1 | Introduction

1.1 Rationale: why investigate Chinese words?

Why is Chinese morphology worth investigating? To many, the very posing of this question will seem to suggest an ironic lack of relevance, due to the common belief that Chinese ‘doesn’t have words’ but instead has ‘characters’, or that Chinese ‘has no morphology’ and so is ‘morphologically impoverished’. The powerful influence that characters have over conceptions of the Chinese language has led many investigators (e.g., Hoosain 1992, Xu 1997) to doubt the existence of words in Chinese. My goal is to demonstrate that speakers of Chinese compose and understand sentences just as speakers of any language do, by manipulating sentence constituents using rules of syntax, and that the smallest representatives of those constituents have the size, feel, shape and properties of words. And while Chinese may not have word forms that undergo morphological alternations such as *give*, *gave*, *giving* and *given*, Chinese does indeed have ‘morphology’, and the morphology that it has is of a most intriguing and enlightening sort.

Understanding how Chinese words are constructed and used is critical for a full understanding of how the Chinese language operates. Chinese native speakers possess implicit knowledge about the structure and use of words. For example, a native speaker knows that you can change *shuìjiào* 睡觉 sleep-sleep ‘sleep’ to *shuìguojiào* 睡过觉 sleep-ASP-sleep ‘have slept’ or *tiàowǔ* 跳舞 jump-dance ‘dance’ to *tiàoguowǔ* 跳过舞 jump-ASP-dance ‘have danced’, but that you can’t in the same way change *jiějué* 解决 undo-decide ‘decide’ / *chūbǎn* 出版 emit-edition ‘publish’ to get **jiěguojué* *解过决 undo-ASP-decide ‘have decided’ or **chūguobǎn* *出过版 emit-ASP-edition ‘have published’. By the same token, the native speaker knows that it is fine to say *tiàodegāo* 跳得高 jump-EXTENT-tall ‘can jump high’ but not **tuīdeguǎng* *推得广 push-EXTENT-wide ‘can push wide’. In this book, I will explain how the native speaker knows these facts about words by describing the form that this knowledge takes. I do this by proposing generalizations that explain the regularities in the creation and use of words, and then offering principled explanations for the exceptions to those generalizations. Following current trends in cognitive science, I shall argue that much of what native speakers know about words and their structure occurs innately in the form of a hard-wired, specifically linguistic 'program' in the brain, and that such hard-wired word structure information is realized in surface form upon exposure to linguistic data.

Following that line of reasoning, Chinese words are worth investigating because they have the potential to tell us a great deal about the universal properties of words in natural language. Chinese words traditionally have been considered uninteresting as objects of morphological investigation because they do not manifest characteristics thought critical to the concept 'morphology' (such as grammatical agreement or morphophonemic and paradigmatic alternation). In the pages that follow I will show that Chinese words are particularly suitable for asking different but equally interesting questions about words – for example, how words evolve, how they come into being via lexicalization, abbreviation or borrowing, and how they pass out of existence through reduction or grammaticalization. Chinese is particularly suited to answer these questions because Chinese word components are relatively easy to isolate, identify and track over time.

Chinese words exhibit other properties that must be understood if we wish to claim a universal characterization of words. For example, to what extent is the concept of 'bound root' – which is important in Chinese (see 3.4) – relevant in other languages? Since Chinese is the world's most widely spoken language, it is clear that any account of language that aspires to a claim of universality – including universals of word structure – must take the Chinese data into account. Chinese words have a story to tell about the degree to which words are susceptible to the algorithms of syntax, and whether there is a definition of *word* that works reasonably well across languages. Using Chinese to address these questions is bound to increase our understanding of universal word properties.

I will demonstrate how the structure I propose for Chinese words goes a long way toward explaining how these words have come to have the shape they now have, resulting in the present designation of Chinese as a language of 'compounds'. If we want to know how Chinese words evolved to take their present shape, it is important to understand how word components evolve to take on the identity they have, and how that identity shifts over time as new words are created and old ones discarded. It would be a mistake to overrely on contemporary data in addressing historical factors, but a good understanding of what is happening in the language now can offer a possible window into the past.

Another important issue this study addresses is the relationship between words and characters in Chinese. Time and again, when I tell people that I work in Chinese linguistics, I get a response like: ‘Oh, Chinese makes sentences by putting characters together, right?’, as if, unlike the rest of the world’s languages, Chinese enables spoken communication by the oral exchange of little visual icons. People for the most part do not really think that Chinese speech communication occurs via ‘characters’, but many *do* believe that the spoken language unit represented by the character – the morpheme – is the unit that is used to create and understand Chinese sentences. This may seem more reasonable than the notion of little visual icons flying through the air among speakers, but it is quite nearly as untenable, as we shall see in 7.2.

