J Psycholinguist Res (2006) 35:405–426
DOI 10.1007/s10936-006-9022-y

ORIGINAL ARTICLE

# Analysis of a Chinese Phonetic Compound Database: Implications for Orthographic Processing

**Janet Hui-wen Hsiao · Richard Shillcock**

Published online: 29 July 2006
© Springer Science+Business Media, Inc. 2006

**Abstract** The complexity of Chinese orthography has hindered the progress of research in Chinese to the same level of sophistication of that in alphabetic languages such as English. Also, there has been no publicly available resource concerning the decomposition of Chinese characters, which is essential in any attempt to model the cognitive processes of Chinese character recognition. Here we report our construction and analysis of a Chinese lexical database containing the most frequent phonetic compounds decomposed into semantic and phonetic radicals according to Chinese etymology. Each radical was further decomposed into basic stroke patterns according to a Chinese transcription system, Cangjie (Chu, 1979 Laboratory of chu Bong-Foo Retrieved August 25, 2004, from http://www.cbflabs.com/). Other information such as pronunciation and character frequency were also incorporated. We examine the distribution of different types of character, the information skew in phonetic compounds, the relations between subcharacter orthographic units and the pronunciation of the entire character, and the processing implications of these phenomena in terms of universal psycholinguistic principles.

**Keywords** Chinese characters · Orthography · Mental lexicon · Chinese database · Visual word recognition

## Introduction to Chinese Orthography

In Chinese orthography, characters are the smallest units of the orthography and can be regarded as the perceptual unit of the orthography (Hoosain, 1991). Chinese

J. Hui-wen Hsiao (✉)
Department of Computer Science & Engineering,
University of California San Diego,
San Diego, USA
e-mail: jhsiao@cs.ucsd.edu

R. Shillcock
School of Informatics,
University of Edinburgh, Edinburgh, UK

characters consist of several individual strokes. There are about 20 distinct strokes in Chinese. A few strokes comprise a *stroke pattern* that appears to be a recurrent orthographic unit of Chinese characters. Some of these orthographic units can be characters by themselves. Units can be constructed recursively to form other composite units. Those units that are integral stroke patterns and cannot be further decomposed into other units have been referred to as *single bodies* (Huang & Hu, 1990; Chen, Allport, & Marshall, 1996).

A subgroup of these single bodies is explicitly listed in Chinese dictionaries as *Bu Shou*, to serve as an index system of the dictionaries. There are currently 189 Bu Shou listed in a simplified Chinese dictionary (Xin Hua, 1979) and 214 Bu Shou in a traditional Chinese dictionary (Mandarin Promotion Council, Ministry of Education, R.O.C, 2000). Chen et al. (1996) further defined the term *lexical radical* as a single body in a given character that is a unit from the Bu Shou and also occurs in its Bu Shou position with respect to the rest of the character. Each character contains exactly one lexical radical, which usually implies the meaning of the character and is also referred to as the *signific* or the *semantic radical*. The remainder of the character, if any, usually informs the pronunciation of the character and is referred to as the *phonetic* or the *phonetic radical*.

In general, there are four different types of Chinese characters: *pictographs*, *indicatives*, *ideographs* and *semantic-phonetic compounds*. Pictographs are depictions of material objects, such as "山" (mountain); indicatives are abstract characters with indicating signs, such as "上" (up). These two types of characters are also referred to as simple characters. The other two types are compound characters. An ideograph is a composition of the meanings of its components. For example, the character "休" (rest) consists of a person on the left (人), and a tree on the right (木), showing a person resting beside a tree. The last type of Chinese character is the semantic-phonetic compound (or simply *phonetic compound*). Its orthography contains both semantic and phonetic radicals. This group of characters comprises about 81% of the 7,000 frequent characters in a Chinese dictionary (Li & Kang, 1993). Most phonetic compounds have a left-right structure. This left-right structure is the most tractable aspect of Chinese orthographic structure, and has been a focus for understanding how Chinese readers recognize Chinese characters.

## Regularity and Consistency

A phonetic compound can be decomposed into two major components: a semantic radical that bears the meaning of the character, and a phonetic radical that typically provides partial information about the pronunciation of the character. For example, the character "沐" means "take a bath" and is pronounced as "mu4" in Pinyin.[1] It consists of a semantic radical on the left, which means "water," and a phonetic radical on the right, which is pronounced the same as the character itself. For current purposes, we call these characters *regular characters*. Some characters have the same pronunciation as their phonetic radical but with a different tone, such as "袖." Its phonetic radical "由" is pronounced as "iou2" in Pinyin. However, "袖" has a different tone—it is pronounced as "iou4." These characters are referred to as *semi-regular characters*. There are also *irregular characters*, which are pronounced with different segments

[1] The Chinese Pinyin system is a spelling system based on the Latin alphabet.

from their phonetic radicals, such as "洒" (sa3) and "西" (xi1). Among irregular characters, some characters may be pronounced similarly to their phonetic radical. They may share an onset or a rime. Hence, there are three subcategories in irregular characters: *alliterating* (sharing an onset), *rhyming*, or *radically irregular* (i.e., no apparent relationship).

A *regularity effect* and a *frequency effect* have been found in the processing of Chinese phonetic compounds: Chinese readers name regular characters faster than irregular characters, and name high-frequency characters faster than low-frequency characters. There is also a frequency by regularity interaction in Chinese, as in English (e.g., Seidenberg, 1985; Hue 1992; Liu, Wu, & Chou, 1996). These effects have been commonly used to examine the cognitive plausibility of computational models (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996).

Another dimension along, which to categorize phonetic compounds concerns the consistency of their phonetic radicals. We call a phonetic compound consistent if all the other characters with the same phonetic radical in the same position have the same pronunciation as this given character.[2] Similarly, a phonetic radical is consistent if its pronunciation is identical to the pronunciation of all characters containing this phonetic radical (Feldman & Siok, 1999). There are approximately 800 phonetic radicals in the Chinese language (Taylor & Taylor, 1983), and 38% of the phonetic radicals are consistent (Zhou, 1978).

## Orthographic Representation in Modelling Chinese Character Recognition

The granularity of Chinese orthography is a fundamental issue in any modelling efforts. The modelling of Chinese character recognition has long suffered from an input representativeness problem due to its complexity (cf. Christiansen & Charter, 2001). There is an ongoing debate about how to represent Chinese characters in a psychologically realistic way. Researchers have previously suggested that Chinese character recognition starts from an analysis of features and the number of individual strokes (e.g., Seidenberg, 1985; Chen & Young, 1989; Perfetti & Zhang, 1991; Yu & Cao, 1992a, b; & Perfetti & Tan, 1999; Xing, Shu, Li, 2002). The critical considerations are the goals of the modelling and where generalization is expected to occur. Models of the processing of alphabetic languages have sometimes used the features of letters (e.g., McClelland & Rumelhart, 1981) and sometimes the letters themselves (e.g., Shillcock & Monaghan, 2001) depending on whether or not the intent is to investigate, for instance, visual generalization between similar letters such as **a** and **g** or pronunciation generalization between letters such as b and d.

In early attempts to model Chinese language processing, researchers usually used strokes to encode orthographic representations (e.g., Perfetti & Tan, 1999; Xing et al., 2002). In recent years, studies have shown that recognition by skilled readers is based upon well-defined orthographic constituents, i.e., single bodies, which are integral stroke patterns that cannot be further decomposed into other units, instead of individual strokes as previously thought (Chen et al., 1996; Zhou & Marslen-Wilson, 1999). In order to reflect this observation and to facilitate modelling, we have constructed a Chinese lexical database containing the 3,027 most frequent phonetic compounds

[2] We treat the same radical appearing in different positions of a character differently when examining consistency, in order to reflect the observation that radicals appear to have positional-specific representations (Taft & Zhu, 1999).

from frequent Chinese characters, which are decomposed into semantic and phonetic radicals according to Chinese etymology (cf. Zhang & Chen, 1716; Harbaugh, 1998). We also decompose all the radicals into 179 basic stroke patterns based on the *Cangjie transcription system*, a Chinese character transcription system which encodes each character according to its orthographic constituents (Chu, 1979).

