

1

Preliminaries

Chu-Ren Huang and Dingxu Shi

A grammar is the system of knowledge of the relation between what people do and what people know when they use a particular language. Since what people know in the context of their language use is often implicit, linguistic theories are proposed as a foundational hypothesis to enable the explicit explanation of a grammar of any particular language. This presents an underlying dilemma in the writing of any grammar. On the one hand, descriptive work is the foundation of any scientific study and is crucial to the language sciences. Modern linguistics emerged as a result of a conscientious effort to move from prescriptive to descriptive studies of language. On the other hand, once any theoretical framework or account is adopted, a grammar becomes prescriptive in the sense that it imposes a set of conceptual primitives and structure of rules prescribed by a sub-set of linguists. How to capture the system of implicit knowledge without prescribing an *a priori* theoretical framework remains the biggest challenge to any descriptive grammar.

A Reference Grammar of Chinese meets this challenge with an empirical approach focused on describing what people do when they use Mandarin Chinese, while allowing generalizations to emerge from our descriptions as well as from the readers' observation of the data. We believe that a keenly observed description of the generalizations and tendencies based on the observation of the extensive data of language use will lead to capturing the implicit knowledge people share without prescribing an explicit rule. To achieve this goal, corpora and Web-extracted examples are used extensively, with an occasional supplementation of made-up sentences. These data have been carefully examined by our authors for their distributional patterns and tendencies. None of the examples cited in this grammar are single instances of language use; rather, they were chosen as an illustrative representation based on a set of similar examples selected by the authors. In other words, this reference grammar is intended to be read like a guide to the Chinese language, mediated by an extensive set of extracted examples for each grammatical point we make. Readers can consult the example database when they read the grammar to strengthen both their understanding of the generalizations and the complexity of language in use.

This grammar assumes a minimal set of theoretical concepts, which include grammatical categories, basic grammatical functions, and some intuitive semantic concepts, such as the thematic roles of agent, theme, goal, etc. A more detailed discussion of the grammatical categories is presented in Chapter 2.

1.1. The Chinese language

Chinese, or Mandarin Chinese, has the most native speakers in the world as well as one of the longest cultural heritages. Mandarin Chinese also has become one of the most learned foreign languages in the world. The 2005 version of *The Language Situation in China* claimed that more than 30 million people in the world were learning Chinese as a foreign language. The need for a linguistically felicitous and accessible reference grammar of the Chinese language is clear and urgent. Authored by leading Chinese linguists in each topic area, this volume serves as a comprehensive and accessible reference grammar of Chinese in that it aims to cover all the important linguistic facts of the language; moreover, these facts are presented in a way that does not presuppose knowledge of a particular linguistic theory or grammar of Chinese.

This grammar provides a synchronic, descriptive grammar of present-day Standard Mandarin Chinese. It shares some of the major design philosophy with that of *The Cambridge Grammar of the English Language* (Huddleston and Pullum 2002).

1.1.1. Standard Mandarin Chinese

Standard Mandarin Chinese has the phonological system of the Beijing variant of Northern Mandarin as its norm of pronunciation. Historically, Mandarin (官话 *guan1hua4*) has been the common language adopted by officials, yet it developed into different variants among the areas where it was primarily spoken, such as Northern Mandarin, Southern Mandarin, and Sichuan Mandarin. In common English usage today, however, Mandarin and Chinese are used interchangeably and loosely to refer to Standard Chinese. The Chinese described in this reference grammar refers to the Standard Mandarin Chinese that is generally accepted in a wide range of public discourse, such as government, education, broadcasting, and publishing. This standard language is referred to as Putonghua (普通话 *pu3tong1hua4* ‘common language’) in Mainland China, Singapore, and Macau, and as GuoYu (国语 *guo2yu3* ‘national language’) in Taiwan, while both terms are used in Hong Kong. Broadly speaking, Putonghua follows Mainland China conventions and GuoYu follows Taiwan conventions, and they do differ from each other occasionally, not unlike the contrast between UK and US English. While focusing on the widely accepted usages as reflected in standard written and spoken Chinese, and on the common usages shared by Putonghua and GuoYu, we will point out significant distinctions when necessary.

