International Journal of Chinese Language Teaching (2025)
Vol. 6 (2) 63-89 https://doi.org/10.46451/ijclt.20250205

Article

Entropy-based Sound-Character Mapping for Chinese Character Learning

Arthur Berg
Pennsylvania State University, USA

Received: 130 July, 2024/Accepted: 25 October, 2024/Published: 1 November, 2024

Abstract
This study introduces an innovative approach to learning Chinese by leveraging unique sound-character relationships. By employing the concept of entropy in sound-character mappings, we provide a systematic method for identifying and categorizing characters based on their phonetic uniqueness. Our approach specifically targets listening and writing skills, focusing on improving dictation abilities by distinguishing between sounds corresponding to unique characters and those associated with multiple characters. This method not only facilitates accurate character writing but also reinforces correct pronunciation, leading to comprehensive improvement in Chinese language proficiency. By providing quantitative measures of the relationship between pronunciations and characters through entropy calculations and integrating these findings into practical learning tools, this study contributes to a more nuanced understanding of Chinese learning. It offers practical applications for both educators and learners, potentially enhancing teaching effectiveness and learner outcomes.

Keywords
Sound-character mapping, phonological awareness, tone recognition, entropy, educational technology

1 Introduction

Learning Chinese as a second language poses unique challenges due to the complex relationship between its phonological and orthographic systems. Unlike alphabetic languages, where letters correspond to specific sounds in a relatively transparent manner, Chinese characters often share the same pronunciation but represent different meanings, leading to a high degree of homophony (Lee & Huang, 2022). This abundance of homophones, combined with the logographic nature of Chinese writing, creates difficulties in character recognition and production for learners (Hsuan, Tsai, & Stainthorp, 2018).

Phonological awareness–the ability to recognize and manipulate the sound structures of spoken language–is crucial for reading acquisition in any language (Tseng et al., 2023). In Chinese, despite its non-alphabetic script, phonological awareness helps learners distinguish between syllables and tones, which is essential for differentiating homophonous characters (Siok & Fletcher, 2001). However, Chinese orthography poses additional challenges due to its low orthographic transparency.

Email: asb17@psu.edu

Orthographic transparency refers to the consistency and predictability of sound-symbol correspondences in a writing system (Ho, Yao, & Au, 2003). In Chinese, many characters contain phonetic components intended to provide pronunciation cues, but these cues are often inconsistent or misleading. This inconsistency means that learners cannot reliably infer a character's pronunciation from its visual form alone, complicating the development of reading and writing skills (Lin et al., 2019). Moreover, the act of writing Chinese characters engages cognitive processes deeply involved in reading. Cao and Perfetti (2016) found greater neural involvement of writing in Chinese reading than in English reading, highlighting the intricate connection between the physical act of writing and the cognitive processes of reading in Chinese. Similarly, Chai and Ma (2022) demonstrated that character writing proficiency significantly predicts reading ability in second language learners, underscoring the importance of integrating writing practice in Chinese literacy instruction.

The prevalence of homophones further complicates language learning. Multiple characters can share the same pronunciation but have different meanings and written forms. Liu and Wiener (2020) found that while homophones can facilitate lexical development by allowing learners to leverage existing phonological representations, they can also cause confusion. Learners may struggle to distinguish between characters that share the same pronunciation, especially without contextual cues (Wiener, Lee, & Tao, 2019). This cognitive load affects learners' production accuracy, as task complexity and prior knowledge significantly impact their ability to produce new words (Liu & Wiener, 2021).

These challenges—phonological awareness, orthographic transparency, and homophones—are deeply intertwined. Low orthographic transparency hinders the ability to connect phonology with orthography, making it difficult for learners to apply phonological awareness effectively (Tseng et al., 2023). The high degree of homophony exacerbates the issue, as learners encounter many characters sharing identical pronunciations, increasing ambiguity in both listening comprehension and character writing.

To address these challenges, a quantitative approach is needed to assess the ambiguity in sound-character mappings. Entropy, a concept from information theory introduced by Shannon (1948), measures the unpredictability or uncertainty within a system. In the context of Chinese language learning, entropy can quantify the degree of ambiguity associated with a given pronunciation.

By calculating the entropy of pronunciations, we can determine how many possible characters correspond to a specific sound and how evenly distributed their frequencies are. High entropy indicates that a pronunciation maps to many characters with similar frequencies (high ambiguity), while low entropy suggests fewer characters or a dominant character (low ambiguity). For example, consider characters like 我 (wǒ), 能 (néng), 水 (shuǐ), 外 (wài), 怎 (zěn), 放 (fàng), 此 (cǐ), and 改 (gǎi). Their pronunciations are uniquely associated with these specific characters, eliminating ambiguity. Conversely, hearing the pronunciations "yì" or "shí" may lead to ambiguity due to the large number of characters sharing these pronunciations.

This metric provides a clearer understanding of the phonological and orthographic challenges learners face. By categorizing syllables based on entropy, educators can tailor instructional methods, starting with low-entropy (less ambiguous) pronunciations and progressively introducing higher-entropy ones. This approach aligns with the scaffolded learning principles (Lightbown & Spada, 2013) and supports integrating tone learning with vocabulary instruction, enhancing pronunciation and overall language proficiency (Liu & Xiao, 2021).

2 Contributions of the Current Study

This study introduces an entropy-based approach to analyze sound-character mappings in Chinese, offering a systematic method to quantify and categorize pronunciations based on their ambiguity. Our contributions are as follows:

1. **Quantitative Analysis of Ambiguity:** We apply entropy calculations to Chinese pronunciations to measure the uncertainty in sound-character relationships. This analysis provides insights into the extent of homophony and its impact on language learning.
2. **Integration with Educational Tools:** We develop specialized flashcards compatible with the Pleco app, incorporating our entropy findings into practical learning resources. These tools are designed to enhance listening and writing skills by focusing on pronunciations with varying levels of entropy.
3. **Implications for Teaching Strategies:** By categorizing syllables based on entropy, educators can tailor instructional methods, starting with low-entropy pronunciations and progressively introducing higher-entropy ones.

By bridging the gap between theoretical analysis and practical application, our study offers a novel strategy to enhance Chinese language proficiency. The entropy-based method provides a new perspective on addressing the complex interplay between phonology and orthography in Chinese, potentially informing both pedagogical approaches and linguistic research. Notably, this approach resonates with the ideas of Yuen Ren Chao, who emphasized understanding the intricate relationships between sounds and characters in Chinese (Chao, 1968).

The subsequent sections of this paper will detail the methodology, findings, practical applications, and implications of our entropy-based approach. We will present an analysis of the most common pinyin and their associated characters, as well as measures of the entropy of character distributions. Our findings will highlight the most common and highest entropy pinyin, and the practical application section will discuss how these insights are implemented through the Pleco flashcards to enhance Chinese language learning.

3 Methodology

3.1 Frequency concepts

In this study, we employ several interrelated frequency concepts that form the foundation of our analysis. Understanding these concepts is crucial for interpreting our methodology and results:

1. Character Frequency: This refers to how often a specific Chinese character appears in texts or usage. Character frequency is typically expressed as a percentage or relative frequency compared to other characters. For example, common characters like 的 (de) or 是 (shì) appear much more frequently than others.
2. Pinyin/Pronunciation Frequency: This is the raw count or absolute frequency of how often a particular pinyin appears in the corpus, regardless of which characters it represents. For instance, the pinyin “de” is very common as it represents several high-frequency characters.
3. Relative Pinyin Frequency: This is the pinyin frequency expressed as a percentage or proportion of the total pinyin occurrences in the corpus. It allows for comparison of pinyin usage across different datasets or corpus sizes. The relative pinyin frequency is particularly important in our study as it helps learners understand the prevalence of certain sounds in spoken Mandarin. For example, there are many zero-entropy pinyin sounds (those that map to only one character), but their frequencies can vary greatly. Common ones like “wǒ” (我) and “dà” (大) are heard frequently, whereas others like “lǎ” (喇) and “lǒu” (搂) are far less common. By reporting the relative percentage for each pinyin sound, we provide insight into which sounds learners are most likely to encounter in real-world usage.
4. Character-Pinyin Pair Frequency: For polyphonic characters (those with multiple pronunciations), we consider the frequency of each character-pinyin pair separately.

This approach captures the nuanced usage of these characters in different contexts. For instance, the character 行 can be pronounced as “xíng” or “háng”, each with different usage frequencies.

5. Cumulative Frequency: This is the sum of relative frequencies up to and including a given pinyin. It helps in understanding how much of the language can be comprehended by learning the most frequent pinyin sounds and their associated characters.

These frequency concepts form the basis of our entropy calculations and help distinguish between the prevalence of characters and their pronunciations. By considering these different aspects of frequency, we provide a comprehensive analysis of character usage patterns and pronunciation variability in modern Chinese.

