2. Chinese Language Basics

This chapter covers some basics of the Chinese language. Most of the information is closely related to Chinese text reading about which you should become familiar. A few topics may not be very relevant for our purposes. They are briefly mentioned because the terminologies and concepts may appear in the documentation of the tools we use.

2.1 Characters (字), Words (词) and Sentences (句)

Chinese is very different from western (Latin) style languages. First of all, the text in Chinese does not consist of alphabet letters; instead it is made up of characters (字). The character is the most basic element in the Chinese language. It is built from strokes written in a virtual rectangle box. Commonly used characters contain from one to twenty, sometimes thirty, strokes. Figure 2.1 shows some examples of strokes and characters containing these strokes.

Stroke	丶	一	丨	丿	㇏	㇆	㇇	㇖	㇟	㇂	㇀	㇜	㇛	㇙
Example Characters	小	素	上	分	人	羽	又	冠	能	代	化	去	女	以
	心	味	山	秒	定	扇	多	軍	見	我	水	公	災	切
	下	平	伐	必	勝	綸	子	完	尾	成	求	云	巢	比
	雨	生	木	爭	天	巾	孫	蛋	巴	戈	北	叁	好	鼠

Figure 2.1 Strokes and characters

Each character has its meanings, which can represent a thing or a concept. The sound of a character contains one syllable, and the pronunciation can be very different when spoken by people from different geographic regions. The Mandarin (国语), also known as Putunghua (普通话), is the pronunciation used by people from Beijing. It is the official standard pronunciation recognized by the governments from both the People's Republic of China (PRC) and Taiwan, the

Republic of China (ROC). Mandarin is estimated to be spoken by about 55% of the whole Chinese language population, and over two thirds of the population in the cities. Most Chinese whose primary language is not Mandarin can nevertheless understand it.

The Chinese character set is known as Han character or Hanzi (汉字) by Japanese and Koreans. Many Chinese characters have been adopted by the Japanese and Korean languages and given the same meanings with different pronunciations. In Japanese and Korean, a person's name is usually written in Hanzi.

It is hard to say how many Chinese characters there are in total. Generally, there are between 5,000 to 15,000 characters in standard dictionaries that people use regularly. The famous 康熙字典 (*Kangxi dictionary*) contains 47,035 characters. The most comprehensive Chinese dictionary that I know of, 中华字海 (*A sea of Chinese Characters*), even lists in excess of 85,000! Fortunately, most of these characters are rarely used; otherwise, people would have to spend all their lives learning the language. The most frequently used 1,000 characters cover about 90% of the usage. If you know 3,500 characters, the estimated number of characters an educated Chinese adult knows, the coverage rate reaches 99.5%.

The next level of elements in the language hierarchy above the 字 is the 词 (*word*). A general definition of the Chinese word is a combination of mostly two or more characters representing a specific thing or concept. Most characters have meanings but they may represent only a vague thing or concept, which is sometimes too ambiguous when used alone. When two characters are put together to form words, their meanings become more specific. Take the following character and words for example: the character 近 means near, close, or recent. When it is used together with another character, they become words with a more specific meaning.

近来 (*recently*), 近代 (*modern era*), 靠近 (*come close*), 近邻 (*neighbor*), 近视 (*near sighted*)

It is even harder to estimate how many total words there are in the Chinese language. A daily use dictionary may contain from 20,000 to 150,000 entries; the most comprehensive one that I know of, 汉语大词典 (*Comprehensive Dictionary of Chinese Words*), has a collection of 370,000! The word should be considered a critical if not the most important part of the Chinese language. It also plays an absolutely crucial role in reading. Even though functionally the Chinese word is very similar to the English word, there is one major difference. Unlike English or any other Latin language, there are no spaces in Chinese text to break different

words apart. Each sentence only consists of characters that concatenate together, with no indication of which characters form a word. It is the sole responsibility of the reader to identify all the words inside the sentence to interpret the meaning of the text. For an MT (Machine Translation) program to get good translation results, it is critical to have a dictionary that contains all the words people commonly use. If an MT program lacks a word in its dictionary, it would have to interpret that word as two or more separate characters. This may work well sometimes, but usually it doesn't. The same situation may even occur for human readers. Sometimes a writer uses a word that is not commonly seen, or is only known to people from a specific geographic area. Readers from other places will have a hard time comprehending the word, even if they know all its constituent characters.