1.1.2. Synchronic description of present-day Chinese

The earliest record of a well-developed system of Chinese writing dates back to more than three thousand years ago. Although linguists do not agree on all the details, the history of the Chinese language can be divided into four stages: 上古汉语 *shang4gu3han4yu3* ‘Old Chinese’ (Shang Dynasty to Han Dynasty, sixteenth century BCE–220 CE), 中古汉语 *zhong1gu3han4yu3* ‘Middle Chinese’ (Southern and Northern Dynasties to Five Dynasties and Ten Kingdoms, CE 220–960), 近代汉语 *jin4dai4han4yu3* ‘Early Modern Chinese’ (Song Dynasty to May Fourth Movement, 960–1919), and 现代汉语 *xian4dai4han4yu3* ‘Modern Chinese’ (1919–present). Throughout these periods, substantial changes took place in all linguistic aspects of Chinese. In terms of grammar, Old Chinese can be identified by the lack of more complex constructions, which developed later. Middle Chinese is the period when many new constructions and forms emerged, including the 把/将 *ba3/jiang1* constructions, the 被 *bei4* passive, and a pronoun system that is very similar to that of present-day Chinese. During this period, Chinese prepositional phrases also moved from the predominantly post-verbal position to the pre-verbal position. In Early Modern Chinese, aspectual markers such as 了 *le0* and 着 *zhe0* and phrasal suffixes such as 的 *de0* and 地 *de0* were widely used. Writing based on vernacular Chinese (白话文 *bai2hua4wen2*) emerged in the Tang Dynasty, but Classical Chinese (文言文 *wen2yan2wen2*) was still used in formal writing until Modern Chinese, particularly after the May Fourth Movement in 1919, when most publications in China started to use the vernacular language.

The historical change of Chinese is of great linguistic importance and interest, but as a synchronic grammar, this volume limits the description to present-day Modern Chinese, especially the language since 1991, because all the generalizations of this grammar are based on corpora with natural Chinese data collected from that time. It is important to bear in mind, however, that conventionalized historical forms are still used in Modern Mandarin Chinese, especially in formal registers. As such, they are part of present-day Chinese and will be covered in this grammar.

1.1.3. Varieties of Chinese

The term World Chineses (全球华语 *quan2qiu2hua2yu3*), though not as common as World Englishes, is becoming more and more widely used with the increasing popularity of Chinese as a second language and with the Chinese diaspora spreading and growing. Despite the same linguistic heritage, Mandarin Chinese in different regions has evolved in different ways as a result of the political, economic, cultural, and social development of each region. Variations can be found in pronunciation, lexicon, and syntax. While it is important to investigate these

differences, this reference grammar aims to present the shared core of grammar. In the rare cases where the variations render the description of a shared core difficult or if the variations present a special challenge to learners and speakers, observations and descriptions will be provided.

The varieties of Chinese also differ orthographically in adopting traditional characters (Taiwan, Hong Kong, and Macau) or simplified characters (Mainland and Singapore). Orthography projects language identity through the formation of a community of practice and may introduce language variations, although strictly speaking it is not part of the grammar. In this grammar, we adopted simplified characters as the most commonly learned system. It is, however, important to note that although a meaning-preserving mapping from traditional characters to simplified characters can be performed without ambiguity, the same cannot be said for mapping simplified characters to traditional characters.

