bi12 家鴿
Dk19 碑誌
Cb06 Cb14 邊際
Ae07 漁民
Bk08 拇指
Bl03 耳屎
Bp18 家譜
Bp17 墨汁
Bi19 雄蜂
Bn03 客廳
Bp20 國旗
Da07 病歷
Bq03 皮衣
Dj04 本錢
Dk03 課卷
Bn08 鋼窗
Bn24 佛寺
Bp18 帳本
Br02 糧草
Bp09 Bo27 Bp36 盒子
Bi18 毛蚴
Di02 盟國
Ca14 年紀
Bn01 茅屋
Dk30 劍術
Bn11 巷子
Bi13 母雞
Bn11 馬道
Dk11 夢囈
Bf02 北風
Af10 局長
Bl01 汗液
Bk08 脚趾
Af02 侍女
Bi19 雌蜂
Af10 首領
Bn22 墓穴
Bq03 棉衣
Bp10 花藍
Br12 奶酪
Bn22 墓碑
Bn23 宗廟
Bn24 佛堂
Dj05 本幣
Dk31 卷軸
Dk18 課表
Dk20 卷帙
Ca09 年底
Cb08 城區
Br04 米湯

DDa21 病狀
CCb18 墳地
BBo25 車輪
DDc04 De04 Df14 眼光
BBl03 眼屎
CCb25 Di02 鄉鎭
DDm04 飯店
BBh02 蘭花
Dd15 敬稱
AAe10 Di11 陸軍
AAe17 丑角
AAg09 肉票
AAj13 喉舌
BBr08 冰糖
BBq01 棉布
BBr13 蜂蜜
Bm04 土塊
BBk03 耳朶
BBp13 南胡
BBh07 荔枝
DDf01 腦海
Bk05 乳房
DDf14 牙慧
DDb06 邏輯
AAn07 鴒兒
Bk03 鼻子
AAe15 Dk03 西醫
BBo15 模具
HHg07 口試
BBo22 輪船
Ab04 孩童
Dk03 課業
Cb18 墓地
Bi16 河蚌
Bk13 顱骨
Bg09 筆跡
Dd05 男方
Bn03 廚房
Ae03 水手
Bm04 泥巴
Bg01 冰塊
Dh02 龍宮
Bo01 摩托
Br12 冰糕
Bq02 棉綢
Br09 蜂糕
Dj07 薪水
Bp33 戒指
Da09 手氣
Bh07 鳳梨
Bf02 麗風
Bd02 Bi10 蟾蜍
Di14 民俗
Dk09 節目
Dk20 史書
Dk27 史詩
Dm04 客舍
Bk11 羽毛
Bh06 北瓜
Dk29 曲劇
Bo31 火箭
Bk14 肺臟
Dm07 劇場
Bk17 唾腺
Dk02 西學
Bk13 頭骨
Bm01 金子
Dk06 漢語
Aa01 Ba01 Da28 東西
Ai01 鼻祖
Da14 Dd02 Di01 Di20 毛病
Ak03 歹徒
Da21 Df05 民情
Bp12 Bp35 木馬
Aj11 貨主
Dj08 軍費
Br14 嗎啡
Br19 油條
Dd15 乳名
Bq03 風衣
Bp23 火柴
Bh09 芝麻
Dk05 Dk06 字母
Eb23 火速
Ab01 Ab04 Af02 Ah14 Y 頭
Bn23 佛殿
Bq05 布鞋
Dk29 歌劇
Bk17 淚\ 腺
Bm04 岩石

BBk09 膀子 Aa01 Bi01 畜生

% **
% VN 100
% (level-3 accuracy):
% -> 37(right)+63(wrong) -> 37/100= 37.00%
% (level-3 with heuristic filtering):
% -> 39(right)+61(wrong) -> 39/100= 39.00%
% (level-2 accuracy):
% -> 45(right) + 55(wrong) -> 45/100= 45.00%
% **

Gb01 想法	Aj15 伴郎	Da08 逆境
Cb22 裂隙	Al04 Ba08 廢物	Bo22 渡船
Bi07 獵犬	Hf07 返鄉	He09 If24 Dj08 賠帳
HHg11 揮毫	Dj08 Hi11 匯款	Fb01 拔腿
Cb21 缺口	Hc28 Hj11 辦事	Hn01 犯罪
Ib09 懷孕	Hj52 投江	If11 獲獎
Hc25 If07 罷官	Hi40 Dk11 笑話	Gb20 留心
Hg09 鍍金	If06 得志	Hg08 Hg10 讀書
EEd06 低級	Hi63 結伴	Hj55 帶孝
HHh05 開球	De04 聽力	Ic01 揮淚\
Gb11 抱怨	Ib15 暈車	Hg19 製圖
DDa08 If03 幸運	Ie14 完工	Id03 赤脚
Hg11 揮筆	Hi10 賀年	Fa01 Fa29 Hi44 Hj12 Ig01 動手
HHi02 省親	Gb10 抱恨	Hk48 散心
DDa09 Ef09 幸福	Hd23 攻子	He08 欠債
Hi55 玩兒	Bm07 試管	Ed24 丟臉

Am01 施主
AAe17 編劇
Hi12 笑語
AAn06 醉鬼
BBd01 織女
AAk05 Ee16 滑頭
Bp13 揚琴
BBp19 執照
BBm09 引柴
CCa04 Cb20 關口
AAf07 閹人
DDa04 疑雲
Ee24 傲岸
DDj05 支票
DDi25 約法
HHm06 刷卷
FFc03 鎖眉
FFa01 還手

Hj04 違約
Ae15 穩婆
Hi15 告警
Af05 可汗
Ga14 醉心
Hg19 設色
Hg21 用藥
Hh03 用光
Ib08 遺尿
Hi18 問罪
Hi14 表白
Fc04 把風
Ie16 誤點
Ka01 要死
Ib03 If22 完蛋
Hi36 援手
Hi39 抗命
L03 賞光

Ha06 問鼎
Ed43 有鬼
Ee42 俗氣
Gb17 關心
Ed07 Ed48 便當
Bn03 暗房
Hg20 看病
Hg20 Hj22 懸壺
If05 揚名
Ib09 有身
Hn10 舞弊
Hi42 頂嘴
Hi19 Hi41 說嘴
Hi19 吹牛
If18 失火
Ee34 傲物
Fc06 Hi12 搶嘴

```
% ************************************************************
%  500 VV
% (level-3):
% -> 210(right)+290(wrong) -> 210/500= 42.00%
% (level-3 heuristic filtering):
% -> 218(right)+282(wrong)-> 218/500= 43.60%
% (level-2):
```