The interplay between these frequency measures is crucial for our study. For instance, a pinyin with high relative frequency but high entropy (mapping to many characters) presents different learning challenges compared to a pinyin with low relative frequency but zero entropy (mapping to only one character). Understanding the frequency of specific character-pinyin pairs within polyphonic characters can guide learners in prioritizing the most common usages.

3.2 Data sources

We utilized two primary data sources to analyze the frequency and characteristics of Chinese characters: the Chinese Character Wiki provided by Dong Chinese (https://www.dong-chinese.com/wiki) and character frequency lists compiled by Jun Da (http://lingua.mtsu.edu/chinese-computing).

The Chinese Character Wiki is a free and open-source dictionary that includes a comprehensive range of information on Chinese characters (Olsen, n.d.). This resource covers stroke orders, pronunciations, definitions, examples, origins, and component breakdowns, making it particularly useful for Chinese language learners. It focuses on commonly used characters, avoiding rare and esoteric ones, which enhances its practical value for learners.

The repository of the Chinese Character Wiki database contains 93,846 entries, but after filtering to include only simplified characters with pinyin frequencies, it is reduced to 2,822 characters. This database provides pinyin frequencies (including respective frequencies for polyphonic characters), character components, HSK levels, number of strokes, and frequency of appearance in movies and books. The comprehensive nature of this database makes it an invaluable tool for learners aiming to improve their proficiency in the Chinese language.

The 现代汉语单字字频 (Modern Chinese Character Frequency List), curated by Jun Da (笪骏, 2004), provides a comprehensive character frequency list for modern Chinese. It includes characters along with their pinyin but does not offer the relative proportions for polyphonic characters. The dataset comprises 9,933 characters with details on their associated pinyins, relative frequencies, and English meanings sourced from the CEDICT Chinese-English Dictionary.

Both datasets were utilized and analyzed to achieve a comprehensive understanding of character usage and pronunciation frequency in modern Chinese. By integrating data from the Chinese Character Wiki and Jun Da’s dataset, we aimed to provide a nuanced analysis of character usage patterns and pronunciation variability.

Entropy calculations require pronunciation frequencies of different characters, which is a primary reason for using the Chinese Character Wiki dataset. Although this resource is comprehensive, the 现代汉语单字字频 dataset supplements the results by providing rare characters associated with the given pinyin in parentheses. It is noted that the characters presented in parentheses do not contribute to the entropy calculation as these characters do not have recorded pinyin frequencies in the Chinese Character Wiki dataset.

3.3 Entropy calculations

Entropy is a measure of uncertainty or unpredictability in a system (Shannon, 1948). In this context, we use it to quantify how predictable a character is from its pronunciation. Lower entropy indicates higher predictability, while higher entropy reflects greater ambiguity. For Chinese syllables, entropy can be expressed as:

$$\mathrm{H}(p) = -\sum_{i=1}^{n} P\left(x_i|p\right) log_2 P\left(x_i|p\right)$$

where $\mathrm{P}\left(x_i|p\right)$ is the probability of the *i*-th character given a specific pronunciation *p*, and the sum is over all characters x_i such that $P\left(x_i|p\right) > 0$.

For example, the pronunciation “dǎ” corresponds uniquely to the character 打, resulting in an entropy of 0 because $P(\text{打}|\text{dǎ}) = 1$ and $P(x|\text{dǎ}) = 0$ for all other characters . Conversely, “shí” corresponds to several characters including 十 (ten), 时 (time), and 实 (real), and others, resulting in a positive entropy value.

An entropy value of 1 is equivalent to two equally likely characters. An entropy of 1 can also be obtained with several characters, though not all equally likely. For example, “dào” is associated with the characters 到 , 道 , 倒 , 稻 , 盗 , and 悼 (excluding other very rare characters) with relative frequencies of 10,331, 2,324, 530, 92, 52, and 11, respectively. After normalizing the frequencies so they sum to one, the entropy of “dào” is calculated to be approximately 1. This suggests that the uncertainty associated with mapping a character to “dào”, devoid of context, is equivalent to choosing between two equally likely characters. Although there are six characters associated with “dào”, the character 到 is the most likely, occurring 77% of the time, followed by 道 and 倒 at 17% and 4%, respectively.

More generally, a pinyin associated with an entropy of *n* would be equivalent to having 2^n equally likely characters associated with the respective pronunciation. This quantitative measure allows us to rank pronunciations based on the ambiguity of their character mappings, providing valuable insights for language learners and educators.

By applying this entropy calculation to all pinyin in our dataset, we systematically quantify the predictability of characters based on their pronunciations. This method enables us to identify zero-entropy pinyin, such as “shuǐ”, which map to single characters and present less ambiguity, as well as high-entropy pinyin, such as “shí”, which map to multiple characters and present greater learning challenges.

4 Results and Discussion

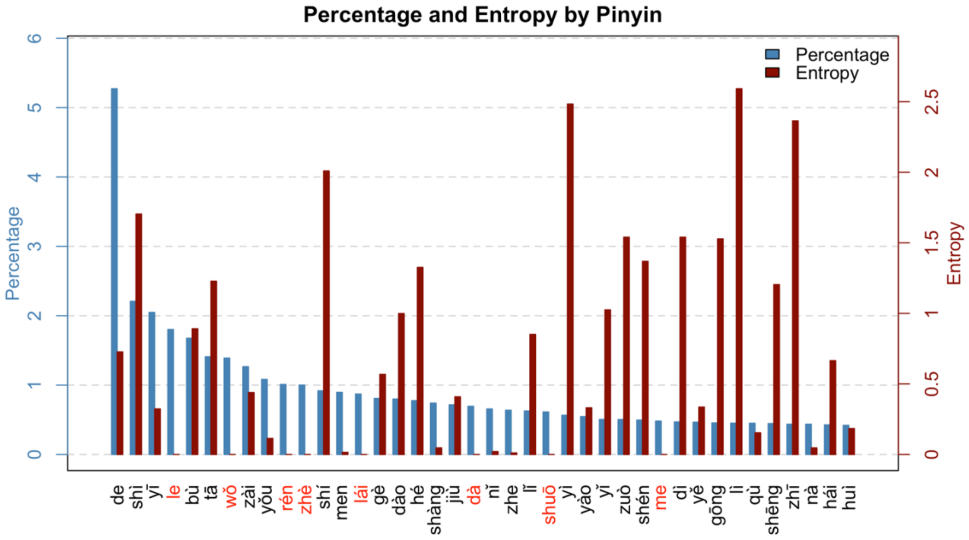
4.1 Entropy analysis of common Pinyin

Our analysis of Chinese character and pinyin frequencies revealed several key insights, presented in three tables and one figure. Table 1 displays the 300 most frequent pinyin along with their associated characters, including rarer characters in parentheses. The table also lists the cumulative percentage of occurrences for each pinyin and its respective entropy value. Including rare characters ensures comprehensive coverage, while cumulative percentages provide insights into the relative commonality of each pinyin. The entropy values offer a quantitative measure of the ambiguity associated with each pinyin, with lower values indicating less ambiguity and higher values reflecting greater uncertainty in mapping a given pinyin to its corresponding character(s).

Figure 1 visually represents the relationship between the frequency of occurrence (percentage) and the entropy of various pinyin in the Chinese language. The graph employs a dual-axis system to display both percentage (left y-axis, dark gray bars) and entropy (right y-axis, light gray bars) for each pinyin along the x-axis. This visualization facilitates a quick comparison between how often a pinyin is used and how ambiguous it is in terms of character mapping. Pinyin with high frequency and high entropy, such as “de” and “shì”, stand out as frequently used sounds with multiple possible character representations. In contrast, pinyin like “le” and “wǒ” show high frequency but low entropy, indicating less ambiguity in their usage. Pinyin with zero entropy are highlighted with slightly darker text labels. This mapping offers a novel way to explore characters and understand the relative abundances of certain pinyin and the number of associated characters, potentially aiding in the development of targeted learning strategies.

Figure 1

Percentage of Occurrence and Entropy Values for Common Pinyin in Chinese



Note: Blue bars represent the percentage of occurrences (left y-axis), while red bars indicate the entropy (right y-axis) for each pinyin.

Table 1

Most Frequent Pinyin, Associated Characters (including rare characters in parentheses), Cumulative Percentages of Occurrences, and Entropy Values.