1.1.4. Chinese dialects vs. Sinitic languages

“Chinese” as the language spoken by ethnic Han people is traditionally divided into seven major groups: Mandarin (or Northern Chinese), Wu, Xiang, Gan, Kejia (Hakka), Yue (Cantonese), and Min. Speakers of different groups are mutually unintelligible in terms of speaking, although they share the same written language and the grammar of written Chinese for each group does not differ substantially from that of Standard Chinese. Such facts bring about the question of whether they should be referred to as dialects or as languages (i.e. Sinitic languages), a linguistic issue with strong cultural, political, and societal implications. Since this grammar concentrates on present-day Standard Chinese, when references to other varieties is necessary, only the language/dialect name will be used, without explicit reference to its language/dialect status.

1.1.5. Descriptive account

The descriptive account of this grammar is succinct and theory-neutral, based on corpus observation and reflecting how the language is actually used. For non-standard or ungrammatical usages, the grammar reports that the usages are rarely or not found in the corpora, rather than providing created examples in contrast to the standard and grammatical examples.

1.1.6. Grammar

We agree with Pullum and Huddleston (2002: 4) that a grammar is divisible into syntax and morphology; the former is concerned with the way words combine to form phrases, clauses, and sentences, while the latter is concerned with the formation of words. Because a word plays a prominent role as a basic unit at

different levels, this grammar pays attention to the definition and identification of words, in addition to other important linguistic facts of the language.

This grammar includes a chapter on classifiers (Chapter 7), a grammatical category less familiar in grammars of English and other Western languages. Classifiers are essential components of noun phrases in Chinese and they represent important selectional and semantic information. In addition, following the conventions established in Chao (1968), this grammar assigns the semantic correspondents of adjectives in English and many other languages, such as ‘small’ (小 *xiao3*) and ‘quite’ (安静 *an1jing4*), as well as the category of (state) verbs, and reserves the category of adjectives for the non-predicative types, such as 超级 *chao1ji2* ‘super’ (see Chapter 10).

1.2. A data-driven and corpus-based reference grammar

While more and more linguistic research and reference grammars have adopted corpus-based empirical approaches, this grammar takes a further step by being both data-driven and corpus-based. The developments in the past thirty years have made it possible for people to access and extract generalizations from corpora. The large-scale corpora accessible for this grammar include the POS-tagged Chinese Gigaword Corpus (1,400 million characters from Mainland China, Taiwan, and Singapore; data collected during 1991 to 2004, Huang 2009) and the manually tagged Sinica Corpus (10 million words collected primarily since 1996, Chen et al. 1996). The availability of large-scale corpora also enables the use of computational tools, such as Word Sketch Engine, to extract grammatical information directly from the corpora. For this grammar, the authors’ expertise and judgments have been greatly enhanced by their access to both the Sinica Corpus and the 2nd edition of the Chinese Gigaword Corpus, through the corpus interface of Chinese Word Sketch (Huang et al. 2005). Most of the examples in this grammar, with very few exceptions, have been carefully selected from the corpora with minimal modification. In this sense, this grammar is the first Chinese grammar written based on corpus data. It is also among the first such reference grammars in the world.

As a reference grammar, our emphasis is to get the facts and generalizations right, especially those facts or generalizations missed or mischaracterized by previous grammars. This goal was achieved through a two-pronged empirical approach. First, the authors had access to the largest available Chinese corpora as well as the most powerful corpus interface. In addition, they were encouraged to consult the Web through Google when in doubt. This ensured that the widest range of language data was accessed. Second, each chapter was drafted by a designated author(s) who has done extensive work on the topic area. After the initial draft, the chapter was presented and discussed at authors’ workshops, and each chapter

6 *Chu-Ren Huang and Dingxu Shi*

also underwent extensive comments and review by at least two other experts. Lastly, each chapter was reviewed and revised by the two chief editors. In sum, each and every chapter reflects the collective research knowledge of four or more leading Chinese linguists in the field, each contributing to the consistency and comprehensive coverage of the facts.