% -> 231(right)+269(wrong) -> 231/500= 46.20%

% **********

Ef02 興盛

Fa10 開鑿

Fc04 偷看

Fc04 Gb08 窺見

Gb08 熟悉

Gb09 寵愛

Gb10 怨恨

FFc04 觀看

GGb03 推想

Gb03 臆度

Fc04 窺探

Hd05 Fa27 切削

Fd06 發顫

Hb06 轟擊

Hi03 邀約

Hi12 亂說

Ie03 對抗

Gb03 揣測

HHi05 會見

Hi33 推委

IId22 綿聯

HHi37 看顧

Id21 緊靠

Id14 搖晃

Id14 搖蕩

FFb01 停步

Fb01 止步

Fa08 採摘

Fc04 回眸

Ga01 哀傷

Gb08 看穿

Fd04 Ea08 蜷曲

Fa01 拷打

Hi33 推卸

Hi32 婉謝

Gb06 渴想

Gb03 估計

Hi05 會晤

Gb10 憎恨

Hi32 婉謝

Gb03 推論

Hi10 問安

Id22 連綿

Id25 攪和

Hi03 邀請

Fb01 Hi01 走動

Fa30 Gb12 Hi12 Id05 傾倒

Id14 搖動

Hn03 搶掠

Fa30 翻倒

Fb01 散步

Hi47 辯誣

Hi40 稱讚

Ie07 未婚

Hi12 暢談
Jd08 消亡
Fa21 沐浴
Hj01 借住
Ie08 歸併
Ie05 烘襯
IId08 陳放
Id10 噴射
Fa01 拍打
HHi12 閒扯
Id11 飄盪
Hi13 演講
Fc04 Hj26 察看
Fa32 Id14 搖動
Fa07 擁抱
Gb03 猜測
GGb10 厭惡
Gb10 仇恨
Fa32 Id14 抖動
GGb10 嫌棄
Hi02 拜見
Hi05 陪同
Gb04 謀求
Gb05 籌劃
HHi12 話別
GGb06 懷念
Hi12 淺談
Ic09 僞裝
Hi40 讚許\
IId02 隆起
Fa01 掠笞
Jd08 消泯
Jd08 消散
Id11 流動
Id14 搖擺\
Ef01 Ef13 淆亂
Id21 毗鄰
Fa01 敲打
Jd08 Hb13 Jd10 消滅
Id12 翻騰
Fd05 跌倒
Ef13 紊亂
Id18 偶遇
Id15 迴旋
Gb04 Df08 希望
Fa30 Gb12 Hi12 Id05 傾倒
Hi41 爭論
Gb10 厭棄
Hi41 辯論
Hj56 參見
Gb03 猜想
Gb04 祈望
Hi32 允許\
Hi12 敍別
Gb10 憎恨
Ic04 哀嘆
Id03 覆蓋\
Ic09 佯死
Gb03 推測
Id03 曝露

Id03 裸露
Je03 沾染
JJe05 授予
JJe08 牽制
Je08 制止
Je04 挑起
Je04 引致
Je08 Gb16 抑制
Je11 警備
Ga17 Je02 Fa29 觸動
Id12 飄舞
Id12 飛舞
GGb06 懸念
IId13 浮盪
IId14 擺\ 刋
Hi12 說笑
GGb06 掛念
Id05 蜷縮
Id02 鼓出
IId18 遇到
IId18 巧遇
FFa09 Id06 折斷
Id04 張合
HHi01 結交
Gb01 深思
Id07 吊掛
IId03 障蔽
FFa19 拆卸
FFc05 聽見
Gb09 憐愛
Ic04 靜默
Je14 Hc07 收受
Je08 壓制
Je14 領受
Je05 給予
Je10 戕害
Je10 傷害
Ga17 Df01 感觸
Id03 遮蔽
Id03 遮蓋\
Id08 擺\]
Gb05 計謀
Hi12 談笑
Id13 飄盪
Gb06 思念
Hi10 慶賀
Id02 突出
Gb09 鍾愛
Id03 遮擋
Id18 撞見
Id18 相逢
Id15 旋繞
Hi02 探親
Id07 矗立
Id07 屹立
Id18 相遇
Hi02 探訪
Id05 隕落
Hi18 詰問
Id05 倒塌

IId07 聳立
IId05 掉落 Hi02 覲見
IId05 脫落
Id18 偶逢
IId06 損壞
Id20 阻塞
IId05 摔下
Id20 蔽塞
Id06 破壞
Id20 閉塞
Id21 靠攏
Id20 梗塞
JJd10 Fa21 掃除
Id21 逼近
IId23 縈繞
Id25 糅合
Jd10 清除
Id21 靠近

Hi18 訊問
Hc24 招募
Id18 巧逢
Id21 貼近
Id20 暢達
Id07 存放
Id07 懸掛
Id07 寄放
Id21 迫近
Id23 纏繞
Id24 扭結
Id25 淆雜
Id25 混雜
Id23 環繞
Id25 糅雜
Id23 圍繞
Id23 環繞

Ca11 Hj63 過去
Hj64 進去
Hj19 混入
Hj64 出來

Hi40 醜化
Hj19 加入
Id02 凹下
Ja01 看成

Dd15 Fc10 Ic01 呼號
Hg19 描繪
Gb04 Gb12 想望
Fc09 呼喚
Id06 毀壞
Gb18 哀矜

Jd11 毀棄
Fa13 張掛
Hc25 任用
Hc24 聘用
Gb03 想見
Gb10 哀怨

Hg11 描摹

Fa01 痛打
Br09 燒賣
Hh04 哀吟
HHc20 講評
Hi25 哀告
Hi25 哀求
FFc10 哀鳴
Fd02 依偎
Fa31 張開
HHi40 挖苦
Gb17 愛惜
Hi02 拜見
HHj38 掩藏
Ga01 哀愁
Dd15 愛稱
GGb09 Gb13 愛慕
Hi02 Hi25 拜託
Fa32 搬動
EEd24 哀榮
Hi21 數落
Ag04 哀鴻
HHk01 拜功\
Hj02 搬遷
Hh01 搬演
FFa15 拾取
Ga01 哀痛
Df05 Gb09 愛戀
GGa01 哀戚
Hi36 幫補
Ie13 頒行
FFa34 Hi60 Hj59 搬弄
Hi36 幫襯
Jd04 保有
Hh01 扮演
Hi38 保育
Hi16 保舉
HHf05 搬運
Jd05 包容
Hi16 保送
HHc07 Hc11 頒發
Hc11 頒佈
Hj07 奔忙
JJd05 Gb18 包括
Hc01 備辦
Hm05 釋放
JJd01 保全
Hc02 Ka15 保管
Ib04 睡醒
EEe39 保守
Hm06 審訊
Hj28 使用
DDi24 Ka15 保證
Hc04 Hj28 使喚
Ic01 涕泣
HHj40 保存
Id20 滲透
Fc05 收聽
Hj15 報效
Hd04 Hj35 Hm07 收拾
Jd01 保持
IId04 伸展
Da14 Da24 收穫
Hi37 保駕
Hc07 收取
He12 報銷
Ia08 爆炸
Hi16 保薦
Hc15 報到
Ia08 Id06 爆裂
Hc15 報廢
Fb01 奔馳
Hc25 備取
HHj40 保藏
Hg16 編譯
Hi62 報復
DDk23 備考
Fd08 奔突
Ib18 崩漏
HHj28 備用
Hm05 保釋
Hg16 Hj58 編造