Let us use the sentence "美国东接大西洋。" for example. It is to be interpreted as: 美国 (*USA*) 东 (*east*) 接 (*connect to*) 大西洋 (*Atlantic*). The readers need to know that 美国 means "the USA," and not the character-by-character interpretation of 美 (*beautiful*) and 国 (*country*). They also need to know that 大西洋 means "the Atlantic ocean," instead of 大 (*big*) 西洋 (*occident*), or 大 (*big*) 西 (*west*) 洋 (*ocean*).

Some words can have multiple meanings; sometimes the meanings are completely unrelated. It is up to the reader to determine what the word means under each circumstance. For example, 大班 means the top class in a kindergarten, and it also means the manager in a dance hall. 小時 stands for an hour, or it can mean in one's young age. 出口 can mean export or an exit. In circumstances like these, readers need to look at the context to determine what the words mean.

Another element in the language is the 成语 (*idiom*), which usually consists of four characters. The English equivalent of 成语 is an idiom or a phrase. It is similar to a word but it is longer and in most cases is used to describe a specific situation. One unique thing about 成语 is that most 成语 have stories behind them. For the purpose of Chinese reading, you can treat them the same as regular words.

One other element that is in the same hierarchy with word is what I call "pattern." I do not have a linguistic name for it. These are characters that are used together, based on certain rules to form words, but actually they are not "words." For example, consider the modification of the following words when the character 们 is added:

我	I
你	you (singular)
老鼠	mouse
我**们**	we
你**们**	you (plural)
老鼠**们**	mice

Each word refers to a living being. The extra character has made it into a plural.

Similarly, the following examples show that adding the character 的 converts a noun into its possessive form (this rule applies to inanimate as well as living objects):

我	I
你	you
我们	we
老鼠们	mice
北京园	Beijing Garden (a restaurant)
我**的 ~**	my ~
你**的 ~**	your ~
我们**的 ~**	our ~
老鼠们**的 ~**	~ of a group of mice
北京园**的 ~**	~ from Beijing Garden

In order to identify these patterns in sentences, the readers (or the MT programs) have to have some kind of intelligence to follow the rules. We can't simply use a dictionary to list all the words in such a situation, because there will be an infinite number of words of this form.

The next level of element above "word" in the hierarchy is the 句 (*sentence*). It is equivalent to the sentence in English with a very loose grammatical rule. Basically, a Chinese sentence consists of words and characters concatenated together. The subject can usually be implied and omitted in a sentence. In many cases, a sentence does not contain a verb. For example, people always say: "晚上很冷。" (*Night very cold*).

Unlike English, there is no capitalization to help us identify proper nouns in a Chinese text. To make things worse, it allows just about any character for the names of people, places, business, etc. Because of this, sometimes it can be a challenge to distinguish a person's name from other part of the sentence.

Remember I said that each character is always pronounced as one syllable. A sentence is made by putting multiple characters together, and the way to speak the sentence is just to concatenate all the single syllable sounds. From what I can see, this is about the only place where the Chinese language is easier to learn than Latin-based languages.

Chinese is an old language and over time people have used it in different ways. Writers like to be free, creative and do not want to obey rules or follow conventions. They also tend to be lazy and dislike the idea of doing extra work to make their writings more understandable. Perhaps writers consider writing to be the work of geniuses. If readers cannot follow the great concepts behind their masterpieces, then the readers must not be smart enough. Some writers like to use obscure words to deliberately make their writings not understandable to the general public; otherwise their works would be considered 俗 (*vulgar*). This way of thinking leaves a large burden on the readers.

2.2 Writing Directions

Chinese text consists of characters in virtual rectangle boxes of the same size concatenate together. Text can go either vertically or horizontally. In the old days when I was still young, people had all the freedom in the writing directions. Horizontal writing could go from left to right or right to left. Vertical writing always goes from top to bottom, but as they form multiple columns, the column could either go from right to left or from left to right (Figure 2.2). This makes the usage of spaces in tighter areas such as newspapers and street signs more effective. However, sometimes this can cause confusion even for native readers.

When text appears in large paragraphs, you can easily tell how they go by just taking a glimpse of the contents. People unfamiliar with the language can still identify it by looking at the spaces at the beginning and end of the paragraphs. You can also tell by the spacing between characters, because that is usually smaller than the spacing between rows or columns. The real challenge is reading text from banners, signboards, classified ads, etc. where there may be only a single row of text. In some interesting cases, text can even be read both ways.