**Chinese Phonetic Compound Database**

In summary, the majority of Chinese characters are phonetic compounds. These characters usually consist of a phonetic radical and a semantic radical. A semantic radical bears information about the meaning of the character; a phonetic radical usually suggests how the character might be pronounced. Most phonetic compounds have a left-right structure. Usually the semantic radical is on the left and the phonetic radical is on the right of the character.

In order to understand how these phonetic compounds are formed, the database reflects the structures of all phonetic compounds. It contains the 3,027 most frequent phonetic compounds from frequent Chinese characters, decomposed into semantic and phonetic radicals according to Chinese etymology. This decomposition is uncontroversial (cf. Zhang & Chen, 1716; Harbaugh, 1996) and involved native-speaker intuitions.

Our analysis of the cognitive implications of the database statistics is based on current research in visual word recognition. Recently it has become clearer that the human fovea is precisely vertically split, and initially the left and right visual fields, either side of the fixation point, are projected contralaterally to the right and left hemisphere (RH and LH) respectively (Sperry, 1968; Fendrich & Gazzaniga, 1989; Fendrich, Wessinger, & Gazzaniga, 1996) and that this anatomical fact has implications for reading (Shillcock, Ellison, & Monaghan, 2000). This fact about the anatomy of the visual system motivates a concentration on the left-right structure of Chinese characters. We have explored elsewhere some of the implications of this anatomical fact for the processing of Chinese characters (Hsiao & Shillcock, 2004, 2005). From the above analysis, it is clear that the typical granularity in Chinese orthography is substantially coarser than that found in English orthography: English four-letter words, the subject of much modelling attention, contain four constituents, but the left-right structured phonetic compounds in our database typically contain two. We can assume that when a subject fixates at the middle of such a character, initially the left half of the character will be projected to the RH, and the right half of the character to the LH. Consequently, the initial processing of the two halves of a fixated character is located in different hemispheres. According to Shillcock et al.'s (2000) approach, in the current study we examine the hypothesis that there is an equitable division of labour between the two hemispheres during reading, which may be reflected in the structure of the lexicon.[3]

[3] Note that in the current study, readers' eye fixation has been hypothesized to be between the two radicals of a left-right structured character. In Chinese text reading, it has been shown that there is no tendency for eyes to land more frequently at a particular position in a character (Yang & McConkie, 1999; Tsai & McConkie, 2003), possibly because the length of a character is too short for the effects to emerge (see Tsai & McConkie, 2003, for a discussion), Nevertheless, we can assume that the OVP of isolated Chinese character reading is similar to that of short English words (O'Regan, 1990), which is close to the centre of a character. An examination of OVP in Chinese character recognition is beyond the scope of the current paper.

Also, in order to reflect the observation that the smallest functional processing units of Chinese character recognition are the well-defined stoke patterns (i.e., single bodies), which repeatedly appear in Chinese characters (Chen et al., 1996), each radical was further decomposed into basic stroke patterns, or single bodies. A similar decomposition had been achieved in the Chinese transcription system, Cangjie (Chu, 1979), and we thus decomposed each radical accordingly. The frequency information (Huang, 1995), the pronunciation (Pinyin), and the information regarding regularity and consistency of each character were also put into separate tables in this lexical database.[4]

## Methodology

All the 3,027 phonetic compounds were put into two separate tables. The first table contained all phonetic compounds with their semantic and phonetic radicals forming a left-right structure, irrespective of whether the semantic radical appeared on the left or right. Characters with a radical that has its main body occupying one side of a character, such as the semantic radical "辶" (chuo4) in "遠" (yuan3), were also included. The criteria used to include these exceptional cases were as follows:

1. The radical of a given character occupies the whole left or right-side of the character and the top or bottom part of the other side, such as "广" (yan3).
2. For the radicals in (1), the part on one side has more strokes than the part on the top or bottom of the other side. In other words, the part on the side constitutes the principal part of this radical, and the radical could be recognized from this information alone. Hence, "广" does not meet this criterion.
3. The principal part of this given radical is not a radical itself. That is, there is no ambiguity among characters with this principal part on the same side. For example, the semantic radical of the character "施" (shi1) consists of the left part "方" (fang1) and the top of the right part. Since there is no existing character with "方" (fang1) on the left side but without the top part on the right-side, "施" meets this criterion.

The exceptional characters that met all the above criteria were included in the first table, which contained characters with a clear left-right structure. These exceptional characters were those with the following semantic radicals: 疒 (chuang2), 辶 (chuo4), 旰 (gan4), 彥 (yan3) and 阝 (yin3). The second table contained the rest of the phonetic compounds, which were non-left-right structured and had a vertical, concentric, or some other irregular structures.

Each radical was then further decomposed into those basic stroke patterns defined in Cangjie. This decomposition was achieved by encoding each radical in terms of a set of predefined basic stroke patterns. We extracted 110 such stroke patterns from Cangjie encoding rules, and coded them with numbers from 1 to 110. For radicals with more than one identical component, such as "林", we used an extra code to represent this geminate component. The spatial complexity of Chinese orthography compelled

[4] The standard query language (SQL) can produce a larger table containing all related information for each character. The database is available on request from the first author.

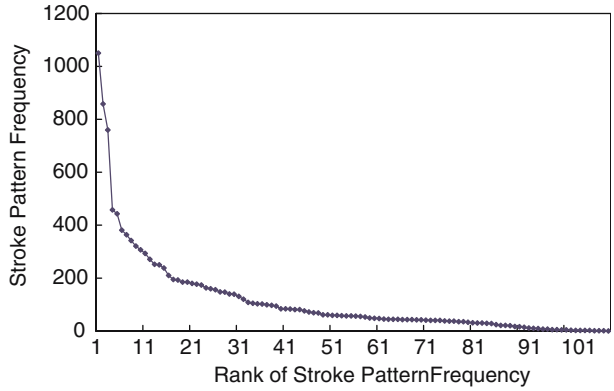


**Fig. 1** Zipf distribution of Stroke pattern frequency among left-right structured phonetic compounds

us to adapt this approach to the representation of geminates, compared with that taken in the modelling of alphabetic spelling (cf. Houghton, Glasspool, & Shallice, 1994.)[5] Both codes for the original and the geminate components were included in the encoding of this given radical. For example, code 73 represented the pattern "口." For the radical "呂," which contains two "口," we used an extra code 128 to represent the other "口." Hence, both codes 73 and 128 were included in the encoding of the radical "呂," whereas only code 73 was included in "口." In this way, we could make sure that the radical "呂" and "口" have one code, or stroke pattern, in common, and at the same time distinguish them with another code. The same applied to triples of the same pattern. For example, if there was a radical with a form "品," an extra code, 147, would be required to represent this triple pattern; Code 73, 128 and 147 hence would all be included in the encoding of this radical. We used numbers 111–179 to code these geminates and triples.

The character frequency information is from the Chinese character list reported by Huang (1995). This list contains information about frequency of usage and number of strokes for each of the 13,060 traditional Chinese characters. The frequency information was taken from a corpus consisting of 171,882,493 BIG-5 Chinese characters, which appeared on Usenet newsgroups during 1993–1994.

Figure 1 shows the distribution of the frequency of the stroke patterns we derived from the Cangjie system, which exhibits a characteristic Zipf curve (Zipf, 1932). This curve is plotted by sorting all stroke patterns according to their frequency of appearance among all left-right structured phonetic compounds, with the most frequent stroke pattern first, and so on.

## Componentiality of Different Character Types

Among the 3,027 most frequent phonetic compounds, there are 2,159 characters with a clear left-right structure (left-right phonetic compounds). This is about 72% of

[5] In Houghton et al.'s study, they used only a single "geminate node" for all words with geminates. For example, this "geminate node" was activated in both of the representations for the word "deer" and "door," although the duplicate letter in the two words were different.