EEd22 奔放
HHe04 出租
DDk23 備註
IId11 奔瀉
HHg16 Hg17 編寫
IId06 崩裂
HHb02 參戰
HHj10 Hj24 Je14 承受
EEe10 超脫
HHb13 撤退
BBe05 Id11 Ed22 奔流
HHc22 Hm07 懲罰
JJd06 充塞
AAe17 編導
HHd12 編織
FFb01 Hj38 奔逸
HHc12 播揚
HHe25 貶謫
HHi14 If23 Jd06 表露
IIh01 變遷
HHi55 Hi60 播弄
HHh03 Hh03 播放
HHj19 Hj20 參加
JJe04 敦促
HHa06 Hb12 Je12 奪取
GGa16 害怕
HHi60 勾搭
HHb03 攻打
HHc20 Hm06 裁判
IIh06 減低
Hb01 備戰
Hi39 Hi50 觸犯
Hb03 抵抗
Hi29 報償
Hi29 報答
Hi40 讒害
Hg17 編錄
Ae16 編審
Hc01 籌辦
Hi14 Ja03 表示
Fd08 衝突
Hi02 報聘
Di20 Hj59 Jd06 表現
If22 崩潰
Hc21 表彰
Hi14 表達
Hg12 Hi02 Hj56 參見
Hg17 編纂
Ee11 放肆
Hc25 貶黜
Fb01 飛奔
Hc14 否決
Hi41 辯論
Hi47 辯難
Hj38 躲閃
Ig02 擱置
Gb08 Jc06 貫通
Hc14 Hm06 裁定
Id07 寄放
Hj66 Id25 攪混
Db01 報應
Ie01 傳佈
Id05 崩塌
Dd06 Di09 Hd12 Hg17 編製
Hg12 參考
Hc18 查問
Hi13 闡述
Dd15 Ja02 稱呼
Gb12 崇尚
Hc25 貶斥
Hg17 編排
Hj38 奔竄
Hh03 播發
Hi21 貶責
Hb08 奔襲
Da02 Ih01 Ih02 變動
Fb01 奔騰
Fb01 奔跑
Ih01 Ih02 變更
Hi41 辯駁
Hi34 奉勸
Ia09 焚燒
Hj08 奮鬥
Fa11 Ib17 Id20 Je07 堵塞
Hi23 過問
Dg05 Jh04 管束
Id15 滾動
Hb03 夾攻
Hc27 繼任
Ha02 Hb03 反抗

Hj47 鑒賞
Hc19 監視
Kd04 仿造
Hd01 建造
EEb23 緩慢
Hi37 護理
Gb08 獲知
He10 Hi27 付出
HHj30 Ie08 合併
Gb06 回憶
AAc17 禁絕
EEb27 Ed39 枯\ 竭
HHe14 If24 虧欠
Fc04 看見
Hf07 Id21 離開
Hi13 論說
Ba07 陪送
HHi58 欺壓
Hc20 評定
Fa01 拍打
Hi49 欺騙
Hh01 排練
Hi46 趨附
Hc24 聘任
Hi61 侵害
Hj25 取捨
Hi25 求見
Hi45 屈從
Hg18 刪改
Ha06 Hn03 攘奪
Hm10 絞死
Fa26 攪拌
Hc12 發動
Ga17 Je02 感動
Ie02 積壓
Hi38 撫養
Hc18 核准
Hi34 Ic05 Je06 激發
Hb08 截擊
Hc05 開創
Ae12 Hb04 看守
Ha02 抗議
Hi26 饋贈
Hg19 描繪
Je12 謀取
Hi58 排擠
Hj61 乞討
He03 搶購
La02 請問
Hi25 求援
Hc07 簽收
Ja05 取決
La03 屈就
Jd10 驅除
Dg05 Gb06 牽掛
Da01 Da14 閃失
Hi34 勸說
Hg18 去除
Gb16 忍耐
Hc05 Hd02 設置
Hi24 打聽
Hd01 改建
Je11 防範
Ed23 Ia12 Ie12 腐敗
Hi06 Ie09 告別
Gb23 悔改
Hb13 擊退
Hb03 進攻
Hf03 疾駛
Id24 糾結
Hc11 刊佈
Fd09 叩拜
Hc25 錄取
Hi54 Ic03 拿捏
Fd08 爬行
Hc20 評比
Hb09 Je08 牽制
He03 傾銷
Hb11 Hn10 侵佔
Hb14 擒拿
Dk11 Hi34 勸告
Hn03 搶劫
Hj51 求歡
Dd04 Hj26 區別
Hj38 潛伏
He15 Ed39 He15 缺乏
Jd10 去掉
Hi34 勸誘
Hn05 殺害
Jd11 捨棄

Je10 傷害	Ka19 擅自	Hg10 涉獵
IIa09 燒毀	Hb04 設防	Je12 攝取
HHb06 射擊	Ga01 傷痛	Hj38 閃避

Appendix B

Character Semantic Head: A List

CSH	CSK code	Character Examples
日 (sun)	強光, 熱燥, 巨遠, 自然物, 一天	明
門 (door)	固定形, 通口, 大, 人造物, 生活用	開
月 (moon)	弱光, 涼靜, 巨遠, 自然物, 三十天	月
肉 (meat)	不定形, 柔軟, 身體部分, 功能	肥
骨 (bone)	不定形, 堅硬, 身體部分, 支撐	髀
貝 (shell)	形色麗, 堅硬, 小, 動物, 材料用	贈
目 (eyes)	圓形, 黑色, 小, 感覺器官, 在臉部	眼
金 (gold)	反光, 硬銳重冷, 自然物, 材料用	錫
木 (wood)	長條形, 穩固實在, 植物, 材料用	棍
韋 (leather)	可塑形, 柔軟堅韌, 人造物, 材料用	韎
水 (water)	無形色, 柔冷液態, 自然物, 生存用	河
火 (fire)	光熱, 常動, 氣態, 自然物, 生存用	燒
米 (rice)	白色, 硬, 細小, 作物, 生存用	糠
土 (earth)	不定形, 多實廣重, 自然物, 材料用	堤
走 (walk)	動態, 脚動體位移, 正常態, 生活用	追
竹 (bamboo)	長條形, 堅韌尖銳, 植物, 材料用	筮
白 (white)	光亮無色, 潔淨, 物表狀態	皎
禾 (cereal crops)	條形, 綠色, 矮小, 作物, 生存用	稻
身 (body)	特定形, 整體, 事物之主要部位	軀

CSH	CSK code	Character Examples
鬼 (ghost)	無形色, 陰冷可怕, 未知態, 危險態	鬼
彳 ()	人互動行爲, 關係態, 生活性	行
牛 (ox)	四足, 有角體大, 動物, 勞力用	牡
舌 (tongue)	扁圓形, 柔軟, 小, 感覺器官, 在口中	舌
風 (wind)	無形色, 可動, 自然態, 在空氣中	颱
舟 (boat)	長形, 可動, 大, 人造物, 交通用	艇
卯 ()	幾何形, 固定, 大, 人造物, 遮蔽用	
示 (show)	行爲態, 恭敬愼重, 祭祀, 對天地鬼神	祭
毋 ()	幾何形, 固定, 大, 人造物, 保護用	
穴 (cave)	內陷形, 固定, 大, 自然物, 人造物	岭
車 (vehicle)	幾何形, 可動, 大, 人造物, 交通用	輪
麥 (wheat)	綠色多芒, 矮小, 作物, 生存用	麵
仝 ()	反常態, 不舒適, 生理現象, 生存態	
犬 (dog)	四足, 矮小, 親和, 動物, 生活用	犬
衣 (cloth)	不定形, 多色多質, 人造物, 蔽體用	襟
巾 (towel)	方形, 多色, 小, 人造物, 蔽體用	巾
虫 (insect)	多變形, 可動, 小, 動物, 各種功能	蟬
刀 (knife)	尖銳利硬, 小, 長, 人造物, 工作用	刀
片 (flake)	薄形, 具兩面, 形狀, 工作用	片
厂 ()	幾何形, 固定, 大, 人造物, 遮蔽用	廠
雨 (rain)	點滴形, 冷, 液態, 自然態, 天候現象	霧
玉 (jade)	不定形, 多色堅硬, 自然物, 材料	玉
歹 (bad)	非常態, 不好, 認知態, 生存態	殘
石 (stone)	不定形, 堅硬, 自然物, 人造物	碑
酉 (wine)	無形色有香, 液態, 醱酵態, 人造物	酸