This widely accepted belief that the morpheme is the unit of spoken language lexical access has coloured the attitudes of many who work in the psycholinguistics of Chinese language processing. For this reason, Chinese language perception and production studies have tended to focus on properties of Chinese orthography.[1] Chinese orthography is valuable because its special characteristics enable us to ask questions about the nature of reading that cannot be asked using other orthographies. But if we want to gain insight into the psycholinguistic properties of Chinese we must also focus on the perception and production of spoken Chinese. To do that requires a precise description of Chinese words and their structure. Some who work in Chinese psycholinguistics assume that words in Chinese cannot be defined easily, or that the concept *word* is somehow not relevant for Chinese. But Chinese forms phrases and sentences as do all natural languages, by using rules of syntax to string together words that are retrieved from a mental lexicon. In order to investigate sentence processing in Chinese, we must be able to identify those words and have an understanding of their properties. Only then can we ask how the on-line natural language processing or the first- and second-language acquisition of spoken Chinese occurs.

[1] A notable exception to this is the work of Xiaolin Zhou and William Marslen-Wilson (e.g., Zhou and Marslen-Wilson 1994, 1995).

1.2 The scope of this work

This volume is a combination of descriptive and theoretical approaches. Following this introductory chapter, I provide criteria for identifying Chinese words in chapter 2, and in chapter 3 I explain why word structure is optimally described in terms of the form class identity of word components and how that may be accomplished. Then I offer a morphological analysis of Chinese words in chapter 4, followed by a universal ('X-bar') analysis in chapter 5 that abstracts the morphological properties of words over different form class categories. In chapter 6, I discuss the phenomenon of lexicalization, including why it explains how the relation between the gestalt word and its constituents varies, and why this is an important factor in understanding how Chinese words have evolved into their present form. The nature of the Chinese mental lexicon is discussed in chapter 7, including how lexical access occurs in speaking, hearing and reading Chinese. Finally, in chapter 8 I offer a summary and some concluding remarks.

The working hypothesis of this book is that the entity 'word' is a real cognitive construct that is also a linguistic primitive in natural language, and that word properties and word-forming algorithms like those proposed for Chinese arise due to universal principles and constraints that apply to all languages, serving to circumscribe the range of possible word types that may occur. This critically involves the notion of lexical primitives (X^{-0}, X^{-1} etc., see chapter 5),[2] the existence and combination of which I propose constitute the universal character of word structure. It is proposed that words in all human natural languages are analysable into these lexical primitives and their concatenation, subject to limited parametric variation.

I shall be referring in all cases to Mandarin Chinese, transcribed using the pinyin system of phonetic romanization and represented using simplified Chinese characters. Also, I'll be dealing for the most part with only two-syllable words. There are many words of three, four and more syllables in Chinese, but I feel better able to investigate

[2] For the purposes of this study, the terms X^{-0} and X^{0} (with negative and non-negative superscripts respectively) may be considered the same. I generally follow the convention of using negative superscripts for morphological objects as a notational device to distinguish them from syntactic objects.

the various aspects of word formation in depth by restricting the data base at present to words consisting of two syllables. To further restrict my data base, in this study I deal for the most part only with complex words formed from noun and verb elements.