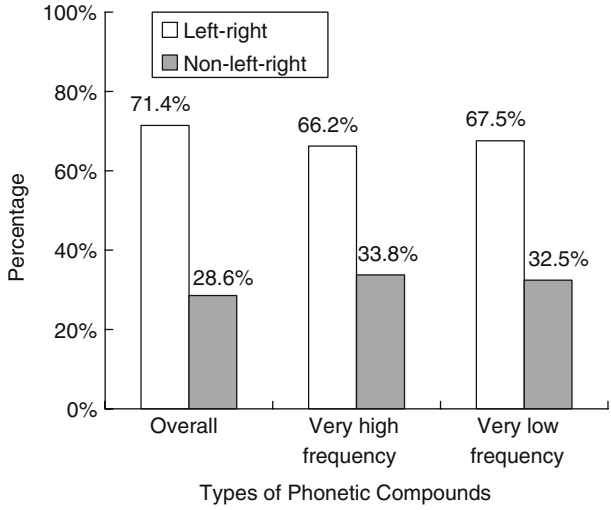


**Fig. 2** Distribution of types of Chinese phonetic compound characters and its interaction with character frequency (the top and bottom 10%)

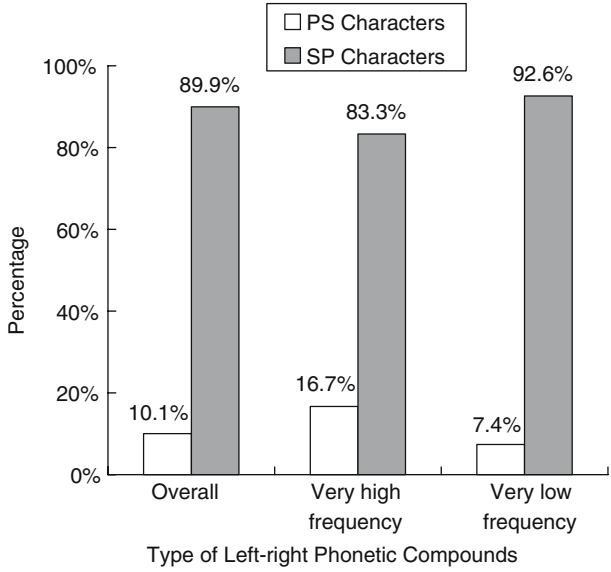


**Fig. 3** Distribution of SP and PS characters interacting with frequency (the top and bottom 10%)

the 3,027 phonetic compounds (Fig. 2). Among these left-right phonetic compounds, there are only 218 characters that have their phonetic radicals on the left side (i.e., PS characters), which is about 10% of the 3,027 left-right phonetic compounds. In other words, around 90% of the left-right phonetic compounds have their semantic radicals on the left and phonetic radicals on the right (SP characters; see Fig. 3).

Figure 2 shows that there are closely comparable proportions of left-right and non-left-right phonetic compounds in the top 10% and bottom 10% of the frequency range

considered. It has been argued that processing of low-frequency word types has greater componentiality and involves rule-governed processing; in contrast, processing of high-frequency word types tends to be holistic, reflecting the access of stored items. For example, it has been claimed that the past tenses of low-frequency English words tend to be accessed componentially, i.e., the—ed rule (cf. Marcus, 1996). The closely comparable proportions of left-right and non-left-right structures among very high- and very low-frequency characters imply an equal degree of componentiality in the processing of these two different structures. Thus, it is not possible to claim, given these assumptions about componentiality, that the left-right phonetic compound is any more or less naturally componential than the non-left-right phonetic compound.

Figure 3 shows a different picture: there tend to be proportionally more minority PS forms in the high frequency range compared with those in the low-frequency range. In other words, given these assumptions about componentiality, processing of SP forms tends to have higher componentiality than minority PS forms. This is in line with "dual-route" theories of rule-governed and non-rule-governed processes interacting with frequency, in a way that is adaptive in storage terms (cf. Marcus, 1996). That is, rule-governed processes can still yield low-frequency outcomes effectively.

## Entropy Analysis

Among all left-right phonetic compounds, there are 252 different radicals on the left of the characters. Some 104 out of the 252 radicals are semantic radicals. On the other hand, there are 857 different radicals on the right, and 843 of them are phonetic radicals. Hence, there is more variation on the right-side of the left-right phonetic compounds. Some radicals can be further decomposed into other radicals. For example, the phonetic radical of the character "沏" (qi1) is "切" (qie1), which can be further decomposed into "七" as the phonetic radical and "刀" as the semantic radical. In total, there are 73 decomposable radicals in the database. If, we do not consider these decomposable radicals, there are 249 different undecomposable radicals on the left-side of the characters, and 831 on the right. In total, there are 888 different undecomposable radicals.

Figures 4–6 compares the entropy of the radicals on the left and on the right of these left-right phonetic compounds. The entropy of the radicals is obtained from equation (1):

$$H(x) = -\sum_{x} P(x) \log P(x), \tag{1}$$

$P(x)$ is the probability of a given radical $x$ appearing in a specific position, that is, on the left or the right of a phonetic compound. In information theory, entropy concerns how much randomness is in a signal, or alternatively, how much information is carried by the signal. It is sensitive to both the probability distribution of different types of events and the total number of events in the signal. The greater the entropy is, the more information the signal carries. We thus use this measure to examine the information distribution within Chinese characters.

How is Fig. 4 to be interpreted? The greater entropy on the right of the figure reflects the fact that the right-side is more variable than the left-side, with this contrast typically reflecting the greater variability of the phonetic radical as opposed to the semantic radical. Can, we interpret this asymmetry further? Figure 5 shows the

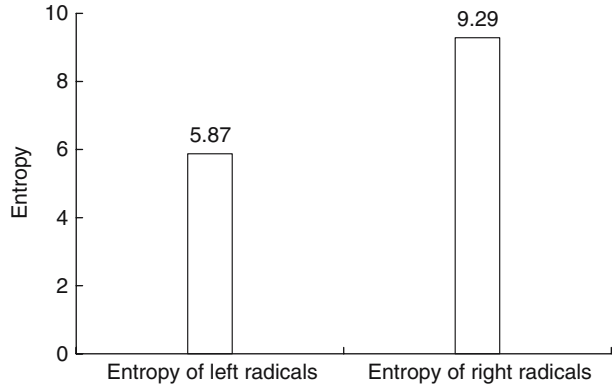


**Fig. 4** Entropy analysis of the radicals on the left and on the right of the left-right phonetic compounds

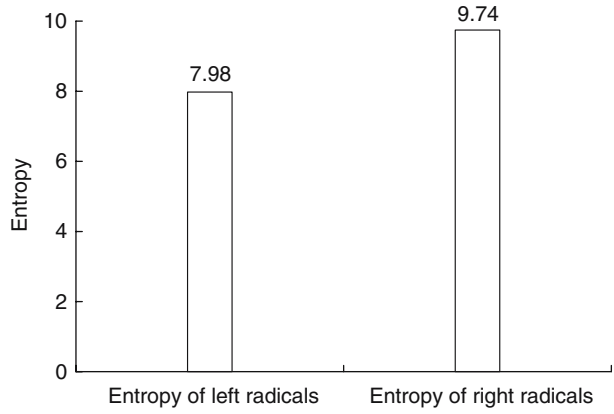


**Fig. 5** Entropy analysis when the graphotactic constraints which exist in the real lexicon are lifted

same contrast when, we lifted the graphotactic constraints, which exist in the real lexicon, and which mean that any one radical on the left of the character can only appear in conjunction with a small subset of radicals on the right of the character, and *vice versa*. That is, we pretended that the lexicon of left-right phonetic compounds contained every possible combination of radicals that can appear on the left and radicals that can appear on the right. Without the graphotactic constraints of real Chinese, every left radical is able to pair with every right radical to comprise a character. If, we compare Figs 4 and 5, we can see that the graphotactic constraints decrease the entropy more on the left than on the right. This implies that, in Chinese, the distribution of radicals on the left is more skewed than on the right. In other words, on the left of the phonetic compounds, there are some very frequent and some very infrequent radicals, whereas the distribution is flatter on the right, indicating that the right half of characters is more informative. This fact can be seen in the Zipf curve (cf. Zipf, 1932) shown in Fig. 7, in which distribution of right radicals is longer and flatter than that of the left radicals. Figure 6 shows the entropy analysis of very high and very low-frequency phonetic compounds (i.e., the top and bottom 10% in terms of frequency). Their left-right entropy distributions are similar to each other, with slightly