Pinyin	Character(s)	Cum. %	Entropy
de	的地得（底）	5.273	0.727
shì	是事世市式士示似视势试适室释氏饰侍誓逝拭（轼嗜仕恃噬柿谥舐視弒螫筮適莳釋試铈謚贳眎篩銔襫鈰飾）	7.481	1.704
yī	一医依衣伊（揖醫漪噫壹咿铱猗欹祎黟袆蛜鷖）	9.529	0.323
le	了（饹）	11.331	0.000
bù	不部步布怖埠（簿钚瓿蔀篰踄）	13.007	0.891
tā	他她它踏塌（遢趿铊祂溻）	14.416	1.228
wǒ	我	15.806	0.000
zài	在再载（縡酨）	17.072	0.439
yǒu	有友（黝酉莠牖卣脩铕羑蜏脜苃羗）	18.154	0.114
rén	人（仁壬鵀魜）	19.164	0.000
zhè	这（蔗浙這鹧柘蟅）	20.165	0.000

Pinyin	Character(s)	Cum. %	Entropy
shí	时实十识石食拾蚀（什炻鲥莳識祏埘辻鉐蝕鼫遈鉽）	21.083	2.009
men	们门	21.979	0.015
lái	来（莱涞徕崃铼萊鶆）	22.850	0.000
gè	个各（铬虼硌箇）	23.658	0.568
dào	到道倒盗稻悼（焘纛帱盜稲衜翢翿）	24.457	0.999
hé	和合何河核荷盒（颌禾劾涸阂阖龢纥菏曷貉盍翮饸龁盇鞨粭鶡敆覈鹖礉頜盉）	25.233	1.326
shàng	上尚（绱）	25.974	0.047
jiù	就救旧舅（疚咎臼厩鹫柩僦柏舊鷲）	26.690	0.409
dà	大	27.384	0.000
nǐ	你拟（旎薿譺）	28.040	0.021
zhe	着蔗（著）	28.678	0.011
lǐ	里理礼鲤（李哩蠡俚澧锂醴娌逦裡鳢悝鱧裏粴禮鋰）	29.305	0.851
shuō	说（說）	29.918	0.000
yì	意义议易益异艺亦亿译役翼忆抑疫毅谊屹（衣逸溢裔懿绎奕邑诣驿翌臆佚轶熠弋弈翊呓蜴薏刈羿缢翳镒峄悒肄挹瘗仡義議怿佾瘞埸劓鎰袘镱殪睪齸詣蘙藙鮨鹝絏藝譯蛡鶂褹裛耴輗饐豷鷊鷁芅謚鐿囈贀驛螠）	30.484	2.483
yào	要药耀钥（藥鹞疟曜籥艞靿葯窔矅箹鑰）	31.030	0.331
yǐ	以已椅乙蚁倚（矣迤旖苡钇锜螘顗齮蟻飲艤舣阤肊釔礒）	31.535	1.025
zuò	作做坐座（凿唑酢祚柞胙怍阼葄）	32.039	1.540
shén	什神甚（鰰）	32.535	1.369
me	么（麼）	33.017	0.000
dì	的地第帝弟递缔（蒂谛棣娣睇碲遞禘菂締釱腣遆諦祶）	33.487	1.540
yě	也野冶（虵）	33.953	0.336
gōng	公工功供攻宫弓躬（蚣恭龚觥肱魟碽龔）	34.408	1.528
lì	力利立历例丽厉励粒隶砾沥荔（莉吏栗笠雳俐痢戾蛎詈俪栎砺莅郦傈枥跞唳粝疠呖溧苈猁疬轹篥坜麗蘽隸涖脷苙磿綟讈蔍蠣赲）	34.859	2.592
qù	去趣（觑阒闃闐）	35.309	0.154
shēng	生声升牲（胜甥笙聲陞苼鼪鍟）	35.755	1.204
zhī	之只知指支织枝芝脂肢汁蜘（祗胝卮栀織隻鴲胑衼禔鼅鳷禵）	36.191	2.364
nà	那纳呐（娜钠捺衲肭納靹衲鈉箬）	36.627	0.047
hái	还孩（骸還）	37.054	0.665
huì	会汇慧绘（惠贿讳晦秽卉诲彗恚喙荟蕙烩蟪缋翙浍阓颒篲譓詯賄諱鉞薉鐬譓）	37.474	0.184
jiàn	间见建件舰剑渐健键箭践鉴荐贱溅（监槛谏僭涧饯腱見毽鑑鍵踐艦薦踺楗瞯諓磵鋻趝瞷瑐）	37.893	2.402
zi	子字	38.312	0.140
zhǔ	主嘱煮拄（属瞩渚麈詝矚砫）	38.730	0.128
xià	下夏吓（厦罅）	39.135	0.330
jiā	家加佳夹茄（挟嘉迦枷袈痂浃珈跏笳葭镓筴麚豭貑服）	39.536	0.997

Pinyin	Character(s)	Cum. %	Entropy
xiàn	现见线限县献陷宪羡羡馅（腺霰苋岘線粯軐県晛縣）	39.936	1.949
wèi	为位未卫味谓慰胃喂畏（猬魏尉蔚渭鳚衛謂霨蝟苿躗硙鏏軎餵磑藯躛錗鮇）	40.335	1.961
duì	对队（兑怼碓隊憝鐓錞镦祋銳陮）	40.723	0.773
guó	国（帼虢掴腘馘膕聝）	41.107	0.000
chéng	成程城承盛诚乘呈惩（澄丞橙裎枨铖塍酲埕郕脭誠絾琤）	41.484	1.466
kě	可渴（坷岢）	41.849	0.066
méi	没梅眉煤枚霉徽酶（媒玫湄嵋楣莓镅鹛郿猸矀蘪鋂禖鎇）	42.214	0.588
hǎo	好（郝）	42.577	0.000
kàn	看（瞰阚磡矙）	42.939	0.000
jì	系计记济技际纪继既季剂寄寂（迹绩祭忌冀妓伎悸暨骥稷髻鲫偈蓟觊霁芰荠鲚計跽繼記洎際紀罽蹟繫穊跡哜鰶臮蟿茍訐穄繫鵋驥鯽鰿鱀）	43.299	3.225
qǐ	起企启岂（稽乞绮杞芑綮豈屺棨邔）	43.657	0.396
jī	机几基击激积迹鸡绩肌饥圾讥（奇玑稽姬畸缉叽矶羁唧跻嵇箕畿乩犄芨屐咭赍齑笄積墼谿雞飢剞踦齏齎績羇鄿虀觭羈鐖禨稘緝羈磯機賷）	44.009	2.624
dōu	都兜（蔸篼）	44.352	0.064
zhōng	中终钟忠（衷盅锺忪螽舯終鐘鈡蹱鈡）	44.693	0.765
xué	学（穴踅泶袕鸴鷽）	45.031	0.000
duō	多哆（咄掇裰）	45.366	0.048
néng	能	45.701	0.000
nián	年黏（粘鲇鲶鮎）	46.035	0.050
zhèng	正政証证郑症挣（帧诤證）	46.367	1.798
xiǎo	小晓（筱篠筊謏）	46.697	0.113
xiǎng	想响享（饷飨響鲞餉饟鯗）	47.027	0.698
xīn	心新辛欣薪芯锌（馨鑫忻歆莘昕鋅）	47.355	1.203
yòu	又右幼诱佑（釉祐柚囿宥蚴鼬侑誘褎褏）	47.683	0.570
huà	话化划画桦（华話繣）	48.008	1.680
dòng	动洞冻（栋恫侗峒胴胨硐霘衕垌詷）	48.331	0.425
jǐ	己给几挤（脊戟麂虮鱾掎）	48.652	1.269
zì	自字（渍恣眦眥胔胾）	48.972	0.548
jìn	进近尽禁劲浸（晋烬靳噤荩觐缙妗進盡赆賮齽祲）	49.291	1.282
bǎ	把（靶钯）	49.610	0.000
tiān	天添（黇）	49.925	0.139
zhǐ	只指止纸址（旨趾徵咫酯芷祉枳阯黹紙轵觗茝絺藢）	50.240	1.561
guò	过（過）	50.550	0.000
zhì	制至治质置智秩掷稚帜（识致志滞挚峙窒炙痔痣蛭郅觯雉栉桎質鸷帙贽陟骘彘轾踬製忮誌铚袠胵芖紩鑕猘秲緻鷙跱銡袟锧稙覟贄鷙隲鴙）	50.853	2.305