I would like to thank for helpful comments or references (in more-or-less chronological order) Yingxing Yin, Joan Bybee, Isabel Wong, Michael Sawer, Dick Anderson, Bill Nagy, Yu-chiao Jade Longenecker, Yu Shen, Yabing Wang, Xiaolin Hu, Tianwei Xie, Carl Pollard, Jim Dew, Vivian Ling, Mike Wright, Taiyuan Tseng, Richard Sproat, Kevin Miller, Chiung-chu Wang, Gary Feng, Shiou-yuan Chen, Bob Good, Chih-ping Sobelman, Jerry Morgan, Georgia Green, Jennifer Cole, Dan Silverman, Hans Hock, Adele Goldberg, Elabbas Benmamoun, Chin Woo Kim, James Tai, Yung-li Chang, James Myers, Jane Tsai, Shou-hsin Teng, C-C. Cheng, Benjamin Tsou, Liejiong Xu, Derek Herforth, Marcus Taft, Xiaolin Zhou, Tongqiang Xu, Charles N. Li, Tsu-lin Mei, Elizabeth Traugott, Wen-yu Chiang, Yuancheng Tu, Si-qing Chen, David Chen, Yan Chen, Shenghang Huang, Yu-min Ku, Kazue Hara, Shu-fen Chen, Gary Dell, Carol Packard, Jose Hualde, Jenn-Yeu Chen, James Yoon, Victor Mair and Stanley Starosta. I would especially like to thank my friend Shengli Feng, two anonymous Cambridge University Press reviewers and two additional anonymous reviewers for giving me valuable detailed feedback on draft versions of the manuscript. Special thanks also to Alain Peyraube for detailed comments on the manuscript and for many valuable references to complex word formation in earlier stages of the Chinese language. Thanks also to Christine Bartels and Kate Brett for having faith in my work, to Citi Potts for excellent copy editing, and to Barbara Cohen for making the index. I would like to thank the University of Illinois at Urbana-Champaign for granting the sabbatical leave allowing me to work on this book, and the UIUC Research Board for awarding the grant that enabled me to complete the project. Finally, I want to thank my fellow family members Carol, Errol, Sam and Eric, whose patience as I worked on this book was always appreciated (though it may not have seemed so at times), and whose dinner conversations have provided an endless font of linguistic and conceptual creativity as well as comic relief.

As the reader goes through this work, in many places it will become evident that I have remained overly simplistic, choosing to sidestep many questions of interest. In some cases I have remained at that level intentionally, because to do otherwise would have resulted in great delays as I tackled problems of detail, and also because the resulting exposition has allowed me to make the points and address the issues I wish to focus on. There are also likely to be logical lacunae and analytical abysses in the interplay of ideas that I have forged in putting this work together. I invite the reader to point these out, and to offer suggestions and criticism.

2 | Defining the word in Chinese

2.1 What *is* a 'word'?: different views

For speakers of some languages, the 'word' is a robustly intuitive notion. But it seems that no matter what the language, we have a hard time providing an exact definition that encompasses all and only those entities that our intuition tells us are words (see, e.g., Anderson 1985b: 153–4). This means that the concept 'word' is nothing if not elusive, and suggests that perhaps there is no concept of word that is universally applicable. Indeed, if there is no cross-linguistic, or universal psycholinguistic evidence for the existence of the word, then we may well doubt the validity of the word as a primitive natural language construct. It could a priori be the case that there is really no such thing in absolute terms as the 'word', and that it is just an artifactual linguistic construct that happens to coincide with salient units intermediate between morphemes and phrases that happen to appear in many of the world's languages.

There is another reason why the possibility that the 'word' is a derived rather than primitive construct may occur to us: words are definable using several disparate linguistic criteria. For some of these criteria considered in isolation, the label 'word' seems strangely inappropriate, since words so defined seem overly abstract, with nothing very 'word-like' about them. Let us take a look at these criteria to see if any of them are closer than others in providing an accurate portrayal of 'word'.

2.1.1 Orthographic word

Probably the most popular conception of the word (especially in languages such as English) is that of the 'orthographic word', that is, the word as defined by writing conventions. It is easy for an English speaker (or a pigeon, for that matter) to segment a written English text into words strictly by the visual appearance of the text, i.e., by picking out the written material that occurs between the spaces. Speakers of English therefore have a strong 'intuition' as to what is and is not a word in spoken language, partly as an effect learned through experience with orthography: in producing written English the speaker/writer must put the spaces in their proper place. This, of course, raises the question of what criteria are used to decide where the spaces go in the first place. It turns out that the criterion that is closest to the orthographic word in English is remarkably close to that of the 'syntactic word' (see 2.1.7 below).

In deciding for the purposes of this study what are words in Chinese, we could safely eliminate the orthographic word for reasons having little to do with Chinese per se – namely, that orthographic words are usually defined using non-orthographic criteria. That is, items are usually selected for membership in the 'orthographic word' category based upon linguistic properties other than the nature of the orthography. In any case, the orthographic word has no relevance specifically for Chinese, since Chinese orthography segments written texts into characters, which generally represent morphemes rather than 'words'.[1]

2.1.2 Sociological word

The term 'sociological word' may be attributed to Chao (1968: 136), and describes a concept that native speakers use to refer to linguistic units of a certain size. Chao defines it as 'that type of unit, intermediate in size between a phoneme and a sentence, which the general, non-linguistic public is conscious of, talks about, has an everyday term for, and is practically concerned with in various ways' (Chao 1968: 136–8). The sociological word is the familiar 'word' in English, and in Chinese, it is the *zì* 字, meaning either the Chinese written character or the Chinese spoken morpheme. The concept of the sociological word will be further discussed in 2.2.