This reference grammar is anchored by illustrative example sentences. All of these corpus-extracted realistic examples are presented in the standard four-line format: the first line consists of the text in Chinese characters; the second line consists of a word-for-word Pinyin transcription; the third line is aligned with the second line to provide a gloss; and the last line provides faithful free translation. The example [1] is the sentence [2b] from Chapter 4.

- [1] 有空的时候,到公园里去走一走,呼吸呼吸新鲜空气。
 you3kong4 de0 shi2hou0 dao4 gong1yuan2 li3 qu4
 have_time DE when go_to park in go
 zou3yi1zou3 hu1xi1hu1xi1 xin1xian1 kong1qi4
 walk_a_walk breathe_breathe fresh air
 ‘When you have some time, have a walk in the park and breathe some fresh air.’

Following linguistic conventions, the example sentences are discussed and grammatical information is explicated in the text immediately before or after the example given. The above example shows that we have ordered the examples in each chapter according to their order of appearance, and use a, b, c, etc. to differentiate a group of similar sentences given under the same example number.

This grammar will be accompanied by a periodically updated online example database in order to supplement the examples of the grammar and to add value to the restriction of the finite number of printed pages. The original database was constructed together with the grammar, when at least twenty example sentences were selected for each linguistic topic described in the grammar. Each sentence is not only annotated with the topic for which the sentence is selected, but also annotated and indexed with all other relevant linguistic topics covered in the grammar.

This reference grammar aims to make the underlying set of linguistic facts from which we built our generalizations sharable with others who may construct a parallel reference grammar with different design criteria. Therefore, in addition to the example database, a citation database was constructed based on the topics of the grammar. The periodically updated database consists of all the bibliography used for this grammar and all the articles from the major journals of Chinese linguistics, including 中国语文 *zhong1guo2yu3wen2* ‘Chinese Language and Writing’

and *Journal of Chinese Linguistics*. With such a database, this reference grammar will remain relevant with respect to the most updated research topics, even if the printed version is not updated as frequently.

1.3. Chinese writing system

The Chinese writing system is the longest continuously used system in the world. The system is composed of characters (汉字 *han4zi4*, kanji in Japanese as well as in common English translation), which are logographic symbols encoding both phonetic and semantic information. Unlike phonological writing systems, each Chinese character is grounded with some conceptual knowledge information that was conventionalized at the time the character was created. Furthermore, the writing system is considered a cultural symbol that unifies the Chinese people speaking mutually unintelligible varieties of Chinese. By this design, the Chinese writing system is not as arbitrary as phonological writing systems, which are common among other languages in the world. This non-arbitrariness, in turn, has allowed the Chinese writing system to reflect more about the grammar of Chinese; hence, some discussion of the writing system in this reference grammar is required.

It is important to debunk the myth that the Chinese writing system consists of Chinese characters only. This may have been the case as recently as fifty years ago; however, most contemporary Chinese dictionaries nowadays include a few hundred so-called alphabetic words (字母词 *zi4mu3ci2*). These alphabetic words are *bona fide* entries in the lexicons of modern Mandarin Chinese, with full grammatical functions in the category to which they belong, as discussed in Chapter 3. These words can be composed of all alphabets (Q *kyu* ‘to have sustainable good texture when chewed on’; HSK *eich-es-kei* Hanyu Shuiping Kaoshi/‘Mandarin Standard Test’; IBM *ai-bi-emu* ‘International Business Machine’), or start with one or more alphabets but end with a character (AA 制 *ei-ei-zhi4* ‘to go Dutch’; K 书 *kei-shu1* ‘to hit the books hard’), or start with a character but end with an alphabet (阿 Q *a1-kyu* ‘a fatalist, a fictional protagonist in Lu Xun’s novel’; 卡拉 OK *ka3la1-ou-kei* ‘karaoke’). Unlike Japanese katakana, these alphabetic words are not restricted to loanwords, although many have loanword origins (especially those referring to new technology or brand names). However, the blended translation alternative remains a desirable alternative, so many loanwords are actually represented by Chinese characters (such as 可乐 *ke3le4* ‘able-enjoy cola’; 爱疯 *ai4feng1* ‘love-crazy iPhone’). It is worth noting that many alphabetic words originate from either Pinyin-based abbreviations or vivid imitations of/associations with the sounds/shapes of the alphabets. Alphabetic words also have a unique linguistic feature in that they do not conform to the phonological integrity of Chinese. The alphabetic parts of the alphabetic words are typically pronounced without an assigned tone, and many





