Take the phrase 人人愛我 for example. When reading from left to right it means "Everyone loves me," but it means "I love everyone" when read from right to left.

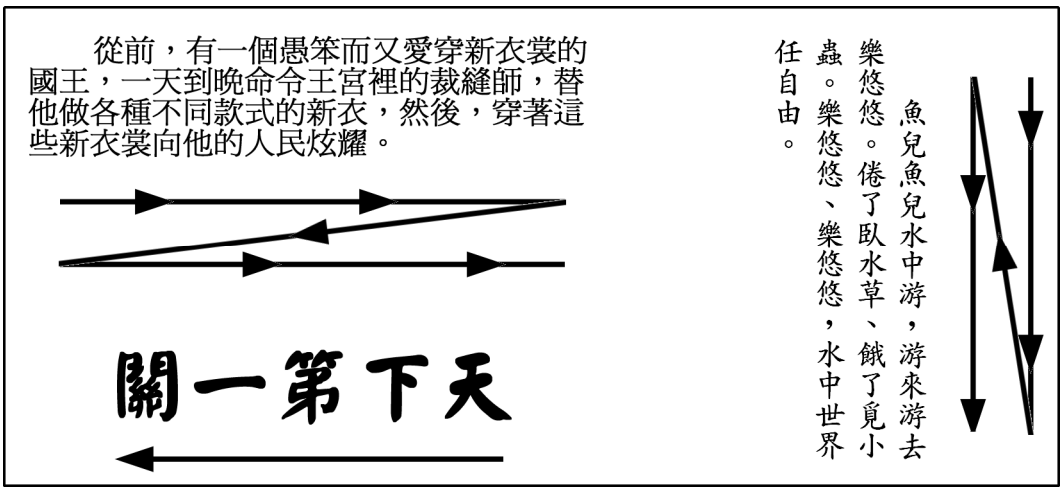

Figure 2.2 Chinese text writing directions

To avoid confusion, both Taiwan and China now have standardized the writing direction of Chinese text. For horizontal writing, they go from left to right for the convenience of mixing with Latin letters and numbers. For vertical writing, the sequence of columns goes from right to left because that is the traditional way used in all the books and other printings that have been in existence for thousands of years. This writing standard certainly makes everyone's life easier and reduces guessing. However, bear in mind that there are many writings that existed way before the introduction of the standard. Horizontal text written before the modern era are most likely to go the other way—from right to left. The other thing to remember is that a standard is just something for people to follow. Even though you can expect it to be obeyed in the printing for books, magazines, etc., there is no guarantee that people will always do the "right" thing when it comes to private or informal usage.

I have trouble understanding the rationale behind the traditional way of writing from right to left. The strokes within each character actually go from left to right; to me it would make more sense for the characters to go from left to right as well. Another inconvenience that I recall from this is doing paint-brush homework when I was little. Because of going from right to left, my hand always smudged the fresh writing I just made on the previous column, and that often resulted in a big mess.

In Taiwan, vertical writing is still the most common method used in books, magazines and other printed literary works. For text in electronics format, such as in websites or e-mail, writing goes horizontally. In mainland China, writing almost always goes horizontally from left to right. Occasionally, you may see vertical text printed by the side of pictures in newspapers, or on spring couplets pasted by the door. In such cases, the vertical columns progress from right to left. For the purpose of reading Chinese, you don't really need to worry about the writing direction if the materials are already in electronic format. When reading text from books, posters, or other non-electronic sources, you have to use a scanner (or camera) and OCR to capture the text. In such situations you may need to know which way the text goes in case the OCR program cannot recognize the blocks correctly. We will discuss this in more detail in the future sections that cover OCR software and handheld scanner devices.

2.3 Pinyin and Zhuyin

All Chinese characters have exactly one syllable in sound when pronounced. The pronunciation varies from region to region, and the differences can be quite significant. Mandarin Chinese is the pronunciation used by people in Beijing, and it is the official standard in both China and Taiwan. Mandarin Chinese has fewer sounds than English. One unique aspect of spoken Chinese is its use of tones. Pitches of voice are used to help differentiate among characters. There are five tones used in Mandarin. Considering that different tones sound differently, there are only about 1,300 unique sounds in Mandarin. Since these will be used for more than 47,000 characters (based on the *Kangxi dictionary*), many characters have the exact same pronunciation. There is no standard method that will tell us how to pronounce characters. Basically, all you can do is memorize the sound of each of the characters as you learn them. A simple rule of thumb is to see if the character contains other simpler characters in it, and then pronounce it using the sound of the simpler character. While this rule does not work well, it is better than a random guess.