Pinyin	Character(s)	Cum. %	Entropy
yuán	原员元源园圆援缘猿（袁垣辕媛沅爰黿圜芫螈塬橼緣鈨贠鶢鼋緑褑薗笎）	51.151	2.384
yòng	用（佣）	51.449	0.000
ba	吧巴爸罢拔叭笆	51.738	1.640
fā	发（髮）	52.027	0.000
jí	及即集级急吉疾辑籍脊（吃极藉嫉棘汲亟笈瘠岌楫芨蒺嵴佶殛戢級鹡蕺蹐鶺脨踖谻蝍緝箿鞊趌）	52.315	2.313
mín	民（岷缗珉玟苠錕旼暋）	52.595	0.000
yàng	样漾（恙怏羕詇）	52.874	0.035
ne	呢呐	53.150	1.000
jiào	教觉叫较轿窖酵（校醮較峤覺徼轎噍藠斠訆覚）	53.425	1.728
qián	前钱潜钳（乾虔黔荨揵钤箝錢蔵鍼軡鰬）	53.694	0.787
shù	数术述树束竖朮（恕墅庶漱戍術澍腧沭豎裋荗鶐）	53.958	2.346
xíng	行形型刑（邢硎饧荥陉鉶）	54.221	1.287
yú	于鱼渔愚愉與娱竽（与予余於逾瑜虞禺俞榆隅渝欤谀盂馀覦腴臾揄畬萸嵛窬顒蝓餘颙雩狳舁好魚隃邘諛輿玃舒鱮芋釪羭）	54.471	0.864
chǎn	产阐铲（谄蒇蕆諂鏟闡闡骣）	54.720	0.113
jīng	经精惊睛晶鲸茎腈（京荆兢菁經旌泾粳驚莖鶄麖秔荊鯨）	54.966	1.408
shè	社设射涉舍摄（慑赦麝歙厍設蔎滠騇）	55.211	1.544
dài	大代带待戴袋逮（贷黛怠殆岱迨玳甙骀绐埭轪艜襶簤貸魼）	55.455	1.938
ér	而儿（鸸粫鲕輀胹）	55.698	0.592
wéi	为维围唯违惟桅（韦帷圩闱潍嵬帏維濻鮠鍏鄬違艜沩艜）	55.940	1.798
diǎn	点典踮（碘點蒧）	56.175	0.192
shǐ	使始史驶屎（矢豕駛鉂）	56.407	1.535
zhàn	战站占蘸（颤绽湛栈菚）	56.639	1.272
rán	然燃（髯蚺髥袡肰）	56.868	0.371
cóng	从丛（淙琮賨誴）	57.094	0.132
xiē	些歇（楔蝎蠍）	57.320	0.154
hěn	很狠（詪）	57.545	0.208
qì	气器弃汽泣砌（妻契迄亟憩讫碛槭葺碶汔礤碱鼜）	57.769	1.525
jiē	结接阶街皆揭（节偕秸嗟疖節階喈祖腱稭萋）	57.993	1.650
xiàng	相像象项巷橡（向項蟓鱌）	58.214	1.011
jiān	间坚监尖肩兼艰歼奸煎（渐浅笺缄鞯間菅犍缣戋湔鹣鞬戋蒹搛鰜鳽閒钘蕑鲣監麗鰹鳒豎蔄鐗鈃湕瀐箋）	58.433	2.047
lǎo	老（姥佬潦铑栳蛯荖）	58.648	0.000
wù	物务恶误悟雾（勿晤兀坞戊鋈骛鹜婺寤焐芴机誤軏霚痦霧阢鶩）	58.860	1.454
zhǒng	种肿（冢踵種腫）	59.070	0.059
kāi	开揩（開锎）	59.279	0.033
biàn	变便遍辩辨辫（汴卞弁變苄缏辯忭覍緶艑）	59.486	1.729
yè	业夜叶页液咽（拽曳谒腋掖邺晔烨靥葉頁鐷緤鍱）	59.692	1.576
quán	全权泉拳（痊蜷诠荃虇铨醛鬈筌鰁鳈詮譔絟辁硂葲銓顴）	59.896	0.804

Pinyin	Character(s)	Cum. %	Entropy
shǒu	手首守（艏）	60.095	0.938
zhòng	中种重众（仲眾諥茽）	60.294	1.421
tóu	头投（骰頭緰）	60.493	0.400
shēn	身深参申伸绅呻（娠莘砷诜糁鲹蓡詵蔘鯵籸紳葠鯓）	60.691	1.376
r	儿	60.886	0.000
tí	提题蹄啼（題缇绨鹈醍荑鳀虒遆稺趧綈鶗騠緹磃蕛）	61.080	1.069
liǎng	两（俩魉裲蜽）	61.273	0.000
cháng	长常场偿尝肠（裳嫦長苌腸徜鲿）	61.465	1.443
zǒu	走（鯐）	61.655	0.000
bèi	被备背贝倍辈狈（惫悖钡蓓焙孛碚鞴鐾褙貝誖邶骳輩鐴鋇）	61.842	1.936
gāo	高糕膏羔（睾皋篙槔鷎皐韟）	62.028	0.141
dàn	但弹担淡旦蛋诞氮（石惮澹啖萏瘅霮禫蜑餤駳誕贉髧）	62.213	1.338
guān	关观官棺（冠倌莞關鰥觀蒄窤）	62.397	1.327
sān	三（叁毵彡鬖糂）	62.578	0.000
yóu	由游油尤犹邮铀（疣鱿猷莜莸繇蝣蚰尢輶鲉莤蝤遊鈾蚘郵鮋蓧）	62.759	1.750
huí	回（蛔洄茴迴鮰）	62.936	0.000
jù	据具句剧巨距聚拒惧俱锯（瞿炬踞遽飓钜苣倨讵醵窭虡屦鐻椇秬鉅鋸簴粔）	63.109	2.610
yuè	月越乐阅跃悦（钥岳粤樾刖钺閲龠瀹躍篗趯軏粵礿躒）	63.282	1.407
jué	决觉绝角掘嚼（脚爵厥诀崛倔抉攫獗蕨蹶谲橛珏噱矍镢桷劂孓絕爝钁觖觼蕝蹷玃芵訣蟨穱絶躩）	63.454	1.464
gěi	给（給）	63.625	0.000
wèn	问（紊汶璺顐）	63.795	0.000
cái	才财材裁（財）	63.965	1.095
shuǐ	水	64.134	0.000
dìng	定订锭（钉铤啶碇腚訂釘饤碠磸錠）	64.301	0.308
fāng	方妨（坊芳枋匚钫邡）	64.468	0.043
yán	言研严延沿炎岩颜盐檐（癌阎蜒筵妍闫硏顏鹽阽埏綖麣簷閻莚訮）	64.635	2.370
zhù	住助筑驻祝柱铸蛀（着注著贮伫杼箸炷苎翥纻貯跓麈疰築苧竚紵鑄駐羜註祩）	64.801	1.648
gēn	根跟	64.967	0.991
suǒ	所索锁琐（唢鎖鏁）	65.132	0.515
dǎng	党挡（谠黨讜）	65.296	0.196
yīn	因音阴姻（殷荫茵湮氤喑陰洇堙铟骃禋秵絪闉駰裀霒銦蒑陻）	65.459	1.129
míng	明名鸣（铭冥茗瞑溟螟暝鳴眳銘眀鄍）	65.622	1.011
èr	二（贰佴貳）	65.785	0.000
wǔ	五武午舞侮捂（伍鹉妩庑忤迕怃仵牾膴碔）	65.948	1.481
qīng	清轻青倾氢蜻（卿輕鲭鶄圊鯖）	66.111	1.915
mìng	命（詺）	66.273	0.000
shi	是事实式识士视势食拾匙（鍦）	66.433	2.254

Pinyin	Character(s)	Cum. %	Entropy
fù	复父负富副付附妇腹赴缚（傅咐赋覆阜驸蝮馥讣鲋福赙負袝賦袝鰒縛訃鍑複縳）	66.591	2.969
yǎn	眼演掩衍（奄俨偃魇兖鼹琰扊郾罨剡龂龑蝘菴鄢黡齗鼴黶）	66.749	0.780
fēn	分纷氛吩（芬酚玢雰紛翂鈖）	66.906	0.526
gé	格革隔骼（蛤阁葛阖嗝镉搿鬲膈鬲閣詥轕裓塥鎘）	67.062	0.883
lù	路陆露录鹿碌（禄赂戮麓漉璐辘箓潞鹭渌逯蓼辂陸簏錄蕗菉盝録騄祿簬逯騼膟騄睩稑醁賂簶鷺）	67.218	1.209
zhēn	真针珍侦（贞斟臻帧桢祯甄箴砧榛針胗椹溱蓁鈫禛鎭鱵貞禎眞）	67.372	0.906
sì	四似饲（食伺寺肆嗣祀巳俟泗笥姒驷汜耜兕覗飼蕼竢禩）	67.524	0.733
bàn	办半伴扮瓣拌（绊辦跘絆）	67.675	1.434
kuài	会快块筷（脍侩狯哙浍鲙郐鄶駃）	67.826	1.091
rèn	任认韧（刃妊纫任恁仞衽認轫葚訒韌袵讱靭餁纴紝軔飪紝軔）	67.975	1.036
dāng	当（铛裆簹）	68.124	0.000
děng	等（戥）	68.272	0.000
ma	吗妈麻嘛蟆（么）	68.418	1.371
zhí	直指值职执植殖（侄蛰踯摭跖蹠縶埴職銕禃膱）	68.563	2.111
xiān	先鲜仙纤掀（酰暹锨跹籼氙祆莶鮮纖縿韱鱻）	68.708	1.013
qí	其奇齐骑旗棋崎（只歧祈鳍琪琦祁祺耆脐岐淇芪麒畦蛴圻颀祇蘄綦亓荠骐萁臍碁蜞饑鯕鵸跂齊軝騎蠐褀鬐魌蚑竒鰭艩粸鳍）	68.850	1.777
jīn	金今禁津斤筋巾襟（矜钅衿觔砛）	68.992	1.920
xìng	性兴幸姓（行杏悻荇興譽莕）	69.133	1.606
xī	西息希吸析悉惜稀牺夕锡溪晰膝嘻熄犀蟋（昔栖熙兮嬉奚螅曦熹蹊羲汐烯蜥晳醯唏淅僖硒歙窸翕浠矽舾穸欷樨郗粞菥豨鼷訢鸂鍚磎糦鏭睎礂鑴俙翖）	69.273	3.276
rú	如蠕（儒茹獳嚅濡薷铷襦颥顬袽駑）	69.413	0.071
biān	边编鞭蝙（砭笾鳊煸邊編邉籩）	69.552	0.641
běn	本（苯畚栦）	69.691	0.000
zuì	最罪醉（蕞）	69.830	0.525
píng	平评凭瓶屏苹（萍坪鲆枰評聠蛢荓缾）	69.966	1.534
jūn	军均君菌（钧筠麇皲軍麏鮶袀硱麐覠莙）	70.101	0.929
dǎ	打	70.236	0.000
fēng	风封丰峰疯锋蜂（枫烽沣酆風葑砜豐鄷犎碸鋒豊）	70.370	2.077
shū	书输殊叔舒疏枢梳蔬（淑倏抒纾菽殳姝摅輸毹紓鵨綀）	70.504	1.808
wài	外	70.638	0.000
zhǎng	长掌涨（鏱仉鞝）	70.772	0.749
shī	师失施诗尸湿狮（虱蓍絁邿詩鍦葹鰤鶳鮖蝨）	70.905	1.958
diàn	电店殿垫奠淀佃惦（甸玷癜钿靛簟電阽坫蜔鈿磹）	71.038	1.061
qī	期七妻欺漆凄凄沏（溪戚栖缉蹊嘁萋桤桼碕郪諆鶈）	71.170	1.647
gǎn	感敢赶杆秆（橄擀鳡簳澉鱤趕稈）	71.302	1.695
xiào	笑效校肖啸（孝詨）	71.433	1.442
jiǔ	九久酒（灸韭玖紤韮）	71.564	1.234