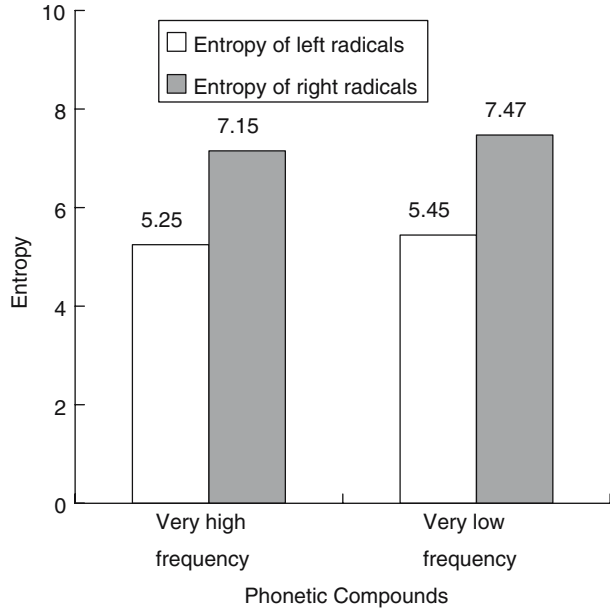


**Fig. 6** Entropy analysis of very high- and very low- frequency phonetic compounds (*i.e.*, the top and bottom 10%)

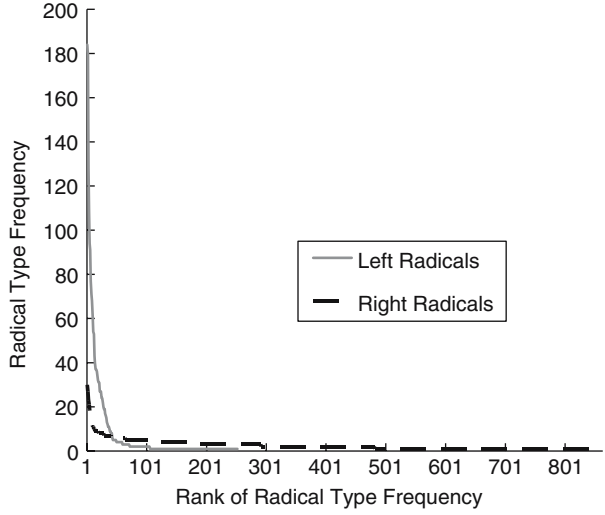


**Fig. 7** Zipf distribution of radical type among left-right structured phonetic compounds

smaller entropy among very high-frequency compounds. This fact is again reflected in the Zipf distribution of radical type frequency (Fig. 7), that is, a flatter frequency distribution among very low frequency compounds than among very high-frequency compounds.

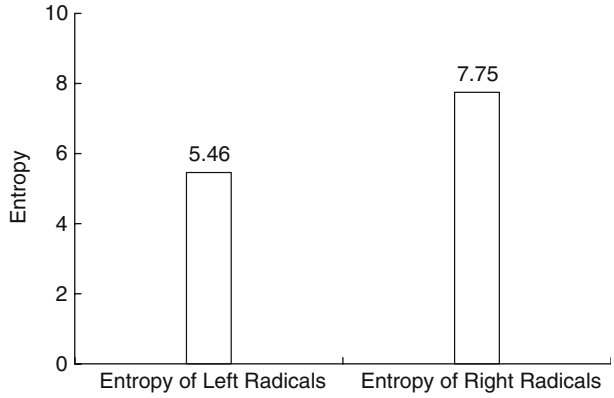


**Fig. 8** Entropy analysis of the radicals on the left and on the right in terms of character token frequency

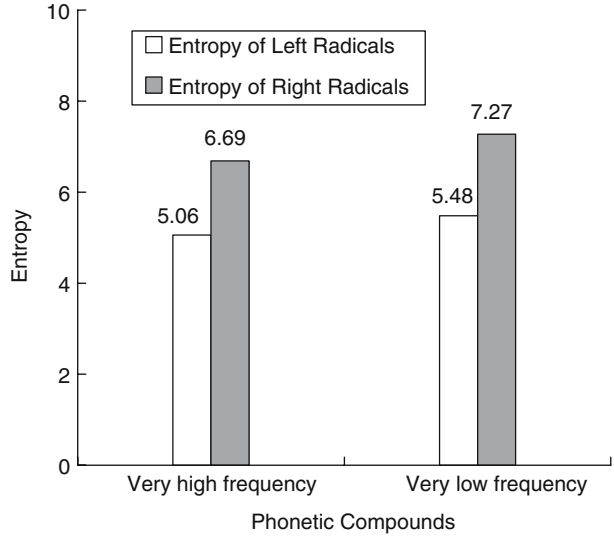


**Fig. 9** Entropy analysis of very high and very low frequency characters (the top and bottom 10%), in terms of character token frequency

Figures 8 and 9 shows the same entropy analysis, but in terms of character token frequency. If, we compare Fig. 8 with Figs 4 and 5, the entropy of right radicals decreases more than that of the left radicals. This decrease means the distribution of right radicals is less flat after taking token frequency into account (see Fig. 10 for the Zipf curve for token frequencies). In other words, the usage of characters, reflected in the character token frequencies, makes the levels of entropy on the left and right of the left-right phonetic compounds more similar. Figure 9 compares this left-right entropy distribution among the top and bottom 10% frequent characters in terms of token frequency. The lower entropy among very high-frequency characters reflects the fact that, in Chinese texts, the radicals of very high-frequency characters tend to be of just a few types (Fig. 9).

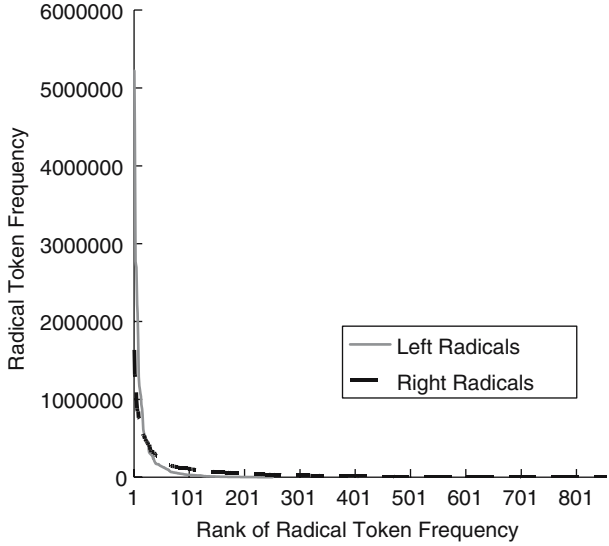


**Fig. 10** Zipf distribution of radical token frequency among left-right structured phonetic compounds

Thus our hypothesis has received support from this analysis: the character set should have evolved so as to produce an adaptive division of labour between the two hemispheres, and thus between the two hemifields, and the two halves of phonetic compounds. The entropy level is more closely equal—but with rather more information being projected initially to the LH—in the set of most frequently occurring characters, exactly where we would expect to find most adaptation.

## Character Regularity and Consistency

As, we have seen, phonetic radicals vary in their relationship with the pronunciation of the entire character. They may signify the pronunciation transparently (i.e., they are regular) or with varying degrees of transparency (i.e., they are semi-regular or irregular); indeed, the irregular category itself breaks down into subcategories of differing degrees of transparency. In addition, any one phonetic radical may be more or less consistent in its relationship with character pronunciation.