Pinyin (拼音), also known as Hanyu Pinyin, is an effort by PRC for Romanization of the Chinese language. The idea is to represent Chinese characters using their sound (based on the Mandarin standard), which can then be represented by Latin alphabets using some simple rules. This standard was first introduced in 1958, but the idea of Romanization is certainly not a new concept. A similar system called Wade (aka Wade-Giles), developed by Thomas Wade and Herbert Giles, has been widely used to Romanize Chinese since the 19th century.

Pinyin was adopted by the ISO in 1979, and is now the most commonly used Romanization standard for Mandarin Chinese. Pinyin is used for its close resemblance to the phonetic system with which people are already familiar. The phonetic system is called Zhuyin (注音), aka Bopomofo, and was invented in 1913.

The Zhuyin system consists of 37 Chinese phonetic symbols and four notations to represent the five different tones in Mandarin. The system contains consonant and vowel symbols. It represents the pronunciation of each character using one consonant and one or two vowels put together, along with a tone notation. The sole purpose of Zhuyin is to provide phonetic annotation of the Mandarin sounds of Chinese characters. We can find the use of Zhuyin in many of the dictionaries published after 1913.

Other than in dictionaries, one major use of Zhuyin is for annotating text so it can be read by children, or people who are starting to learn the language. As of today, Zhuyin is still used in Taiwan, Hong Kong, and some overseas Chinese communities.

The way to use Zhuyin is to place these symbols on the right-hand side of each character. Figure 2.3 shows an example of text printed with the Zhuyin symbols. One other use of the Zhuyin system is to input Chinese characters into the computer.

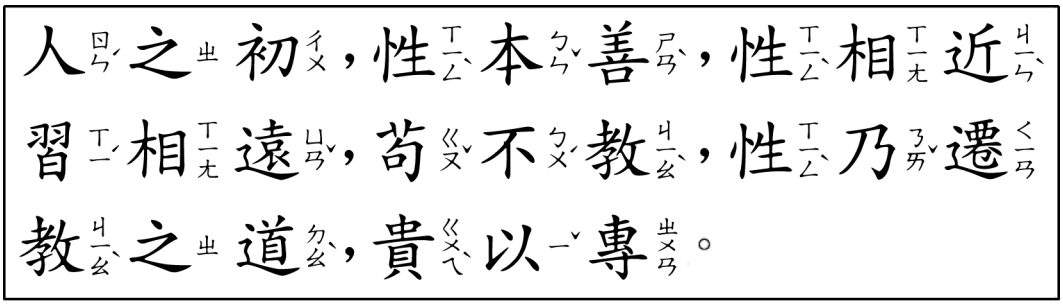

Figure 2.3 Text with Zhuyin annotations

One major drawback with the Zhuyin system is that it uses a special set of symbols that is unfamiliar to foreigners. From what I can see, the symbols are not strongly correlated to Chinese characters either. It has always been a mystery to me why people choose to use them. Pinyin closely resembles Zhuyin, but it uses the Roman alphabet to replace the phonetic symbols. As such, it can be used for both Romanization and phonetic annotations. When used for annotation, it is usually put on top of the characters, as shown in Figure 2.4.

rén zhī chū xìng běn shàn xìng xiāng jìn
人 之 初，性 本 善，性 相 近
xí xiāng yuǎn gǒu bú jiào xìng nǎi qiān
習 相 遠，苟 不 教，性 乃 遷
jiào zhī dào guì yǐ zhuān
教 之 道，貴 以 專。

Figure 2.4 Text with Pinyin annotations

Figure 2.5 shows a list of phonetic symbols used in Zhuyin and a mapping of Pinyin.

Consonants

ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ	ㄐ	ㄑ	ㄒ
b	p	m	f	d	t	n	l	g	k	h	j	q	x

ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ
zh	ch	sh	r	z	c	s

Vowels

ㄧ	ㄨ	ㄩ
i y yi	u w wu	u y yu

ㄚ	ㄛ	ㄜ	ㄝ	ㄞ	ㄟ	ㄠ	ㄡ	ㄢ	ㄣ	ㄤ	ㄥ	ㄦ
a	o	e	e ie	ai	i ei	ao	u ou	an	n en	ang	ong eng ing	er

Figure 2.5 Zhuyin phonetic symbols and their mappings in Pinyin