Pinyin	Character(s)	Cum. %	Entropy
jié	结节杰洁截捷竭睫（桔劫诘颉桀偈拮孑碣婕羯結讦疖絜蛣鲒蕺蜐詰）	71.694	1.470
fǎn	反返（攵）	71.823	0.154
bìng	并病（摒竝）	71.952	0.927
ge	个格哥歌搁	72.081	0.655
bǐ	比笔彼鄙（匕俾吡妣筆秕舭聛貏枈粃）	72.210	0.821
què	却确雀（鹊阙榷阕碻悫鵲闕）	72.337	1.092
wén	文闻纹蚊（雯阌玟聞闅閿鼤紋閺螡魰）	72.463	0.599
fǎ	法（砝）	72.588	0.000
zěn	怎	72.713	0.000
tīng	听厅（汀烃聽綎鞓聴）	72.838	0.186
jing	经静晴	72.962	1.020
sī	斯司思私丝撕嘶（厮咝蛳锶鸶缌澌鷉絲鼶颸禠鷥緦禗）	73.086	1.743
hou	候	73.209	0.000
fàng	放	73.331	0.000
bié	别（蹩穪莂）	73.453	0.000
jiě	解姐（飷）	73.575	0.431
zhēng	正争挣睁怔蒸（症征铮筝狰徵峥钲箏篜）	73.696	0.930
wú	无蜈（吴吾毋芜梧浯鼯鹀禑铻莫祦莁）	73.816	0.049
yǔ	与语予雨宇羽屿（禹與齬俁庾圄窳伛語圉瘐貐頨蝺萸祤藇）	73.936	1.936
xiāng	相香乡箱厢镶（湘襄骧芗缃鄉葙纕蘘緗鑲）	74.056	1.580
dǎo	导倒岛蹈捣（祷禱隝）	74.175	1.237
wàng	望往忘妄旺（盳迋）	74.293	1.568
bì	必避毕币闭壁臂蔽碧毙痹痺（比泌辟弊陛庇婢敝璧弼裨愎贲蓖跸毖哔薜嬖畀铋祕篦睥髀濞閉蓽襞荜筚狴禆鷝邲閟躄滗苾庳觱诐箅鵯綼祕蜌鼊鴓詖髲篳畢肶繴贔鞞鉍駜鷩縪驆鮅）	74.409	2.301
xi	西系息	74.525	1.508
liàng	量亮辆谅晾（踉靓諒）	74.640	1.367
cì	次刺伺（赐莿賜）	74.754	0.516
chē	车（砗車）	74.867	0.000
dù	度杜渡肚镀（妒蠹芏詑鍍秺）	74.980	1.016
kē	科颗棵磕瞌蝌（柯苛珂轲窠嗑颏髁稞疴蚵薖顆钶窼跒邁軻頦）	75.093	1.108
dōng	东冬（咚氡鸫岽鶇蝀笗鶫菄）	75.205	0.598
tiáo	条调（迢笤龆苕蓚髫鲦蜩鋚鰷蓨鯈）	75.316	0.514
bǎi	百摆柏（佰捭襬）	75.427	0.686
lián	联连怜廉帘镰（莲涟濂臁鲢裢蠊奁連鐮蓮聯簾鎌鬑聨）	75.537	1.474
nán	难南男喃（楠難）	75.646	1.529
xì	系细戏隙（夕阋翕饩細禊舄虩绤钑鬩餼）	75.754	1.514
jìng	境竟静竞敬镜径净（劲靖痉胫迳靓鏡婧獍靜競脛逕竸）	75.862	2.721
jiè	界介借届戒诫（解藉芥疥蚧骱褯誡衸蚧）	75.969	1.344
wán	完玩顽丸（烷芄纨蚖貦）	76.076	0.795

Pinyin	Character(s)	Cum. %	Entropy
shān	山扇衫珊杉（删栅煽姗跚苫潸舢芟膻钐纔埏羶釤鯅龏）	76.182	0.331
tài	太态汰（泰钛肽酞鈦粏舦）	76.288	0.796
qíng	情晴（擎氰檠黥）	76.393	0.217
huó	活（和）	76.498	0.000
tǐ	体（袛體）	76.603	0.000
liú	流留榴硫（刘瘤浏琉遛馏镏鎏旒骝鰡飗駵飀鹠藰驑）	76.707	1.188
chī	吃哧蚩（痴嗤笞魑媸螭鸱郗眵鴟鵄）	76.811	0.232
xí	习席袭媳（锡褶檄習隰觋郋襲鰼騽霫蓆薂鎴）	76.915	1.313
kè	克客刻课（恪嗑缂氪溘锞課骒艐袔碦礊騍）	77.019	1.850
měi	美每镁（浼羙）	77.122	0.906
yù	与育预域遇玉欲愈御狱誉郁豫裕吁寓（语谷喻浴谕毓蔚驭聿煜芋峪熨钰昱阈妪鹬饫鬻鹆蜮聿預穀燠遹鬱蓣馭矞禦鋊鳿轝礜譽軉鴥籲狗鈺閾籞隩鐍驈衙裒錥礇蕷瑀）	77.224	3.128
guǒ	果裹（猓馃椁蜾綶菓粿）	77.326	0.179
dao	到道	77.426	0.086
bào	报暴抱爆豹（瀑刨鲍趵菢袌）	77.526	1.606
kǒu	口	77.625	0.000
huò	或获货惑祸（和豁霍蠖藿嚯镬禍貨臛靃鑊）	77.723	1.273
fēi	非飞啡（菲妃绯扉蜚霏鲱飛鯡騑裶緋）	77.821	1.026
guāng	光（胱咣銧桄硄趪）	77.919	0.000
mén	门（扪钔門穈虋）	78.017	0.000
chuán	传船（椽遄舡）	78.114	0.987
fú	服福佛符伏幅浮扶俘辐蝠凫（缚袱弗拂芙孚氟匐苻茯蚨郛芾涪砩菔罘蜉怫莩桴绂祓黻複幞绋鮄葍綍艴虙袚韍輻）	78.211	2.912
xū	需须虚吁嘘（墟顼胥戌圩虛須繻盱魆谞諝譃虗魖鬚蕦頊）	78.307	1.380
zǒng	总（偬總蓯総蓗）	78.403	0.000
lǐng	领岭（領）	78.499	0.260
dé	得德（锝）	78.594	0.331
nín	您（恁）	78.689	0.000
jiāng	将江疆僵姜浆缰（豇薑礓鳉螿鱂）	78.784	1.139
jǐn	尽仅紧谨锦（瑾馑槿堇卺緊廑謹堇菫錦）	78.879	1.592
gāng	刚钢岗纲冈缸（扛杠肛罡綱鋼鎠釭）	78.973	1.531
dòu	斗豆逗（读窦痘饾餖竇鬪脰荳）	79.067	0.683
wěi	委伟尾伪纬苇（唯萎娓猥诿痿炜隗玮韪艉鲔洧頠薳緯葦骫磈硊趡骫蔿諉蜼鮪）	79.159	1.816
zī	资姿滋（仔兹咨吱孜淄辎谘龇髭孳缁粢赀锱訾嵫資趑鲻茲鯔菑觜趦茈薋菑粢諮鎡貲禌鄑齜）	79.251	0.328
zào	造躁燥灶噪皂（唣譟趮艁竈）	79.342	0.663
zhǎn	展盏崭（斩辗搌飐盞醆）	79.433	0.363
mù	目木幕牧穆墓慕（募暮睦沐苜钼仫雮霂）	79.523	1.774
chǎng	场厂（敞昶氅惝鋹）	79.613	0.999