We will explore the claim that the relationship between a phonetic radical and the pronunciation of the whole character is interpretable in terms of universal psycholinguistic principles governing the relationships between spoken words, as addressed principally by priming experiments. On the basis of interference between homophones in tasks involving pronunciation judgements and semantic judgements, it has been shown that the pronunciation of a Chinese character is activated early in recognition and has been argued to be integral to the lexical access of the character (Perfetti & Zhang, 1995). Given the existence of short-range phonological priming between similarly pronounced words (e.g., Collins & Ellis, 1992), we can interpret the role of the phonetic radical in a character as answering the question "what other character pronunciation would facilitate the pronunciation of the current character." We have examined the regularity and consistency of the 3,027 most frequent Chinese

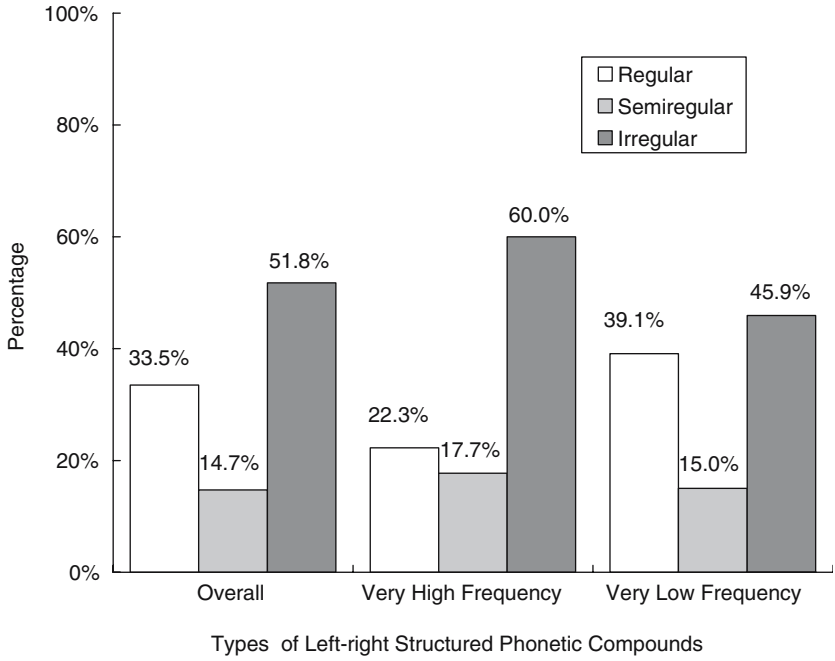


**Fig. 11** Distribution of regularity of left-right structured phonetic compounds and its interaction with character frequency (the top and bottom 10%)

phonetic compounds. As shown in Fig. 11, overall only about 33% of the left-right structured Chinese phonetic compounds are regular and about 52% are irregular, including alliterating, rhyming and radically irregular characters (Fig. 12). The best clue to a character's pronunciation is something that has the same pronunciation as the character. Hence, the relationship between a regular character and its phonetic radical is identity priming: the radical is trivially the best clue to the character's pronunciation. As expected, from the "memory" argument rehearsed above, the percentage of regular characters is larger among low frequency characters than high-frequency characters. The same comparison for non-left-right structured phonetic compounds is shown in Fig. 13. These characters have structures different from the left-right structure, such as a vertical structure (e.g., 炁), a concentric structure (e.g., 圖), and others irregular structures (e.g., 望).[6] Figures 11 and 13 are similar with regard to the distribution of the different degrees of regularity across the frequency range, and we do not discern any important differences.

In both Figs. 11 and 13, the smallest category of phonetic compound is the semi-regular one, in which phonetic radical and character pronunciation differ only by tone. There is good reason to regard this category as a subcategory of the completely regular phonetic compounds. Although tone is a proper part of the phonology of a tone language (see, e.g., Van Lancker & Fromkin, 1973, 1978, albeit for Thai), its perceptual processing in speech judgements of words and nonwords, and in homophone

[6] Note that the phonetic radicals of non-left-right structured phonetic compounds tend to have small combinability and orthographic alterations (e.g., the phonetic radical of "有" is "又," which has been altered to "𠂇").

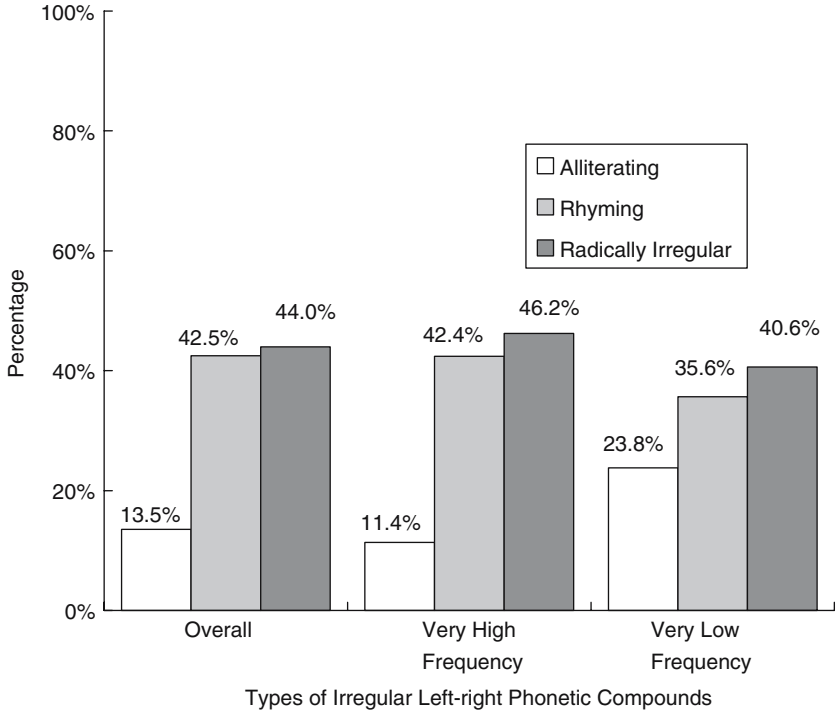


**Fig. 12** Distribution of types of irregular left-right structured phonetic compounds and its interaction with character frequency (the top and bottom 10%)

judgements of written characters is qualitatively different from that of segmental processing, being slower and more prone to misperception (see Cutler & Chen, 1997, albeit for Cantonese, in which the tone system is more complex; Repp & Lin, 1990; Taft & Chen, 1992; Spinks, Liu, Perfetti, & Tan, 2000). Tone in informal continuous speech is typically not perceived in strict temporal conjunction with the carrier word alone, but may require information about the speech contours over the previous and subsequent word(s) (Xu, 1994, 2001). Cutler (1986) has shown that lexical stress in English is not used prelexically to constrain lexical access (listeners automatically activate both meanings in homophone pairs such as *fore*arm and fore*arm* when either is heard). Although tone is much more important in Chinese, and is crucial for word identification, its perception is inherently slower and less reliable than that of segment perception. Indeed, Chen, Chen and Dell (2002) argue on the basis of an implicit priming task that the syllable minus the tone can act as a planning unit at the phonological level. From this perspective, we might categorize the regular and semiregular phonetic compounds together as being segmentally identical, together constituting around half of the phonetic compounds overall. Additionally, we can expect the semiregular category to be small in part because the number of tones is very limited.

If we take a closer look at the distribution of different types of irregular characters among the left-right structured phonetic compounds, interestingly, more than half of irregular characters still share some segments with their phonetic radicals (Fig. 12).