CSH	CSK code	Character Examples
弓 (bow)	弓形, 有彈性弦, 人造物, 攻擊用	弓
角 (horn of animal)	尖鋭形, 堅硬, 自然物, 人造物	角
子 (son)	後代, 小, 認識態, 陽性	孩
魚 (fish)	頭尾鰭身, 水生, 水生動物, 生存用	鯽
矛 (lance)	尖鋭形, 堅硬且長, 人造物, 攻擊用	矛
阜 (mound)	山形, 人聚處, 自然物, 人造物	
邑 (capital)	不定形, 人聚處, 建築物, 生活用	邑
人 (human)	兩足直立, 可動, 角色, 同類	休
食 (eat)	不定形, 動態, 加工品, 生存用	食
气 (air)	無形色, 氣態, 自然物, 生存用	氣
心 (heart)	心形, 柔韌, 小, 生理器官, 生存用	心
手 (hand)	手形, 柔韌, 小, 生理器官, 生活用	抓
口 (mouth)	口形, 柔軟, 小, 生理器官, 生存用	吃
足 (foot)	足形, 柔韌, 長, 生理器官, 生活用	跑
馬 (horse)	四足, 高大, 善動, 動物, 工作用	驤
髟 (hair)	細密形, 深色, 生理組織, 保護用	髮
耳 (ear)	耳形, 柔韌, 小, 感覺器官, 聽覺用	聽
艸 (grass)	細條形, 柔韌, 小, 植物, 生存生活用	草
山 (mountain)	山形, 穩重, 厚大, 自然物, 生存環境	岳
女 (female)	兩足直立, 人類, 同類, 陰性	嫂
糸 ()	長細形, 柔韌, 微, 人造物, 生活用	絲
田 (farming)	方形, 地表, 大, 空間範圍, 生存用	畦
囗 ()	空間, 巨大, 空間範圍, 地域	圍
黑 (black)	無光之色, 不可見, 物表狀態	黯
刂 ()	空間位移, 動態, 空間行爲, 道途	
言 (speech)	無形色, 概念, 概念溝通, 生活用	說