Pinyin	Character(s)	Cum. %	Entropy
zǔ	组祖阻（诅俎組詛）	79.702	1.246
lùn	论（論）	79.790	0.000
xǔ	许（栩诩許糈鄦醑訏盨）	79.878	0.000
gàn	干（赣淦绀旰紺）	79.966	0.000
dí	的敌笛涤嘀（迪狄嫡翟荻籴镝觌踧篴糴靮鬄）	80.053	0.583
zhě	者（褶赭锗）	80.140	0.000
nèi	内（那）	80.226	0.000
shòu	受授售兽寿瘦（狩绶鏉醻膄）	80.311	1.303
gòng	共供贡（貢）	80.395	0.427
gèng	更	80.479	0.000
nóng	农浓（侬脓哝膿秾農醲禯）	80.563	0.413
qīn	亲侵钦（衾骎駸親鋟綅嵚）	80.646	0.866
àn	案按暗岸黯（胺犴豻）	80.728	2.014
huǒ	火伙（夥钬）	80.809	0.876
yīng	应英鹰婴（樱莺瑛鹦膺缨嘤罂璎撄譍鷹锳軈譻甇）	80.890	0.962
gòu	构够购（勾垢媾诟彀覯遘購詬雊詢覯）	80.971	0.971
wàn	万（玩腕萬翫鄤脕輐）	81.051	0.000
sù	诉速素肃宿塑（溯粟愫簌谡夙嗉僳涑觫蔌骕餗訴肅遡驌藗鷫粛膆）	81.130	1.994
zǐ	子仔紫姊籽（梓滓笫訾秭茈）	81.209	0.820
qū	区趋驱屈躯岖蛆（曲蛐袪诎匚駿麴祛趨胠驅麹麯軀鰸）	81.288	0.584
duàn	断段锻缎（椴煅鍛緞葮）	81.366	1.330
mā	妈抹（蚂嬷）	81.444	0.065
li	里理力利李哩狸	81.521	1.831
yáng	阳洋杨扬羊（疡佯炀陽烊蛘徉飏諹钖暘颺鸉鍚）	81.598	2.091
fu	服夫负付妇腐傅咐甫袱	81.675	2.531
fèn	分份奋愤粪忿（偾糞膹鲼）	81.752	2.162
qiú	求球（仇囚裘酋虬泅遒俅逑鰽巯犰蝤赇虯絿蛷銶訅肍觓）	81.828	0.942
rì	日（鈤）	81.904	0.000
jiū	究纠揪啾（鸠赳阄鬏糺鳩）	81.980	0.654
jiǎn	简检减剪捡碱俭拣茧（柬睑锏翦笕謇蹇戬硷裥趼簡囝谫鹼鹻枧瞼 鬋醎蠒）	82.056	2.309
gǔ	古股骨谷鼓钴（滑贾蛊鹄汩鹘毂诂牯嘏罟瞽臌愲脵鹽穀）	82.131	2.264
xiě	写血（藛）	82.205	0.310
qiáng	强墙（蔷樯嫱薔艢）	82.279	0.747
bǎo	保宝堡饱（葆褓鸨飽賲蕵）	82.353	1.225
zhāng	张章樟（彰璋漳蟑鄣獐嫜餦麞粻鏱蔁騿）	82.426	1.023
ài	爱碍隘（艾唉暧嗳嗌砹嫒瑷礙鴱靉賹鑀薆）	82.499	0.412
guo	过	82.572	0.000
guǎn	管馆（莞錧筦館輨錧）	82.645	0.716

4.2 Zero-Entropy Pinyin

Table 2 provides a detailed examination of practically zero-entropy pinyin, filtered to include only pinyin sounds with a frequency of at least 1%, resulting in a list of 213 pinyin. This threshold focuses on the most common and practically relevant pinyin for language learners. These pinyin are calculated to have an entropy of zero based on data from the Chinese Character Wiki, indicating a one-to-one correspondence between the pinyin and a single character. This makes them highly predictable and unambiguous in their usage. To ensure comprehensive coverage, the list includes rare characters from the 现代汉语单字字频 dataset, provided in parentheses, where no frequencies are available in the Chinese Character Wiki dataset. While some pinyin may not be strictly zero-entropy due to the presence of these rare additional characters, they are considered “practically” zero-entropy from the language learner’s perspective.

These zero-entropy pinyin offer an excellent starting point for educators designing Chinese language curricula. Introducing these unambiguous sound-character pairs early in the learning process allows students to build confidence and establish a strong foundation for more complex character recognition tasks. This approach aligns with instructional strategies that emphasize early successes to motivate learners (Lightbown & Spada, 2013).

Table 2

Practically Zero-Entropy Pinyin with a Frequency of at Least 0.01%

Pinyin	Character(s)	Percentage
le	了（饹）	1.802
wǒ	我	1.390
rén	人（仁壬鵀魜）	1.010
zhè	这（蔗浙這鷓柘蟅）	1.001
lái	来（莱涞徕崃铼萊鶆）	0.871
dà	大	0.694
shuō	说（說）	0.613
me	么（麽）	0.482
guó	国（帼虢掴腘馘膕聝）	0.384
hǎo	好（郝）	0.363
kàn	看（瞰阚磡矙）	0.362
xué	学（穴踅泶袕鸴鷽）	0.338
néng	能	0.335
bǎ	把（靶钯）	0.319
guò	过（過）	0.310
yòng	用（佣）	0.298
fā	发（髮）	0.289
mín	民（岷缗珉玟苠鍲旼䁕）	0.280
lǎo	老（姥佬潦铑栳蛯荖）	0.215
r	儿	0.195
liǎng	两（俩魉裲蜽）	0.193
zǒu	走（鯐）	0.190
sān	三（叁毵彡鬖糂）	0.181
huí	回（蛔洄茴迴鮰）	0.177
gěi	给（給）	0.171
wèn	问（紊汶璺顐）	0.170
shuǐ	水	0.169
èr	二（贰佴貳）	0.163
mìng	命（詺）	0.162
dāng	当（铛裆簹）	0.149
děng	等（戥）	0.148
běn	本（苯畚翉）	0.139
dǎ	打	0.135
wài	外	0.134
fǎ	法（砝）	0.125
zěn	怎	0.125
hou	候	0.123
fàng	放	0.122
bié	别（蹩襒莂）	0.122
chē	车（砗車）	0.113
huó	活（和）	0.105
tǐ	体（袛體）	0.105
kǒu	口	0.099
guāng	光（胱咣銧桄硄趪）	0.098