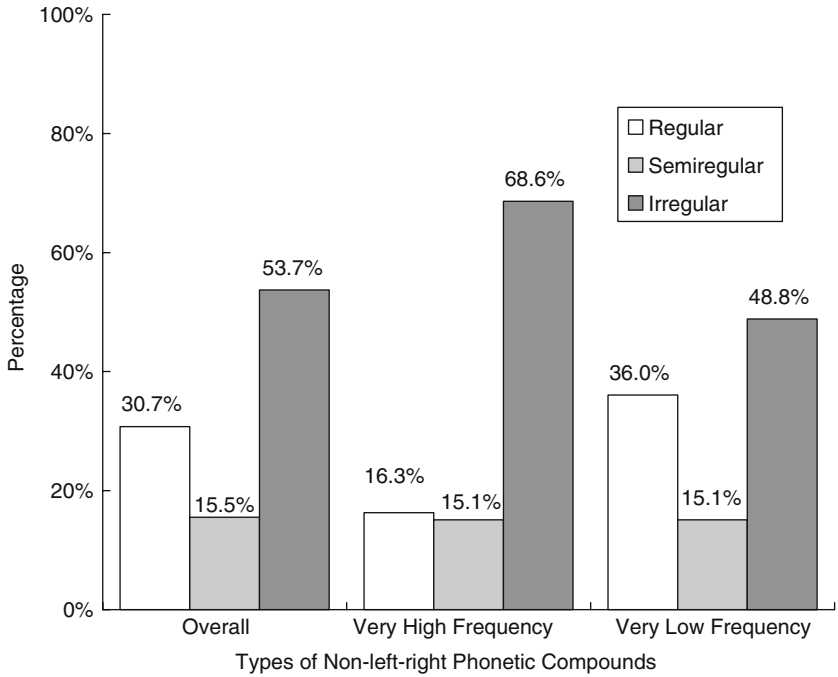


**Fig. 13** Distribution of regularity of non-left-right structured phonetic compounds and its interaction with character frequency (the top and bottom 10%)

In other words, only 23% of all left-right structured phonetic compounds have a phonetic radical, which has a radically different pronunciation. A similar distribution can also be found among irregular non-left-right phonetic compounds (Fig. 14). It suggests that, overall, phonetic radicals may not be as poor an indication of Chinese character pronunciation as is often thought. The distribution of types within the irregular phonetic radicals may be understood in terms of priming relationships between words. Note that one of the largest categories is the one in which the phonetic radical rhymes with the pronunciation of the whole character. There is a much smaller category of alliterating phonetic radicals, which share an onset with the pronunciation of the whole character. There is a substantial literature showing the salience of the rime in the phonological representation of words (see, e.g., Dumay et al., 2001). In phonological priming experiments, overlap at offset tends to lead to facilitation of the target (Radeau, Morais, & Segui, 1995; Monsell & Hirsh, 1998; Slowiaczek, McQueen, Soltano, & Lynch, 2000; Dumay et al., 2001; see also Norris, McQueen, & Cutler, 2002, for a discussion of strategic effects). If, we equate the rhyming phonetic radicals with such facilitation, then their preponderance in the irregular phonetic radicals can be understood.

In contrast, phonological priming studies tend to show inhibition when only the onset is shared (Monsell & Hirsh, 1998; Radeau et al., 1995). The alliterating phonetic cues would thus seem to be less preferable as cues, and this may explain the relative proportions of the two categories. We may extrapolate from this reasoning to say that it is adaptive for the more frequently used characters to contain a larger proportion of the better cues and a smaller proportion of the poorer cues. Thus, the high-frequency

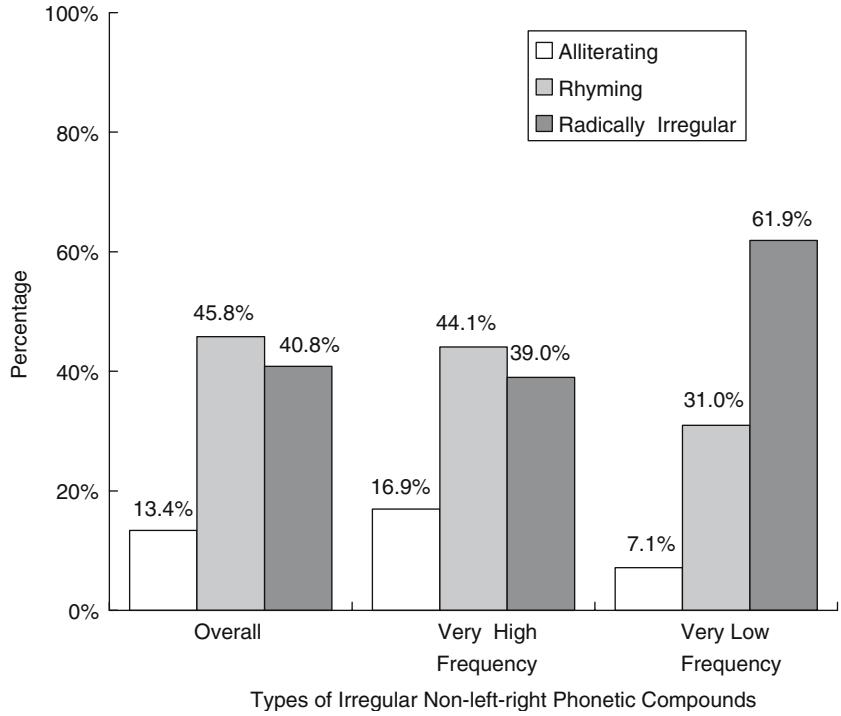


**Fig. 14** Distribution of types of irregular non-left-right structured phonetic compounds and its interaction with character frequency (the top and bottom 10%)

irregular characters contain a larger proportion of rhyming characters and a smaller proportion of alliterating ones (although the interaction of the alliterating ones with frequency is based on rather small absolute numbers).[7]

In discussing the regular left-right phonetic compounds, we used a memory explanation, saying that the low frequency pronunciations relied on rule-like decomposition; this decomposition was absolute—the low-frequency character was segmentally identical to its phonetic radical. In the irregular characters, pure decomposition does not lead to the correct answer. In these circumstances, the low-frequency characters and their pronunciations survive due to interactions occurring with other words. For instance, radically irregular pronunciations may be consistent across several instances of the phonetic radical, or there may even be wider systematicity between these apparently poor clues to pronunciation. One testable hypothesis raised by this observation is that although any one radically irregular phonetic radical only has an arbitrary relationship with the pronunciation of a character, in which it occurs, there may be a systematic relationship between the set of radically irregular phonetic radicals and

[7] The interaction of percentage of alliterating characters with character frequency does not hold for non-left-right structured irregular phonetic compounds, as there is a larger proportion of alliterating ones among the high-frequency irregular characters than low-frequency ones. But note that the absolute number of alliterating characters is even smaller than that of the left-right structured phonetic compounds. If we compare the top 50% high and the bottom 50% low-frequency characters among the non-left-right irregular phonetic compounds, there is still a larger percentage of alliterating characters among low frequency characters than high frequency characters (14.49 vs. 12.45%).

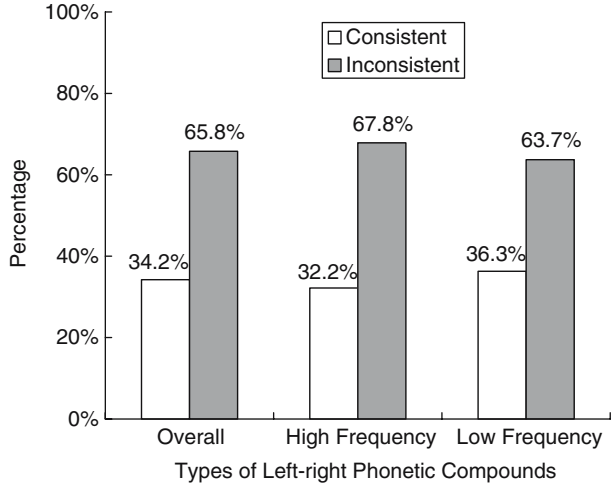


**Fig. 15** Distribution of consistency of left-right structured phonetic compounds and its interaction with character frequency (the top and bottom halves)

the corresponding set of pronunciations of the whole characters. For example, in the phonological representation space, the distribution of the pronunciations of the radically irregular characters may have a similar structure to the distribution of the pronunciations of their phonetic radicals; in other words, there may be a systematic mapping between these two distributions (cf. Shillcock, Kirby, McDonald, & Brew, *submitted*). (Testing this hypothesis is beyond the scope of the current paper.)

Figure 15 bears out the suggestion made above concerning consistency. There is a significant increase in consistency in the low frequency stratum compared with the high frequency stratum. ($\chi^2 = 3.865$; $df = 1$; $p < 0.05$) The most comprehensive way of testing the role of factors such as character frequency, radical frequency, consistency, range of segment identity in onset, nucleus and coda and so on, is to explore a connectionist mapping between the orthographic form of a character and its pronunciation (see Hsiao & Shillcock, 2004) In such modelling, we see that the regular and the semiregular characters pattern together in terms of the difficulty of learning the mapping, and the three categories of irregular character also pattern together, being somewhat harder to learn than the regular relationships, but not behaving strikingly differently between one another.