Appendix C

Character Ontology

0 OBJECTIVE DOMAIN

00 ENTITY

000 BEING

00000 SUBSTANCE

00000A CHEMICAL : 金銀銅鐵錫氧碳鋁磷砒氫氮

00000B ELECTRIC : 電波磁

00000C AQUATIC : 海洋湖泊沼澤藪池蕩潭淀江

00000D TOPOGRAPHY : 陸原塚谷野坪坎坷嶽岳巒岫峰漠

00000E FORM : 灰岩石磐碌塵粉砂沙玉琨瓊瑤瑗末土埃

00000F BORDERING SUBSTANCE : 磯岸湄埼墩灘島洲嶼渚礁

00000G MATERIAL : 材釉埴墐壌赭墡堊鹵潟泥

00000H ENERGY : 油沼焱樵木碳煤炭酒柴薪

00000I LIVING ORGANISMS : 人獣鳥魚貝蟲籘禽畜牲樹

00001 FLORA

00001A FOOD PROVISIONS : 稲禾穭糯糠豆麥粳糧秣餱稓粟穀_ 玉米

00001B VEGETABLES : 蔬菜蒜薯匏葫蘆茄_蘿蔔_葱瓜

00001C FRUITS : 檸檬枇杷葡萄柿蕉榴柑桔橘_柳橙_棗梨椰李_楊桃

00001D PLANTS FOR BEAUTY : 曇菊蘭棠荷松柏榕楓楊柳玫瑰桂

00001E TRUNK : 芽苗秧柄蒂榦桿柯梃標株枚稿穰廱秸莖枝稈棳根_芒_刺

00001F BUILDING MATERIAL : 茅荊棘楚棕櫚竹杉槐梓籐

00001G TECHNICAL CROPS : 茶黍梧桐橡蔗菸桑栗棉麻

00001H PHARMACEUTICAL : 種莢糠花英華秀穗果核蕊絮苞艾

00001I AQUATIC : 藻苔覃_昆布

00001J WEEDS : 草莽茅

00010 LIMBS

00010A LIMBS : 胸手爪指掌身體軀躬背頭首面臉脖頸顱

00010B STRUCTURE : 隔膜筋骨腮骼骷髏\[肋髑腱肌脊凶顱額

00010C BODY SURFACE : 甲角犄鱗殼皮膚胼胝革韋羽毛髮眉鬍鬚鬢翎

00010D BODY FLUIDS : 絲膠膟膏血衄巇髓腺乳奶

00010E SENSE ORGANS : 腦目眶瞳眸瞼耳心意眼睛

00010F ORGANS : 臟脾胰腎腑胃膽腸心肺肝

00010G UPPER LIMBS : 臂膀胛肘胳肱肱腕膊髆肩

00010H LOWER LIMBS : 腿胯髀髖膊髕膝脚足腓脛

00010I EXCREMENT : 涕屎糞便唾涎垢精卵汗痰淚\

00011 ANIMALS

00011A BREEDING : 鵝鴿鴳鵪雞鴨兎豬羊牛豚狗犬貓馬

00011B BIRDS : 鴻鷗鶫鶘鴛鴦鶴鶤鷺鸛鸚

00011C FOUR-FEET : 猩狒虎猿猴獅狐狸狼狽豹\

00011D WATER ANIMALS : 蚌蠣膃肭蝦蟹蝸蠔蜆介豚龜鱉_玳瑁_蛤蜊_牡蠣_蠔螺蚶蜆黿

00011E FISH : 鱷鯨鯽鱖鰱鱮鯉魨鰍鱔鰳鱠魷鰻鯧鯊鱟鯽鱒鯣鰆鱸鮑

00011F REPTILES : 蚯蚓蠍蝎蜈蚣蜘蛛蛇蟒蝮

00011G FLYING INSECTS : 蛾蝴蝶蚊蜉蝣蠅蟑螂螟蚱

00011H REPRODUCTION : 精卵蛋蝌蚪胚胎蛹蛆孑孓

00011I LEGENDARY : 龍鳳麒凰麟鯤角夔饕餮鵬

00011J MICROORGANISM : 菌黴_酵母_病毒

0 OBJECTIVE

00 ENTITY

001 ARTEFACT

00100 FOOD/CLOTHING

00100A FOODSTUFFS : 飯菜米麵餌粥糜餬糕餅粿饅

00100B MEET FOOD : 膾膴胾臠醓醢膿膰胙魚肉

00100C CONDIMENTS : 糖醋醬油鹽餡餖

00100D EATING STYLE : 宴伙饔齋葷素筵秀酒席膳餐\

00100E DRINKS : 飲茶酒汁_啤酒_茗

00100F GIFTS : 砒霜_鴉片

00100G FOOTWEAR : 靴舄襪蹝鞋屨履屣屧屐屜

00100H HEADGEAR : 帽冠笠冕帽盔簽鑐鎧韘韝髮鬄纂巾

00100I MATERIALS FOR CLOTHS : 布絮綢麻棉絲毛皮革絨紬巾冪縵_尼龍

00100J COSTUMES : 飾刺繡妝

00100K WEARING APPAREL : 衣服衫褝襖褧袞裳褲裝袍褂裙

00100L PARTS OF A GARMENT : 領_口袋_袖襟

00101 CONSTRUCTION

00101A ENVIRONMENT : 庭園城郭墟莊院場埕坪場

00101B PRIVATE RESIDENCE : 房屋閣寓舍宅室墅樓廈廬

00101C PUBLIC BUILDING : 館店社庫牢廠鋪棧肆攤窯寮塔坊倉

00101D AFTER DEATH : 冥壙窀穸塚厝墓墳棺槨柩碑塋塚墦\

00101E BUILDING PART : 牆壁階檣基梁柱樁垣堰堤塘壩欄杆柵闌_樓梯_廊

00101F CHANNEL : 港橋徑道路途阡陌埂嶝衢衖衕溝渠_胡同

00101G CONSTRUCTION MATERIAL : 磚瓦木石土泥沙_水泥_瀝青

00101H VEHICLE : 船舟艦筏車蹕機艇舳艫輪舢

00101I COMPONENT : 局閂輪軸轂轅轄軏楫軌榫

00101J CULTURAL INSTITUTE : 寺廟堂祠館校所

00110 RESIDENCE

00110A GENERAL : 物品具器

00110B EATING UTENSIL : 筷箸柊叉匙匕碗盤碟缽罏瓢杯觴

00110C KITCHEN TOOL : 爐灶墟鑪筴鍋甑釜錡鑁壺鼎

00110D CONTAINING UTENSIL : 缸桶罈匋鬲瓶盆罐槽盂鑵

00110E FILL UTENSIL : 籠_籮筐_箱匣盒筐函箋籃簍簞筒袋

00110F DECORATION : 釵簪笄珈鐲釧圭玦環璧_蠟燭_珮\

00110G HOUSEHOLD UTENSILS : 線繩枴杖棍棒鏡鑑梳巾傘鐘錶扇縷\

00110H FURNISHINGS : 桌檯案椅凳屜鎖_屏風_凳座几架櫥櫃

00110I WASTE MATERIAL : 屁尿屎糞盯聹垢渣滓粄泔垢屑

00110J TOYS : 牌球棋骰盧牌弈傀儡偶毽箭

00110K BEDDING : 蓆簾帳幕枕被褥床墊席

00110L CLEANING UTILITIES : 帚_刷_畚箕

00111 WORK

00111A WRITING MATERIAL : 稿紙簿册筆聿墨卷帙尺籤書報畫硯

00111B MUSICAL INSTRUMENT : 琴笛鑼鈸鈴鼓梆鐘鏞笙簫笳琯

00111C POTTERY : 陶瓷玻璃蠟漆瓦磚鏝鋼銑

00111D AGRICULTURAL TOOL: 誄耜鋤耙杪犁鏟鍬鎬鐮刈網羅

00111E FISHING TOOL : 罟罾鉤叉釣竿罻網

00111F MACHINE : 鏈鐐鍵鐲鐶樞鈕鑰匙銀鎖

00111G WEAPON : 彈炮槍砲鎗弓戟干戈劍盾鏢箭刀刃

00111H MEASURING APPARATUS : 秤斗儀尺規表針計圭臬板桿錘_砝碼

00111I CRAFTWORK TOOL : 刀斧剪槌錘隼銼鉋鏝圬鋏鉗鑷砧椎錐釘

0 OBJECTIVE

01 ABSTRACTION (象 SYMBOL、數 NUMBER、性 GENDER、徵 CHARACTER、關係 RELATION、意義 MEANING、條件 CONDITION、利害 INTEREST)

010 SPECIFICATIONS

01000 SYMBOLIC DEFINITIONS

01000A CLASSIFICATION : 界門綱目科屬種類乾兌離疇

00100B ORDER : 甲乙丙丁戊己庚辛壬癸子_首_次初

00100C NUMBER : 一二三四五六七八九十廿壹貳參肆侮陸柒捌玖拾

00100D TIME : 年季月日時分秒前昨今明期旬更辰節旬春夏秋冬曆

00100E SPACE : 上下前後左右內外中東西旁邊側面_附近

00100F MEASURE UNIT : 碼米浬哩呎吋里尺寸分

00100F RATE : 率比例倍折扣

00100H CURRENCY : 毫分鎊塊元毛角鍰兩貫弔

01001 GRAMMAR

01001A MODALITY : 必定應該得能夠要願甘肯

00101B IS-verb : 是爲然即乃係有非否匪甭

00101C PERSON PRONOUN : 我灑俺敝予余偺朕孤你妳

00101D LOCAL PRONOUN : 這那此彼哪

00101E ADVERB : 傯緊遑忽乍驟溘突驀猝漸