of them represent syllables which are not part of the Chinese phonological repertoire, such as *emu* (from IBM), *kei* (from K 书), and *kyu* (from Q, 阿 Q). They simply represent a conventionalized way to pronounce these alphabets. These alphabetic words can be a noun, a verb, or an adjective, and the alphabet-plus-character template seems to be the most productive pattern of neologism.

1.3.1. A brief history of Chinese script

The Chinese character script has been continuously used for more than 3,000 years. The oracle bone script (甲骨文 *jia3gu3wen2*) of the Shang Dynasty (1600–1046 BCE) is the earliest surviving evidence of a well-developed writing system of Chinese and is directly related to the subsequent Chinese scripts. Oracle bone inscriptions are found mainly on turtle shells (甲 *jia3*) or ox and other large animal bones (骨 *gu3*), hence 甲骨文 (*jia3gu3wen2* ‘shell-bone-writing’). These inscriptions are the records of the answers from the divinatory practice of the royal family communicating with their ancestral spirits. Evolving from the oracle bone script, bronze inscription script (金文 *jin1wen2*) is found on ritual bronze vessels of the late Shang Dynasty to Zhou Dynasty (1100–403 BCE). After the Zhou Dynasty, the writing systems of different parts of China diverged until the Qin Dynasty (221–207 BCE) unified China and set the Qin variant of seal script (篆书 *zhuan4shu1*) as the national standard script. For easier and faster writing in government bureaucracy, clerical script (隶书 *li4shu1*) was adopted and it gradually replaced seal script in the Han Dynasty (206 BCE–220 CE). Clerical script is structurally and rectilinearly very similar to the modern Chinese scripts and is thus considered the ancestor of modern scripts. Clerical script then evolved into standard script (楷书 *kai3shu1*) and was replaced by standard script during the Southern and Northern Song Dynasties (420–589 CE). Since then, standard script has been used as the standard form of orthography in China.

Two crucial observations can be made based on the four example characters shown in Table 1.1 below. First, many characters are decomposed into components (such as 明 *ming2*, which consists of the two components 日 *ri4* ‘sun’ and 月 *yue4* ‘moon’), and these components may or may not be characters themselves. This point will be addressed in more detail in the next section. Second, all of the different historical scripts are variants, and the internal structure of each character is largely preserved between script changes. For example, 明 *ming2* ‘bright’ is composed of the two components 日 *ri4* ‘sun’ to the left and 月 *yue4* ‘moon’ to the right. Regardless of how much the graph representing the character changes, both the composition and the left–right structure of these two components remain the same. The consistency of this component composition relation can also be observed to have occurred over time for 渔 *yu2* ‘to fish,’ as in Table 1.1. Furthermore, even for non-decomposable characters, such as 水 *shui3* ‘water’ and

Table 1.1 Evolution of Chinese scripts¹

	'water'	'to fish'	'up'	'bright'
Oracle-bone script				
Bronze script				
Small seal script				
Clerical script				
Standard script				

¹ The Chinese scripts were extracted from Academia Sinica's Database of Chinese Characters Composition (漢字構形資料庫) at <http://cdp.sinica.edu.tw/cdphanzi>

上 *shang*⁴ 'up,' it is possible to see that the internal structure of the critical components of the graph stays the same. It is with this consistency of the internal component composition relation that we can show that the Chinese writing system is a single, continuously used system. Any historical text, regardless of the style of script it is in, can be directly mapped to any other script to be read.