Pinyin	Character(s)	Percentage
mén	门（扪钔門糜虋）	0.098
zǒng	总（偬總蓯総蓗）	0.096
nín	您（恁）	0.095
lùn	论（論）	0.088
xǔ	许（栩诩許糈鄦醑訏盨）	0.088
gàn	干（赣淦绀旰紺）	0.088
zhě	者（褶赭锗）	0.087
nèi	内（那）	0.086
gèng	更	0.084
wàn	万（玩腕萬翫鄤脕輐）	0.080
rì	日（鈤）	0.076
guo	过	0.073
shǎo	少	0.071
nǎ	哪（那）	0.068
nǚ	女（钕）	0.068
nǚ	女（钕）	0.068
tōng	通（嗵）	0.067
ná	拿（镎拏誽秅）	0.067
cǐ	此	0.067
a	啊（阿）	0.066
tè	特（忒忑慝螣铽脦貣）	0.066
gǎi	改	0.065
shuí	谁（誰脽）	0.064
gāi	该（赅垓陔該荄賅）	0.063
bái	白	0.063
qiě	且	0.063
yuǎn	远（遠）	0.062
liù	六（陆遛馏鹨霤飀）	0.062
mǎn	满（螨）	0.062
ya	呀	0.061
rè	热	0.061
gào	告（诰锆郜誥祮禞祰）	0.060
pǎo	跑	0.059
ràng	让（讓）	0.058
sǐ	死	0.056
liǎn	脸（敛琏裣臉羷）	0.056
hǎi	海（醢胲酼）	0.055
zhǔn	准	0.054
zhěng	整（拯）	0.054
tuán	团（抟蓴糰鱄）	0.053
pǐn	品（榀）	0.052
kōng	空（崆倥箜躻䳍）	0.051
cūn	村（邨皴）	0.048
mǐ	米（眯靡弭敉芈脒粎葞）	0.047
rù	入（褥缛洳溽蓐鳰込）	0.047
mǎi	买（買荬蕒）	0.046
shōu	收	0.046
zhuǎn	转（轉転）	0.045
zán	咱	0.044
tiě	铁（帖鐵驖）	0.044
běi	北	0.043
tou	头	0.043
ye	爷	0.040
kǔ	苦	0.039
děi	得	0.038
guǎng	广（犷）	0.038
huài	坏	0.037
cǎo	草（艸）	0.037
zuǐ	嘴（觜）	0.037
zēng	增（曾憎缯罾矰譄鄫繒磳）	0.036
suí	随（遂绥隋隨綏）	0.036
chuān	穿（川氚巛）	0.036
bu	不	0.033
su	诉	0.033
zuǒ	左（佐）	0.033
zháo	着	0.033
suī	虽（尿睢濉荽雖眭鞖簑）	0.033
tuī	推（忒蘈蓷）	0.032
kào	靠（铐犒鮳）	0.031
wēn	温（瘟蕰蕴鳁缊辒緼薀縕）	0.030
qing	情	0.029
xuě	雪（鳕）	0.028
hǎn	喊（罕蔊檻）	0.028
cún	存	0.028
lěng	冷	0.027
dǐng	顶（鼎酊頂艼鐤鼑）	0.027

Pinyin	Character(s)	Percentage
you	友	0.026
chuàng	创（怆）	0.026
qin	亲	0.026
guài	怪	0.025
shěng	省（眚）	0.024
chū	初（出樗貙齣）	0.024
zhuā	抓（挝膼髽）	0.024
kùn	困（睏）	0.023
shú	熟（赎孰塾秫）	0.023
chuáng	床（幢）	0.023
réng	仍（礽）	0.023
dǒng	懂（董箽）	0.023
quē	缺（阙炔）	0.022
niang	娘	0.022
shao	少	0.022
luàn	乱（釠）	0.021
bian	边	0.021
cài	菜（采蔡縩）	0.021
xuè	血（谑）	0.021
nan	难	0.021
biāo	标（彪镖飙镳膘飚骠杓彯贆飑麃飆鑣鏢臕脿藨）	0.021
xuǎn	选（癣選）	0.020
tai	太	0.020
du	度	0.020
si	思	0.020
pá	爬（扒琶耙杷筢）	0.020
bai	白	0.020
chūn	春（椿蝽輴）	0.020
nòng	弄	0.019
duǎn	短	0.019
tòu	透	0.019
tái	抬（台苔跆薹邰炱骀鲐臺菭颱）	0.018
yìn	印（荫胤窨茚隂酳）	0.018
mǒu	某	0.018
dāo	刀（叨氘忉裯釖魛）	0.017
lè	乐（勒叻泐仂鳓簕詘）	0.017
yé	爷（揶）	0.017

Pinyin	Character(s)	Percentage
ren	人	0.017
niáng	娘	0.017
ròu	肉	0.017
tǎo	讨（討）	0.017
zhuī	追（椎锥骓隹錐）	0.016
niú	牛	0.016
na	哪（呐）	0.016
shùn	顺（瞬舜順蕣）	0.016
mō	摸	0.016
tuǐ	腿（脮蹆）	0.016
duān	端	0.016
làng	浪（莨蒗）	0.016
tuì	退（褪蜕煺蛻）	0.015
nǔ	努（弩砮）	0.015
shǎn	闪（陕掺睒閃陝）	0.015
shǎn	闪（陕）	0.015
xia	下	0.015
huān	欢（獾驩貛讙）	0.015
chōu	抽（紬瘳篘）	0.015
jie	姐（价）	0.014
duī	堆（鎚鴭）	0.014
fǒu	否（缶缻）	0.014
hā	哈（铪）	0.014
tào	套	0.014
pāi	拍	0.013
pèng	碰	0.013
rǎn	染（冉苒）	0.013
gǒu	狗（苟枸笱岣耈耇）	0.013
qióng	穷（琼穹邛蛩跫銎筇茕竆蛬藑瞏赹）	0.013
fěn	粉（瞓黺）	0.013
zuān	钻（躜躦）	0.012
tōu	偷（鍮）	0.012
di	弟	0.012
xǐng	醒（省擤）	0.012
hùn	混（诨溷）	0.012
liǎ	俩	0.012
cā	擦	0.012
bí	鼻（荸）	0.011

Pinyin	Character(s)	Percentage
niǎo	鸟（袅茑嬲蔦）	0.011
rěn	忍（棯荏秹）	0.011
nào	闹（淖臑閙）	0.011
kuān	宽（髋）	0.011
sōng	松（嵩淞凇崧忪菘）	0.011
cū	粗（麄麤）	0.011
mà	骂（蚂杩唛祃罵駡禡）	0.011
lóu	楼（喽髅娄偻蝼蒌耧謱蔞髏鞻）	0.011
suō	缩（嗦莎梭唆娑蓑挲嗍羧睃縮桫鮻髿）	0.011
hǔ	虎（唬浒琥虓）	0.010
pán	盘（磐蹒蟠盤膰蹣踫）	0.010
quān	圈（悛腃）	0.010
pén	盆（湓）	0.010
nai	奶	0.010
nù	怒	0.010
féi	肥（腓淝）	0.010
nuǎn	暖	0.010
sàn	散	0.010

4.2 High-Entropy Pinyin

Table 3 presents high-entropy pinyin with a frequency of at least 0.1%, highlighting the most ambiguous sound-character relationships in common Chinese usage. This threshold includes less common but still relevant pinyin, capturing a wider range of complex phonetic relationships while avoiding extremely rare cases. The four highest entropy pinyin are “xī”, “jì”, “yù”, and “fù”, with “xī” topping the list at an entropy of 3.276. This entropy value is equivalent to having approximately 10 equally likely characters (2^3.276 ≈ 10). In practice, it corresponds to 18 commonly used characters such as 西 (west), 息 (rest), 希 (hope), and others, along with an additional 45 rare characters. These high-entropy pinyin illustrate significant ambiguity in sound-to-character mapping, reflecting the rich complexity of the Chinese writing system.

Our entropy-based findings complement and extend previous research on orthographic transparency in Chinese. Studies by Siok and Fletcher (2001) and Ho et al. (2003) have shown that characters with higher transparency are easier for learners to acquire. However, these studies often rely on binary categorizations, which may limit their applicability to diverse educational contexts. In contrast, our entropy measurements provide a finer granularity for predicting potential learning difficulties by assessing predictability on a continuous scale. High-entropy pinyin identified in our study align with what would traditionally be considered “opaque” in orthographic terms, but our method enables ranking these challenging sound-character relationships, potentially informing more targeted instructional strategies (Lin et al., 2019; Tseng et al., 2023) over a 6-year period, in the relationship between character reading ability and orthographic awareness in Chinese from the first year of kindergarten to the third year of primary school in two separate samples: the kindergarten sample of 96 children was assessed three times in the first, second, and third years of kindergarten (K1, K2, K3. Identifying zero-entropy pinyin provides a data-driven approach to recognizing highly transparent orthographic units, potentially refining how characters are introduced in curricula.

An interesting observation arises with the neutral-tone pinyin “shi”, associated with characters like 是，事，实，and others. While many of these characters are not typically pronounced with a neutral tone in isolation, they frequently appear with neutral tones in common multi-character words (e.g. 还是 [háishi], 故事 [gùshi], 结实 [jiēshi], 试试 [shìshi], 认识 [rènshi], and 护士 [hùshi]). This underscores the importance of considering character pronunciation within the context of word formation and natural speech patterns, rather than in isolation.

The high-entropy pinyin highlighted in this analysis present challenges for learners, requiring a nuanced understanding of context and usage to correctly identify the intended character. However, they also offer opportunities for developing advanced language skills, particularly in character recognition and contextual comprehension. Focusing on these high-entropy pinyin can enhance learners' ability to disambiguate characters based on context, a crucial skill for achieving higher levels of Chinese language proficiency. Future research could empirically test the effectiveness of incorporating this entropy-based approach into language learning curricula.