00101F PARTICLE : 幾奈曷奚何詎況之乎也者的嘛嗎哩呢唷

00101G ONOMATOPEIA : 唧咪喃咕咻嗄咿噹咚嘟吁哞汪喵吱啾喔嘎_淅瀝

00101H CONNECTIVE : 又且連予與和及同跟並但_或者_否則_因爲_所以_於是_此外_豈

00101I PREPOSITION : 因以替代給乘趁趕沿順照

00101J CLASSIFIER : 趟回次番度服劑頓口份屆塊

01010 ROOTS

01010A ORIGINS : 宇宙時空元本末性命智慧能精

01010B GENDER : 公母正負陰陽男女雌雄牡

01010C GEOMETRY : 點線面體橫直斜曲形樣狀

01010D POWER : 能力電磁熱聲氣味物質朕

01010E KNOWLEDGE CAUSE : 眞假公私稟賦體用因果事業席

01010F ETHNIC GROUP : 漢滿蒙回藏蠻狄羌蕃韃靼華

01010G REDIDERE : 省市縣鄉鎭村里區路巷弄郡府州

01011 PHENOMENA

01011A ASTRONOMIC : 景天昊穹霄日月林地山川宇宙星辰鈞

01011B SURFACE : 痕斑繡紋鏽玟玷瑕瑜轍跡皺漬縷\

01011C FORM : 角方矩格稜橢珠球蛋邊面圓錐

01011D SITUATION : 質數量爆炸籟氣氛音響勁勢功\

01011E LIQUOR : 氣汽液漿汁泡瀋膠韌韌糊

01011F SOLID : 窟窿圈眼口孔坎窪穴堀

01011G FLUID : 潮汐波浪濤瀾澐漩渦漣漪

01011H VAPOR : 風冰風露雲雯靄霧霾雨霖霪雪雹雷霜虹煙雷電

01011I CULTURAL : 風俗情文化習慣

0   OBJECTIVE

01   ABSTRACT

011   APPLICATION

01100 MESSEGE + INTERFACE

01100A INDICATION : 符號模圖表言語資料文字概念

01100B IDENTIFICATION : 性別姓名字號甫籍貫住址

01100C SIGN : 招牌旗幟麾帘纛榜徽碑銘碣

01100D VALUEABLE : 價值誼恩仇惠利害毒壽功\

01100E TEXTUAL UNIT : 句段節章題篇檔輯場景曲

01100F TEXTUAL STYLE : 序箴跋詩詞訓詁韻諺謎彖賦令曲

01100G DOCUMENT : 帖牒稿信函箋牋牘簡柬札

01100H CERTIFICATE : 旨詔證照徵憑狀例單據執

01101 GAINS AND LOSSES 情緒 (對個人切身意義之事件) EMOTION

01101A (Characterized by) GOOD LUCK : 幸福吉祥瑞慶祺禎禧泰禔運

01101B (Characterized by) SOCIAL VIRTUES: 道德廉恥勤儉誠信仁恕愛篤實諾忠義孝悌禮

01101C (Characterized by) ARTS : 藝術文學圖畫雕塑庭園音樂曲調歌

01101D (Characterized by) MONETARY VALUES : 金銀財寶貝珠款錢鈔帑幣貨

01101E SUFFERING from DISASTER : 苦災厄旱澇潦禍殃燹殍饉

01101F SUFFERING from SICKNESS : 疾瘩癬痔瘡腫瘤病疵恙暍皰疥癆瘀痢瘴症癌

01101G SUFFERING from INBORN MISFORTUNE : 眇瞍瞽瞎瞶盲聾聵啞殘廢痴

01101H SUFFERING from LIFE MISFORTUNE : 孤獨煢罪辜瘋癡癖鰥寡孀\

01110 SOCIAL AFFAIRS

01110A LEGAL AFFAIR : 憲法律制刑規則秩序契標條例約款紀訴訟

01110B POLITICAL AFFAIR : 門派社黨品邦國宗族家氏

01110C BUROCRATIC : 府院部司署課組校所局科

01110D POSITION : 揆長吏僚員帥特簡薦委使夫手官師家者校將

01110E DOMESTIC : 政軍經濟武戎戰稅賦稼穡捐

01110F OCCUPATION : 士農工商教警軍兵醫漁牧妓鴇

01110G BORDERING : 郊野限圻際徼範疇鄰疆域邊界緣

01110H FEUDAL : 黌宬朝廷宮闈殿闕泮邸寺

01111 HUMAN RELATIONS

01111A CONSANGUINITY : 父母爸庭考媽慈妣娘婆伯爺兒女祖孫

01111B AFFINITY : 岳夫妻妾姬妗荊配偶伉儷婆姑姨婿侄甥嫂舅

01111C RELATION : 親戚嫡胞胄嗣裔表眷乾朋輩鄰仇敵恩友

01111D RELIGION : 仙佛神祇妖怪魔鬼僧喇嘛耶穌_阿拉_靈_菩薩_魂魄

01111E GENERAL TERM : 丈翁傖叟嫗媼漢郎佬婦娃媒妁孩童嬰兒

01111F MORAL BEHAVIOUR : 聖賢俠彥豪傑儐客盜匪宄徒賊_英雄_漢士鬼儈_郎中

01111G PRINCIPAL AND SUBORDINATE : 紳曹官庶民姝僑闆倌僕役奴隸主

01111H FEUDAL TERM : 皇帝王君后嬪妃_太子_駙馬宦閥公侯伯子爵首魁主東領

01111I GENERAL : 娶嫁婚逑配贅醮祭祀禓禋

1 SUBJECTIVE

10 PERCEPTION : 刺激 (stimulus)、分辨 (distinction)、狀態 (state)、感受 (feeling)、程度 (degree)、印象 (impression)、情緒 (emotion)。

100 SENSATION

10000 VISIBLE SENSATION

10000A BRIGHTNESS : 曖晞曈朏明朗亮昭暗晦暝\

10000B CHROMATIC : 色黃綠藍紫白素灰黑玄烏黛盧紅赤丹

10000C LIGHTNESS : 皓皎暉晰犖霽顯蒙龍瞞眛

10000D QUALITY : 巴粘淡乾燥純粹雜濃稠黏

10000E SHAPE : 平直縱橫斜豎立正歪偏隆彎曲坦

10000F BODY FORM : 厚薄狹仄高低粗細矮扁大菲小胖瘦微龐

10000G SITUATION : 岩嵌崢崎崇嶔巍峨嶒嵩巖

10000H STATE : 澄湛清潔淨澈渾濁污洪沛

10001 OTHER SENSATIONS

10001A GUSTATORY : 酸甜苦辣鹹澀

10001B TACTILE : 細嫩滑粗澀暖煴溫涼熱燠皮緻稹密嚴精縝疏弛酥硬_軟柔牢固剛

10001C OLFACTORY : 香馨芬芳臭臊腥羶腓郁焦菲

10001D ORGANIC : 餓渴癢楚僵腫嘶啞瘖噎骾

10001E FATIGUE : 憊眩勞昏暈恍惚疲睏倦累

10001F AUDITORY : 響嘈謐悄闃希靜默岑寂幽

10001G LONELY SENSATION : 仃單孑蕭寥拮据稔乏安穩_孤獨

10001H PERPLEXING : 迷僻惘困惑漠湎秘密怪詫

10010 AFFECTION :The psychological features of the mental state of a living org
the psychological features that arouse an organism to action .

10010A PLEASURABLENESS : 謐恬逸享悠閑嫻喜樂愉悅怡歡欣慆爽愜舒舒適暢

10010B CHERISH : 緬愐懷念嚮記掛惦冀盼待憶感觸

10010C DESIRE FOR HUMAN : 寵愛好戀慣倖嗜耽貪婪饞欲想喜慾

10010D DESIRE FOR MATERIAL : 節儉省撙吝嗇惜慳苛慷慨儲滿足

10010E NERVOSUS : 忐忑_緊張_戒警

10010F ANNOYING : 愁頹怏沮嘆煩悶憂惆悵憮惝\

10010G EMBARRASSING : 窘尬尷糗

10010H MENTAL SUFFERING : 悱惴悁悒忡鬱怛戚惾悲傷哀

10010I DETEST : 嫌藐厭惡憎懣忮罣慍怒憤恨責怪怨懟

10010J REGRET : 懊悔_嘆嗟_欷噓_慨喟

10010K HESTITATE : 猶豫

10010L DISDAIN : 鄙漠藐蔑睨傲輕嗤

10010M BENUMB : 麻痺茫迷惑疏離

10010N RESPECT : 佩服尊敬器重

10010O JEALOUS : 嫉妒羨慕

10010P SYMPATHY : 憐憫同情諒

10011 （認知）印象 (impression caused by others behavior stimulus)