1.3.2. Structure of Chinese characters

A character is the smallest meaningful unit of the writing system in Chinese, compared with a morpheme, which is the smallest meaningful unit, and a word, which is the smallest unit with independent syntactic functions. In contrast to phonographic languages, such as English, that are mainly composed of symbols that encode phonetic values only, the character-based writing system of Chinese is featured as logographic in that it is mainly composed of logographic symbols that encode both phonetic and semantic values. Phonetically, each Chinese character represents a syllable, compared to English where a letter or a group of letters represents a phoneme. Semantically, a Chinese character usually encodes a lexical concept, which allows it to stand for the same (or similar) meaning regardless of language changes and variations.

The misconception that Chinese characters cannot be learned without rote memory covering the stroke-by-stroke order of all the strokes of a character has both added to the notoriety of Chinese (and to the myth of the complexity of Japanese, since it uses kanji as one of its complex writing systems) and lent

support to the proposal to convert Chinese to an alphabetic writing system. However, studies have shown that Chinese characters are composed of components (部件 *bu4jian4*). Each component can in turn be composed of smaller components or, eventually, a fixed number and order of strokes. What this means is that recognizing and writing a character only requires knowledge of the components of a character as well as how these components are put together once the basic components are known. There is also a general rule of the order of the composition of left-first, before top-first, and outside-in when other rules do not apply. For instance, the character 明 *ming2* ‘bright’ is formed by the composition of 日 (on the left) first, and 月 second. The character 盟 *meng2* ‘alliance’ starts with the same 日+月 sequence, with the third component 皿 *min3* ‘basin’ last and at the bottom. The character 萌 *meng2* ‘to sprout’ is formed with the grass radical 艹 *cao3* on top, followed by the same 日+月 sequence. These component sequences are largely preserved through the evolution of different scripts (including most cases of simplified characters) and even apply to some regional glyph variants. For instance, 峰 and 峯 *feng1* ‘peak’ are variants of the same character and they can be described by the same component composition rule of 山+峯 *shan1 + feng1*, except that one variant follows the left–right order while the other follows the top–down order.

A Chinese character is not only formally composed of components, but its formal composition also follows rules of internal composition. 说文解字 (*shuo1wen2jie3zi4*, 121 BCE, literally, *Explanations of simple graphs and analyses of composite graphs*) compiled by the Eastern Han scholar Xu Shen was the first comprehensive dictionary to analyze the structure of Chinese characters. Xu Shen proposed six principles of Chinese character composition, of which four are firmly established in modern philology: pictographic (象形 *xiang4xing2*), ideographic (指事 *zhi3shi4*), semantic–semantic composition (会意 *hui4yi4*), and semantic–phonetic composition (形声 *xing2sheng1*). Pictographic characters such as 日 *ri4* ‘sun’ and 月 *yue4* ‘moon’ resemble the objects in the physical world. Ideographic characters such as 上 *shang4* ‘up’ and 下 *xia4* ‘down’ represent abstract ideas. A semantic–semantic compound is typically composed of two or three pictographic or ideographic characters and encodes a combination of the meanings of the characters. For example, 明 *ming2* ‘bright’ is a combination of the pictographic 日 *ri4* ‘sun’ and 月 *yue4* ‘moon,’ while 森 *sen1* ‘forest’ is composed of three 木 *mu4* ‘tree,’ which by itself is a pictographic character. Semantic–phonetic compounds typically consist of a phonetic unit and a semantic unit. For instance, 妈 *ma1* ‘mother’ is composed of the radical 女 *nu3* ‘woman’ and the phonetic 马 *ma3* ‘horse,’ representing the phonetic part, suggesting the sound of 妈 when the character was created. The semantic–phonetic composition is considered to be the most frequently used principle, estimated to represent more than 90 percent of the characters; moreover, the radical–phonetic