We have also compared the distribution of regularity among SP and PS characters (Fig. 16). Among PS characters, only 34% are regular or semiregular, compared with 50% among SP characters. This difference implies that the phonetic radicals in SP characters may be better indicators of pronunciation than such radicals in PS characters. We have reported that there is a significant regularity effect among SP characters but not among PS characters through connectionist modelling (Hsiao & Shillcock, 2004). This difference can be explained by the relatively high percentage of regular characters among SP characters.

Why should the exceptional PS structure have survived in the face of the dominant SP structure? We tested the hypothesis that the existence of the PS structure is adaptive in that it increases the variety of radicals on the left-hand side of the characters, thereby increasing the entropy on the left of the left-right phonetic compounds and

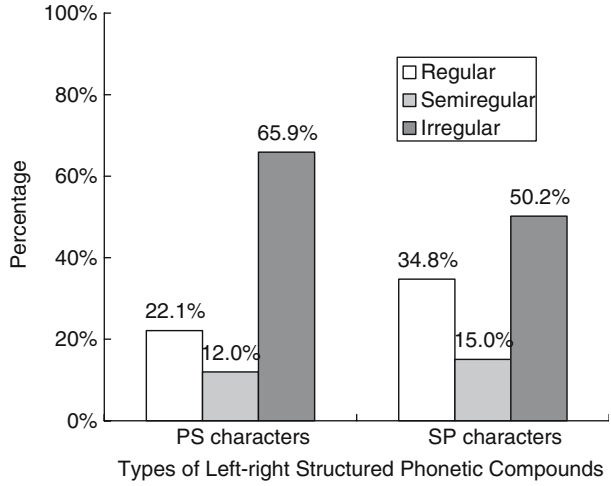


**Fig. 16** Distribution of regularity among SP and PS characters

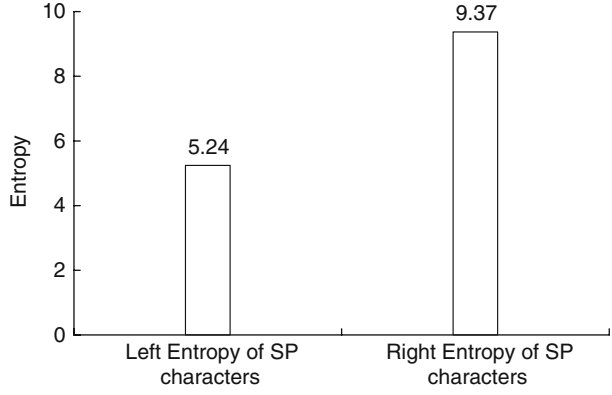


**Fig. 17** Entropy analysis of the radicals on the left and on the right of the SP characters

making the levels of entropy on the left and right sides more similar, in accord with our hypothesis about the division of labour between the hemispheres. Figure 17 shows the entropy levels for the SP characters alone. Compared with Fig. 4 showing the entropy levels for all of the characters, Fig. 17 is more uneven: the existence of the minority PS characters tends to offset this unevenness and make the left-right distribution of information more equitable.

## Conclusion

In computational modelling of Chinese language processing, input representativeness has long been problematic because of the lack of resources concerning Chinese character decomposition. Although there has been recent progress in behavioural studies, showing that character recognition by skilled Chinese readers is based upon

well-defined orthographic constituents instead of individual strokes (Chen et al., 1996; Zhou & Marslen-Wilson, 1999), this lack of resources has stalled the progress in computational modelling and statistical analysis. Through the construction of this Chinese lexical database, we have put diverse information about Chinese characters together, in order to examine the relations between subcharacter orthographic units in the pronunciation of the entire character, to stimulate and facilitate research in Chinese, and to bring research in Chinese to the same level of sophistication as that in English.

From the database, we have examined the structures of the 3,027 most frequent phonetic compounds. About two-thirds of them are left-right structured. We have shown that in spite of the less-frequent and irregular structures, and higher percentage of irregular pronunciations among non-left-right structured phonetic compounds, there are closely comparable proportions of left-right and non-left-right phonetic compounds in very high-and very low-frequency characters. This implies an equal degree of componential processing in the two different structure types. On the other hand, in the comparison of SP and PS characters, we see a greater degree of componential processing in SP characters, which occupy a much larger proportion of phonetic compounds, have higher percentage of regular pronunciations, and tend to have lower character frequency than PS characters. In other words, processing of the SP forms tends to have higher componentiality than the minority PS forms.

According to the results from the entropy analyses, in terms of both type and token frequencies, there is more variation on the right of the characters. In other words, in Chinese, the right half of characters is more informative than the left. This information bias to the right may just be the result of the cultural evolution of the language, but it does reflect the fact that the LH is typically more powerful than the RH. (It is also tempting to note the alignment of the phonetic information in the character with the phonological processing typically found in the LH, but there can be no demonstration of a causal connection.). Despite the typical dominance of the LH, it is adaptive to have a more or less equal division of labour between the hemispheres. We have shown that the typical usage of characters, reflected in the token frequencies, makes the levels of entropy on the left and right of the left-right phonetic compounds more similar. The existence of the minority PS characters has the same effect, offsetting the skew of the majority SP characters and making the left-right distribution of information more equitable.

Regarding character regularity and consistency, we have shown that the relationship between a phonetic radical and the pronunciation of the whole character is interpretable in terms of universal psycholinguistic principles governing the relationships between spoken words. Due to the qualitatively different perceptual processing of Chinese tones compared with segmental processing, and the fact that the syllable minus the tone can produce implicit priming at the phonological level (Chen et al., 2002), semiregular characters can be treated as a subcategory of the completely regular phonetic compounds. The "memory" argument explains the interaction between character regularity and frequency: low-frequency pronunciations rely on absolute decomposition, in which the character is segmentally identical to its phonetic radical. Whereas for irregular characters, whose pronunciations cannot rely on pure decomposition, a large proportion of them share the same offset with their phonetic radicals, and a small proportion share the same onset with their phonetic radicals. This distribution may be explained by the offset facilitation and onset inhibition in phonological priming experiments. The radically irregular characters survive due to the consistency of their phonetic radicals. We have also proposed that there may be a systematic

relationship between the radically irregular characters and their phonetic radicals. In the comparison between SP and PS characters, we have shown that SP characters have a higher percentage of regular characters than PS characters. This fact may explain the significant regularity effect found in SP characters, but not PS characters, in the modelling work (Hsiao & Shillcock, 2004).

From the analyses of the Chinese phonetic compound database, we have not only understood the substructures and distribution of different types of characters, but also the implications of these substructures and distributions for the orthographic processing of Chinese characters. Together with the existing resources for alphabetic languages such as English, this database can thus help us to examine the similarities and differences between radically different orthographies and arrive at a better understanding of processing universals in reading.

## References

Chen, J. Y., Chen, T. M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language, 46*, 751–781.

Chen, M. J., & Young, Y. F. (1989). Reading Chinese: A holistic or piecemeal process? In a. F. Bennett & K. M. McConkey (Eds.), *Cognition in Individual and social Contexts*. Amsterdam: Elsevier Science Publishers.

Chen, Y. P., Allport, D. A., & Marshall, J. C. (1996). What are the functional orthographic units in Chinese word recognition: the stroke or the stroke pattern? *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 49*, 1024–1043.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences, 5*, 82–88.

Chu, B. (1979). Laboratory of Chu Bong-Foo. Retrieved Aug. 25, 2004, from http://www.cbflabs.com.

Collins, A. F., & Ellis, A. W. (1992). Phonological priming of lexical retrieval in speech production. *British Journal of Psychology, 83*, 375–388.

Cutler, A. (1986). "Forbear" is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech, 29*, 201–220.

Cutler, A. & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics, 59*, 165–179.