2.1.3 Lexical word

Another common conception of 'word' we might call the *lexical word* (termed the *listeme* by Di Sciullo and Williams 1987: 1), which incorporates the 'listedness' characteristic of lexical items. That is, the lexicon is traditionally seen as that component of the grammar that contains

[1] Of course the orthographic definition of 'word' does work, albeit tautologically, for romanized Chinese, since in romanized Chinese the goal is generally to put spaces between words rather than between morphemes.

all that is not predictable, and must therefore be stored in a memorized list. To that extent, 'words' are those idiosyncratic, arbitrary pairings of sound and meaning that cannot be generated by rule 'on line' that we file away in memory for use in the performance of a speech act.

The 'listedness' criterion is neither sufficient nor necessary to define 'word', because it is common to have both 'listed' items that are not words (e.g., idiomatic phrases or 'listed syntactic objects', Di Sciullo and Williams 1987: 5) and words that are not 'listed' (e.g., large numbers of complex words in languages such as Turkish or Italian that are productively constructed using members of affixation paradigms, and are not likely to be stored away as 'listemes'). The concept of the lexical word is popular because it most closely comports with the idea of 'listing as a dictionary entry' that is popularly taken to be a defining criterion for 'word', and because it overlaps almost completely with the orthographic word discussed above.

The lexical definition of 'word' is not useful as a defining concept in our investigation of Chinese for just this reason: the 'listedness' criterion fails to include many Chinese words created by rule (see 7.4.1) and improperly includes many things approximating Di Sciullo and Williams' 'listed syntactic objects'. So while it will be interesting to keep this notion in mind – especially when it comes to the time to consider the structure of the Chinese lexicon – for the time being we will set aside the concept of lexical word.

2.1.4 Semantic word

A definition using semantic criteria is one of the most traditional ways of characterizing the notion of 'word'. The *semantic word* is sometimes equated with the idea of a 'unitary concept'. Sapir (1921/1949: 25) portrayed the word as 'the outward sign of a specific idea, whether of a single concept or image or of a number of such concepts or images definitely connected into a whole'. Baxter and Sagart (1997; citing Dowty, Wall and Peters 1981) characterize the semantic word as the 'basic expression' of formal semantics, a form with a semantic value such that such expressions may combine to form complex expressions, but may not be further decomposed into subexpressions (Baxter and Sagart 1997).

The semantic definition of word is one that strongly appeals to intuition – many people probably feel they have an idea of what a 'basic concept' might be, even if it is not uniquely definable either within or among speakers. However, the notion of semantic word is only minimally useful, because reducing concepts to their semantic primitives is a notoriously difficult exercise. Even if it were possible to come up with a list of such semantic primitives, examining them independently of their phonological form actually gets us no closer to defining 'word', since the concept of 'word' crucially requires reference to phonetic form. And once we relate those semantic primitives to phonological forms, what we get is a minimal pairing of form and meaning – an entity that is closer to the traditional morpheme than to the word.[2]

2.1.5 Phonological word

The *phonological word* is a 'word-sized' entity that is defined using phonological criteria. Chao (1968: 153–4) considers the existence of potential pauses – the places in a sentence where it is possible to pause naturally – to be a phonological criterion for the definition of word boundaries in Chinese (for a more general application of the concept, see Anderson 1985b: 150–2). But 'word' as defined by the phonological criterion of potential pause turns out to be of little use, since, like the orthographic and lexical definitions of 'word', this criterion turns out largely to be based upon other (i.e., syntactic, morphological or prosodic phonological) criteria. That is, the reason 'pauses' cannot go where a speaker feels it is inappropriate to place them is because their placement would violate the constituency of a syntactic, a morphological or a (otherwise defined) phonological word.

More recently the phonological definition of word has been based upon the domain of phonological rule application, or the output of a phonological rule. Dai (1997) gives examples of phonological word boundaries, defined by the application of a phonological rule. Baxter and Sagart (1997) give examples of accent (Czech) and sandhi (Sanskrit) phenomena, as well as stress units in Swahili, Polish and ancient Greek conditioned by independently defined word boundaries. The phonological word has also been characterized in prosodic terms, with Duanmu (1997) using phonological tone and stress evidence to distinguish words and phrases in modern Chinese.

[2] Thus we do refer to semantics when defining the morpheme, and make use of semantic criteria when we discuss the concept of 'semantic head' in 5.5.2.