10011A（意性）: 愎佶倔頑訑拗怯憨懦苟木羞臊_緬靦_怯赧

110011B（質性）QUALITY : 才雋智慧睿聰敏庸伶俐黠\

10011C（體性）FIGURE : 帥俊倩豔秀壯醜陋肥胖瘦

110011D（行性）APPEARANCE : 瀟灑倜儻莊嚴肅穆逍遙雍

110011E（良性）TAME : 嫻馴婉乖淳澹樸朴謙賢善

110011F（表性）OUTER CONDUCT : 邋遢猥褻儇佻騷酷妖冶嬌

110011G（感性）FEAR : 驚懼駭噤囁嚅愕憚恐悚怖

110011H（惡性）FEROCITY : 蠻虐凶戾狠暴兇刁奸狡猾

110011I（嬌性）COQUETRY : 俏嗲嬌

1 SUBJECTIVE

10 KNOWING

101 STATE

10100 MOTION (gravitation、operating force、anti-operating force, etc.)

10100A ROTATING MOTION : 轉迴旋迴斡咼縈翻滾緣傾

10100B DISPLACING MOTION : 運艤晃裊盪宕騰舞娑飄浮移曳蕩擺\

10100C FORMED MOTION : 長漲伸延縮舒展脹張膨脖凸凹陷隆鼓崛突

10100D FORCE MOTION : 閃噴射鋋爆炸迸崩怦決潰濺溢

10100E ENERGETIC MOTION : 下降落隕墜墮掉上升昇冒謝塌坍垮倒

10100F DYNAMIC MOTION : 震振碰砸撞搗舂碾軋輾輘

10100G FLYING MOTION : 飛翔翩翥翾頡頏飄繙颳颺

10100H AQUATIC MOTION : 漏流淌汩漫淋瀉泄溢氾濫湧奔潺涓洄湲滲沔\

10101 STATE OF CHANGE/TRANSFORMATION

10101A TIME STATE : 古昔曩早先初常晚遲新陳

10101B SPACE STATE : 曠闊敞博袁魁京喬穹廣堯

10101C ENERGY STATE : 映照耀曬炫熔鑠煬沸滾焚爆炸

10101D FORCE STATE : 強勀弱倏疾威猛狂勃厲劇

10101E CONTAINING STATE : 滿盈彌光空枴罄充深覃湛淺幽

10101F CHAOS STATE : 繁紛紜縟叢簡紊亂糅蓬鬆散漫

10101G THRIVING STATE : 菁茂茁葱蔚荒蕪萎孳蓼榮盛繁華

10101H INCEPTING STATE : 妊娠生產誕殖種孵娩孕滋

10101I CUASE-EFFECT STATE : 導致引勾挑招使讓

10110 STATE OF INTERFACE

10110A OPEN : 開闢闔張翕通暢透

10110B BLOCK : 關閉阻封堵杜遏淤堵壅塞隔閡滯蔽卡鯁軋梗

10110C CONNECT : 連接銜襲嗣聯結亙繼續賡攏合分毗鄰綿

10110D (觸界) : 支承觸及搭挨貼偎傍靠倚懸吊湊掛

10110E (形界) : 互介跨翕併並比峙臨界

10110F (體界) : 糾結交叉紐締綜綿絮纏繞繚旋圍環拱瑩扭

10110G (狀界) : 浸漬浞泡沾涵淹溺浴渲染

10110H VECTOR : 墊襯枕掌堆卡疊托壁亭屹聳矗豎

10110I QUANTATIVE : 湊兼攤磊多複少幾

10110J MIXTURE : 混雜摻攙淆糅和

10111 STATE OF RECOGNITION

10111A（時識）: 起始啓創肇終止既竟竣艮完結

10111B（事識）: 經歷_虔妥順怕惕畏凜懍崇逆

10111C（形識）: 現曝露呈暴敞彰赤袒粲裸

10111D（形識 2）: 蒙蔽遮擋籠罩屏障埋覆蓋\

10111E（狀識）: 破碎齏敝解斷裂綻消銷泮毀損絕散滅缺

10111F（意識）: 好優良佳嘉壞劣莠歹竄對安貞潔蕩名

10111G（力識）: 動靜開關停頓住休息止卻快慢疾緩徐

10111H（判識）: 勝成捷負敗_敗北_足綽夠敷得失違

10111I（利識）: 資珍貴靖貧窮窶富裕饒卑

10111J（緣識）: 巧_偶然_緣_緣分_偶碰遇撞_邂逅_遇逢

1 SUBJECTIVE

11 EXCITABILITY

110 (行爲能力)