Dumay, N., Benraiss, A., Barriol, B., Colin, C., Radeau, M., & Besson, M. (2001). Behavioral and electrophysiological study of phonological priming between bisyllabic spoken words. *Journal of Cognitive Neuroscience, 13*, 121–143.

Feldman, L. B., & Siok, W. W. T. (1999). Semantic radicals in phonetic compounds: implications for visual character recognition in Chinese. In J. Wang, A. Inhoff & H. Chen (Eds.), *Reading chinese script* (pp. 19–35). London: Erlbaum.

Fendrich, R., & Gazzaniga, M. S. (1989). Evidence of foveal splitting in a commissurotomy patient. *Neuropsychologia, 34*, 637–646.

Fendrich, R., Wessinger, C. M., & Gazzaniga, M. S. (1996). Nasotemporal overlap at the retinal vertical meridian–Investigations with a callosotomy patient. *Neuropsychologia, 34*, 637–646.

Harbaugh, R. (1996). Chinese characters and culture. Retrieved Aug. 25, 2004, from http://www.zhongwen.com/.

Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity: a case study of Chinese*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Houghton, G., Glasspool, D., & Shallice, T. (1994). Spelling and serial recall: insights from a competitive queuing model. In G. D. A. Brown & N. C. Ellis (Eds.), *Handbook of spelling: theory, process and intervention*. (pp. 365–404). Chichester, UK: John Wiley.

Hsiao, J. H., & Shillcock, R. (2004). Connectionist modelling of Chinese character pronunciation based on foveal splitting. *In Proceedings of the Twenty Sixth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hsiao, J. H., & Shillcock, R. (2005). Differences of split and non-split architectures emerged from modelling Chinese character pronunciation. *In Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Huang, S. K. (1995). Frequency counts of BIG-5 Chinese characters appeared on Usenet newsgroups during 1993–1994. Retrieved March 16, 2004, from http://www.geocities.com/hao510/charfreq/.

Huang, Z., & Hu J. Z. (1990). *The comprehensive study of chinese words*. Wu Han: Middle China Normal University Press.
Hue, C. W. (1992). Recognition processes in character naming. In H. C. Chen & O. J. L. Tzeng (Eds.), *Language processing in chinese* (pp. 93–107). Amsterdam: North-Holland.
Li, Y., & Kang, J. S. (1993). Analysis of phonetics of the ideophonetic characters in Modern Chinese. In Y. Chen (Ed.), *Information analysis of usage of characters in modern Chinese* (pp. 84–98). Shanghai: Shanghai Education Publisher (in Chinese).
Liu, I. M., Wu, J. T., & Chou, T. L. (1996). Encoding operation and transcoding as the major loci of the frequency effect. *Cognition, 59*, 149–168.
Mandarin Promotion Council, Ministry of Education, R.O.C. (2000). Dictionary of Chinese Variants. Retrieved Sept. 25th, 2004, from http://140.111.1.40/.
Marcus, G. F. (1996). Why do children say "breaked"? *Current Directions in Psychological Science, 5*, 81–85.
McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review, 88*, 375–407.
Monsell, S., & Hirsh, K. W. (1998). Competitor priming in spoken word recognition. Journal of Experimental Psychology: *Human Perception & Performance, 24*, 1495–1520.
Norris, D. G., McQueen, J. M., & Cutler, A. (2002). Bias effects in facilitatory phonological priming. *Memory & Cognition, 30*, 399–411.
O'Regan, J. K. (1990). Eye movements and reading. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes*. (pp 395–453). Amsterdam: Elsevier.
Perfetti, C. A., & Tan, L. H. (1999). The constituency model of Chinese word identification. In J. Wang, A. Inhoff & H. Chen (Eds.), *Reading chinese script* (pp. 115–134), Erlbaum: London.
Perfetti, C. A., & Zhang, S. (1991). Phonological processes in reading Chinese characters. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 633–643.
Perfetti, C. A., & Zhang, S. (1995). Very early phonological activation in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 24–33.
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.
Radeau, M., Morais, J., & Segui, J. (1995). Phonological priming between monosyllabic spoken words. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1297–1311.
Repp, B. H., & Lin, H. (1990). Integration of segmental and tonal information in speech perception: a cross-linguistic study. *Journal of Phonetics, 18*, 481–495.
Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19*, 1–30.
Shillcock, R., Ellison, T. M., & Monaghan, P. (2000). Eye-fixation behavior, lexical storage, and visual word recognition in a split processing model. *Psychological Review, 107*, 824–851.
Shillcock, R., Kirby, S., McDonald, S., & Brew, C. (*submitted*). Exploring systematicity in the mental lexicon.
Shillcock, R. C., & Monaghan, P. (2001). The computational exploration of visual word recognition in a split model. *Neural Computation, 13*, 1171–1198.
Slowiaczek, L. M., McQueen, J. M., Soltano, E. G., & Lynch, M. (2000). Phonological representations in prelexical speech processing: evidence from form-based priming. *Journal of Memory & Language, 43*, 530–560.
Sperry, R. W. (1968). Apposition of visual half-fields after section of neocortical commissures. *Anatomical Record, 160*, 498–499.
Spinks, J. A., Liu, Y., Perfetti, C. A., & Tan, L. H. (2000). Reading Chinese characters for meaning: the role of phonological information. *Cognition, 76*, B1–B11.
Taft, M., & Chen, H. C. (1992). Judging homophony in Chinese: the influence of tones. In H. C. Chen & O. J. L. Tzeng (Eds.), *Language processing in Chinese*. Amsterdam: North-Holland.
Taft, M., & Zhu, X. (1999). Positional specificity of radicals in chinese character recognition. *Journal of Memory and Language, 40*, 498–519.
Taylor, I., & Taylor, M. M. (1983). *The psychology of reading*. New York: Academic Press.
Tsai, J. L., & McConkie, G. W. (2003). Where do Chinese readers send their eyes? In J. Hyona, R. Radach & H. Deubel (Eds.), *The Mind's eyes: cognitive and applied aspects of eye movements* (pp 159–176). Amsterdam, Netherlands: North-Holland /Elsevier Science Publishers.
Van Lancker, D., & Fromkin, V. A. (1973). Hemispheric specialization for pitch and tone: Evidence from Thai. *Journal of Phonetics*, *1*, 101–109.
Van Lancker, D., & Fromkin, V. A. (1978). Cerebral dominance for pitch contrasts in tone language speakers and in musically untrained and trained English speakers. *Journal of Phonetics, 6*, 19–23.

Xin Hua Dictionary. (1979). Shanghai: Shang Wu Press.
Xing, H., Shu, H., & Li, P. (2002). A self-organizing connectionist model of character acquisition in Chinese. In W. Gray & C. Schunn (Eds.), *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.
Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America, 95*, 2240–2253.
Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics, 17*, 1–31.
Yang, H. M., & McConkie, G. W. (1999). Reading Chinese: some basic eye movement characteristics. In J. Wang, A. Inhoff, & H. C. Chen (Eds.), *Reading Chinese script: a cognitive analysis* (pp 207–222) Hillsdale, N. J.: Erlbaum.
Yu, B. L., & Cao, H. Q. (1992a). A new exploration on the effect of stroke number in the identification of Chinese characters. *Acta Psychologica Sinica, 24*, 120–126 (in Chinese, with English abstract).
Yu, B. L., & Cao, H.Q. (1992b). The effect of the stroke-number disposition on Chinese character recognition. *Psychological Science, 4*, 5–10 (in Chinese, with English abstract).
Zhang, Y. & Chen, T. (1716). *Kangxi Dictionary* (1716). Japan: Tongwen Bookstore version.
Zhou, X., & Marslen-Wilson, W. (1999). Sublexical processing in reading Chinese. In J. Wang, A. Inhoff & H. Chen (Eds.), *Reading chinese script* (pp. 37–63). Erlbaum: London.
Zhou, Y. G. (1978). Xiandai hanzihong shengpangde biaoyin gongneng wenti [To what degree are the "phonetics" of present-day Chinese characters still phonetic?]. *Zhongguo Yuwenm, 146*, 172–177.
Zipf, G. K. (1932). *Selective studies and the principle of relative frequency in language*. Cambridge, MA: MIT Press.