11000 ORGANS ABILITY

11000A EATING : 吃食啖饌茹齋飲喝啜飼喂吸攝品嚐嚼嚥吞

11000B SEEING : 看視瞧瞻望觀眺矚盯瞅瞄瞰盯眈眸顧見

11000C HEARING : 讀詠誦唸歌唱謳呼叫嚷喚喊嚎嚷喧號吼鳴啼嘶唳噪

11000D MOUTH : 嗝呃嗆吹哮喘嗤噓歔含叼咬舔叮噙舐噬嗑螫呼吸吐

11000E（意理）: 夢魘旅遊逛娛戲玩耍撩弄

11000F ORGANIC FUNTION WITH EMOTION : 瞪瞠睜睚瞋瞑瞇乜膘眨抿覷皺蹙顰噘嘟

11000G PHSIOLOGICAL : 睡眠寢盹寐甦醒覺歇憩瞌

11001 PHYSICAL (BODY AND LOWER LIMBS)SKILLS

11001A GOING : 行走爬趴攀匍匐陟登渡跑步徒

11001B STANDING : 站立佇企廁坐跏趺跪仰翹

11001C BODY ACT : 親摟擁抱吻掙扎攔擋妨礙攬

11001D MOVE OF LOWER LIMBS : 踏踩蹀踹蹈踩踐跆蹍履蹬踢踮跨

11001E GALLOP : 駸馳騁騖騵驅騑騃駢驃猋

11001F SPECIAL MANNERS OF MOVING : 踽迤邐躚踱躂逗遛跼蹌蹣躥偃仆跤撐摔撲豕跌蹣跚\

11001G DIRECTIONAL MOVE : 進赴趨晉入出退往離返去衝鑽

11001H LOAD : 負荷載承扛背撐挺挑擔馱

11001I LEAN : 傾靠偎依躺趴臥

11001J CURL : 蠕滾翻蜷曲仰彎磕扭收縮

11001K STRETCH : 伸展延

11001L QUIVER : 哆嗦_顫抖慄噤_觳餗

11001M HEAD MOVEMENT : 抬舉昂點仰翹

11010 SKILLS OF UPPER LIMBS

11010A MOTION : 抬拱招揚搧指舉揭

11010B SHAKE : 揮昜搖抖甩擺\

11010C TOUCH : 搭捫摸撫按捺托撮撥捋摩挲摁

11010D CONTROL : 拿把扼握持捧拋撇扔丟投擲撒掬搯扼執操

11010E CUT : 斬劈砍剖切割剉削刮剷絞剪刻剁

11010F PRICK : 戳刺插塞嵌捅扎掖扦

11010G HIT : 擊打揍搥捶擂撾鞭抽敲拷掠撻撲拍摑摜

11010H CHANGE FORM : 攤摺捲卷折扳拗撅綑捆綁撥扭擰

11010I CHANGE STATE : 採摘擷束縛撈抄揣填塞堙搾剝撕挖扒揭掘鑿

11010J CHANGE DIRECTION : 撿拾搬移挪舉擎提曳拎推收拉扯拖排挽掣拔擢

11010K CHANGE SURFACE : 揩拭擦抹拂撢掃抆塗敷搽剃攪扮摻攙搓揉撥

11010L SCRATCH : 抓捉取搔捕攫

11010M WITH FINGER: 掐捏拶

11010N VIA WATER : 撈舀

11011 INTELLECTUS (the ability to learn and reason, to think abstractly)

11011A OBSERVE : 勘看視覽伺瀏聆聽聞嗅觀閱\

11011B ASSIDERE (to estimate the value) : 計辨認驗證証鑒據鑑估算

11011C REASONING : 推想思憶遺忘忖省猜揣測臆

11011D REPRODUCE : 描擬倣摹

11011E COMPREHENSION : 懂憬悟明曉會知識宜可符

11011F REPRESENTATION : 著譜撰述抄寫記錄簽謄繕誌刻

11011G PLAN : 圖謀企擬營擴拓祈禱祝搜求索

11011H DISCRIMINATION : 挑選遴甄揀比較揆媲校斟考判

11011I RESTRAINT : 待熬捱禁忍耐憋努律斂專收拘束制壓控限克抑羈縛掣遏

1 SUBJECTIVE

11 REACTION

111 EXPERIENCE

11100 LIVING SKILLS

11100A COOKING : 烹飪炊煮熬蒸燜餾燉燴煲

11100B BODY AFFAIRS : 穿披褪捋戴頂飾妝扮剃沐洗刷淋浴滌澡漱盥

11100C HOME AFFAIRS : 修葺剪搞弄補衲鉤裁編縫清掃

11100D LIVING AFFAIRS : 寓居宿棲羈留泊屯駐航划

11100E AGRICULTURAL : 佃墾播種植培育犁耕耔耘鋤_灌溉

11100F TECHNIQUE : 建築砌垙礪鍛煉鑄冶淬鍍

11100G EDUCATIONAL : 教育陶毓訓導誨迪學習摹

11100H (物事) : 劃刊輯編印纂剟設舖鋪裝畫繪摹放陳擺\

11101 COMMUNICATION

11101A IDEAS INTERCHANGE : 說道曰云講敍謂述稱告訴闡言語話談吭\

11101B CASUAL CHAT : 咕嚕_嘮叨_囔囔_聒聒_囉唆_支吾_聊扯

11101C REPORT : 報告曉喻誥詔通稟諭

11101D CONSULTING : 允論辯爭詰問訊諮詢商議咨討答應許\

11101E ADMIRING : 請讚嘉謝贊賀捧媒妁賞譽奉恭敬慶祝誇推薦褒獎頌許\

11101F POLITE : 叩拜跪揖鞠躬覿覲謁詣宴晉參訪探拜會省晤見

11101G CONCERN : 囑咐叮嚀諒宥恕饒勸諫勵勉

11101H EXPRESSING POSITIVE EMOTIONS : 笑莞哂嬉鬧噱_莞爾_呵

11101I EXPRESSING NEGATIVE EMOTIONS : 哭泣啼慟號嚎_欷噓_哽咽

11101J DENOUNCE : 訌譴責訐訾罵詈斥叱譏諷訓誡誨貶嘲

11101K SWINDLE : 訛搪僭佯誇譎吹詮詔唬騙敷衍_應付_纏詐欺脅迫

11101L SILENCE : 緘噤默

11101M CURSING : 詛咒

11101N SLUGGISH IN WORDS : 呆楞痴頓

11101O HYPOCRITICAL : 僞佯裝

11101P ANTAGONISM : 逼迫壓榨

11101Q SPREAD : 傳播揚散布謠訛

11110 DEALING WITH THINGS

11110A with VOLITION : 殉誓決拼操克剋恃逞貫徹

11110B (處行) : 戒捨棄罷消減免刪禁戢革委卸

11110C (處位) : 反對_背叛仇攻攘防守抵抗峙

11110D with. AFFAIRS : 傳送授受辦執掌託設施務

11110E with. OBJECT : 領取收納供給付寄匯輸遞獲存授予賦

11110F with. WARE: 貿兌換買購贖賣售貨販花欠還_賄賂

11110G with. FORCE : 擾侵犯闖搏鬩搗蹂躪戰鬥

11110H with. FEROCITY : 偷竊盜扒奪搶劫掠拐篡括

11110I CAPTURE : 擒捉斥獵屠逮

11111 THE WAY GETTING ALONG WITH PEOPLE

11111A DISPATCHING PEOPLE : 使駕馭統率御攬宰治管轄用任斥撤

11111B COMMANDING PEOPLE : 宣佈諭喩示譲俾命令吩咐

11111C EMPLOYING PEOPLE : 聘雇徴募傭擢薦役掖代替任聘請招

11111D COERCING PEOPLE : 監督排抑箝挟控馴懾鎮

11111E SERVING PEOPLE : 恭尊敬順従服伏遵循皈依待謹

11111F HARMING PEOPLE : 打懲罰囚拘錮坑拷箠鞭笞處毒撃

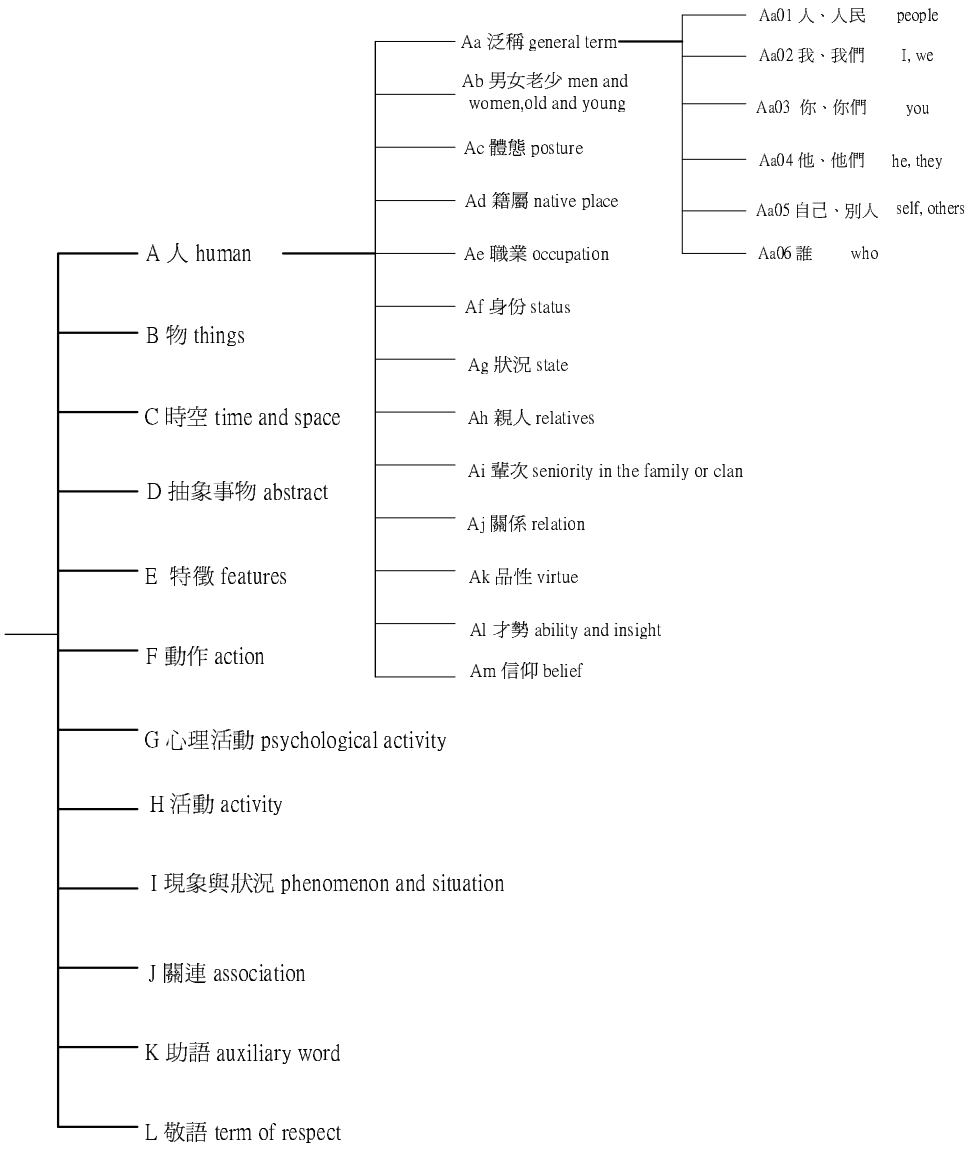
11111G KILLING/HURTING PEOPLE : 勦剿誅殲夷毀滅泯宰殺殊屠決戮戕傷害損誤

11101H SOCIALIZED WITH PEOPLE : 謝辭拒駁推絶酬交往締結_搭訕_處邀約陪伴迎對待

11111I PROTECTING PEOPLE : 保護袒戍衛廕庇佑顧輔佐呵養

Appendix D

A Section of Semantic Classification Tree of CILIN


A 人 human
B 物 things
C 時空 time and space
D 抽象事物 abstract
E 特徵 features
F 動作 action
G 心理活動 psychological activity
H 活動 activity
I 現象與狀況 phenomenon and situation
J 關連 association
K 助語 auxiliary word
L 敬語 term of respect
Aa 泛稱 general term
Ab 男女老少 men and women,old and young
Ac 體態 posture
Ad 籍屬 native place
Ae 職業 occupation
Af 身份 status
Ag 狀況 state
Ah 親人 relatives
Ai 輩次 seniority in the family or clan
Aj 關係 relation
Ak 品性 virtue
Al 才勢 ability and insight
Am 信仰 belief
Aa01 人、人民 people
Aa02 我、我們 I, we
Aa03 你、你們 you
Aa04 他、他們 he, they
Aa05 自己、別人 self, others
Aa06 誰 who