Table 3

High-entropy Pinyin with a Frequency of at least 0.1%

Pinyin	Character(s)	Percentage	Entropy
xī	西息希吸析悉惜稀牺夕锡溪晰膝嘻熄犀蟋（昔栖熙兮嬉奚螅曦熹蹊羲汐烯蜥皙醯唏淅僖硒歙窸翕浠矽舾穸欷樨郗粞菥豨鼷訢鸂鍚徯糦鏭晞磎鑴俙翖）	0.140	3.276
jì	系计记济技际纪继既季剂寄寂（迹绩祭忌冀妓伎悸暨骥稷髻鲫偈蓟觊霁芰荠鲚計跽繼記洎際紀罽蹟繫穊跡哜鯚臮蟿茍訐穄繫鵋驥鯽鰿鬾）	0.360	3.225
yù	与育预域遇玉欲愈御狱誉郁豫裕吁寓（语谷喻浴谕毓蔚驭聿煜芋峪熨钰昱阈妪鹬饫鬻鹆蜮聿預穀燠遹鬱蓣馭矞禦鋊鴪轝礜譽驈鴥籲狗鈺閾籞隩鐍驈衘裛鋊礇蕷瑀）	0.102	3.128
fù	复父负富副付附妇腹赴缚（傅咐赋覆阜驸蝮馥讣鲋福赙負袝賦袝鰒複訃鍑復縛）	0.158	2.969
jìng	境竟静竞敬镜径净（劲靖痉胫迳靓鏡婧獍靜競脛逕競）	0.108	2.721
jī	机几基击激积迹鸡绩肌饥圾讥（奇玑稽姬畸缉叽矶羁唧跻嵇箕畿乩犄芨屐咭赍齑笄積墼谿雞飢剞踦齏齎績羇鄿虀觭羈鐖璣禨緝覊磯機賫）	0.352	2.624
jù	据具句剧巨距聚拒惧俱锯（瞿炬踞遽飓钜苣倨讵醵窭虡屨鐻椇秬鉅鋸簴粔）	0.173	2.610
lì	力利立历例丽厉励粒隶砾沥荔（莉吏栗笠雳俐痢戾蛎詈俪栎砺莅郦傈枥跞唳粝疬呖溧苈猁疠轹篥坜麗麜隸蒞脷苙磿綟讈藶蠣赲）	0.451	2.592
yì	意义议易益异艺亦亿译役翼忆抑疫毅谊屹（衣逸溢裔懿绎奕邑诣驿翌臆佚轶熠弋弈翊呓蜴薏刈羿缢翳镒峄悒肄挹癔仡義議怿佾瘞埸劓鎰袘镱殪睪齸詣藙蓺鮨鷁絏藝譯蛡鶃褹裛耴輗饐豷鶍鷊芅謚鐿讛賢驛螠）	0.566	2.483
jiàn	间见建件舰剑渐健键箭践鉴荐贱溅（监槛谏僭涧饯腱見毽鑑鍵踐艦薦踺楗瞷諓礀鋻趝繝珔）	0.419	2.402
yuán	原员元源园圆援缘猿（袁垣辕媛沅爰鼋圜芫螈塬橼緣鈨贠鶢黿縁褑薗笎）	0.298	2.384
yán	言研严延沿炎岩颜盐檐（癌阎蜒筵妍闫研顏鹽阽埏綖巖簷閻莚訮）	0.167	2.370
zhī	之只知指支织枝芝脂肢汁蜘（祗胝卮栀織隻鴲胑衼禔鼅鳷禵）	0.436	2.364
shù	数术述树束竖朮（恕墅庶漱戍術澍腧沭豎裋蒁鶐）	0.264	2.346
jí	及即集级急吉疾辑籍脊（吃极藉嫉棘汲亟笈瘠岌楫芨蒺嵴佶殛戢級鹡蕺蹐鶺膌踖鈒蝍鍓簎鞊趌）	0.288	2.313
zhì	制至治质置智秩掷稚帜（识致志滞挚峙窒炙痔痣蛭郅觯雉栉桎質鸷帙贽陟騭彘轾踬製忮誌铚袠胵芖紩鑕豒秷緻鷙跱銡袟锧稙覟贄騺隲鴙）	0.303	2.305

Pinyin	Character(s)	Percentage	Entropy
bì	必避毕币闭壁臂蔽碧毙痹痺（比泌辟弊陛庇婢敝璧弼裨愎贲蓖跸毖哔薜襞畀铋袐篦睥髀濞閉蓽躄荜筚狴禆鷩邲閟甓滗苾廦鷝诐箅鵖綼袐蜌鼊熶詖髲篳畢肸鐾皕鞸鉍駜鷩縪驆鮅）	0.116	2.301
shi	是事实式识士视势食拾匙（鳾）	0.160	2.254
zhí	直指值职执植殖（侄蛰踯摭跖蹠縶埴職銕禃膱）	0.145	2.111
fēng	风封丰峰疯锋蜂（枫烽沣酆風葑砜豐鄷犎碸鋒豊）	0.134	2.077
jiān	间坚监尖肩兼艰歼奸煎（渐浅笺缄鞯間菅犍缣篯湔鹣鞬戋蒹搛鰜鶼閒钘菺鰹監麗鰹鰜豎蕑鐧釿湕瀐箋）	0.219	2.047
shí	时实十识石食拾蚀（什炻鲥莳識祏埘迚鉐蝕鼫遈鉽）	0.918	2.009
wèi	为位未卫味谓慰胃喂畏（猬魏尉蔚渭鳚衛謂霨蝟菋讆硙鏏軎餵磑蔚躗錗鮇）	0.399	1.961
shī	师失施诗尸湿狮（虱蓍絁邿詩鍦葹鰤鰤鲺蝨）	0.133	1.958
xiàn	现见线限县献陷宪羡羡馅（腺霰苋岘線粯軐県晛縣）	0.400	1.949
dài	大代带待戴袋逮（贷黛怠殆岱迨玳甙骀绐埭轪艜襶簣貸䲦）	0.244	1.938
bèi	被备背贝倍辈狈（惫悖钡蓓焙孛碚鞴鐾褙貝誖邶骳輩鞞鋇）	0.187	1.936
yǔ	与语予雨宇羽屿（禹與龉俣庾圄窳伛語圉瘐貐頨蝺萸栩蘌）	0.120	1.936
jīn	金今禁津斤筋巾襟（矜钅衿觔矜）	0.142	1.920
qīng	清轻青倾氢蜻（卿輕鲭鶄圊鯖）	0.163	1.915
kè	克客刻课（恪嗑缂氪溘锞課骒艐衎碦礊騍）	0.104	1.850
shū	书输殊叔舒疏枢梳蔬（淑倏抒纾菽殳姝摅輸毹紓鵨綀）	0.134	1.808
zhèng	正政証证郑症挣（帧诤證）	0.332	1.798
wéi	为维围唯违惟桅（韦帷圩闱潍嵬帏維潿鮠鍏鄬違艬沩艜）	0.242	1.798
qí	其奇齐骑旗棋崎（只歧祈鳍琪琦祁祺耆脐岐淇芪麒畦蛴圻颀祇蕲綦亓荠骐萁臍碁蜞鐖鯕鯕跂齊軝騎鬐褀鬐魌蚑竒纃艩粸蠐）	0.142	1.777
yóu	由游油尤犹邮铀（疣鱿猷莜莸繇蝣蚰尢輶鲉莤蝤遊鈾蚘郵魷蓧）	0.181	1.750
sī	斯司思私丝撕嘶（厮咝蛳锶鸶缌澌鷉絲鼶凘褫鶑緦禗）	0.124	1.743
biàn	变便遍辩辨辫（汴卞弁變苄缏辯忭覍緶艑）	0.207	1.729
jiào	教觉叫较轿窖酵（校醮較峤覺徼轎噍藠毀訆覚）	0.275	1.728
shì	是事世市式士示似视势试适室释氏饰侍誓逝拭（轼嗜仕恃噬柿谥舐視弑螫筮適莳釋試铈謚贳眂篩銕襫鈰飾）	2.208	1.704
gǎn	感敢赶杆秆（橄擀鳡簳澉鱤赶稈）	0.132	1.695
huà	话化划画桦（华話繣）	0.325	1.680
jiē	结接阶街皆揭（节偕秸嗟疖節階喈袓腱稭蓵）	0.224	1.650
zhù	住助筑驻祝柱铸蛀（着注著贮伫杼箸炷苎翥纻貯跓麈疰築苧竚紵鑄駐羜註祩）	0.166	1.648
qī	期七妻欺漆凄凄沏（溪戚栖缉蹊嘁萋桤柒碕郪諆鶈）	0.132	1.647
ba	吧巴爸罢拔叭笆	0.289	1.640
xìng	性兴幸姓（行杏悻荇興臖莕）	0.141	1.606
xiāng	相香乡箱厢镶（湘襄骧芗缃鄉葙纕蘘緗鑲）	0.120	1.580
yè	业夜叶页液咽（拽曳谒腋掖邺晔烨靥葉頁鐷緤鍱）	0.206	1.576

Pinyin	Character(s)	Percentage	Entropy
wàng	望往忘妄旺（眖迋）	0.118	1.568
zhǐ	只指止纸址（旨趾徵咫酯芷祉枳阯黹紙轵觝茝絺藢）	0.315	1.561
shè	社设射涉舍摄（慑赦麝歙厍設蔎滠騇）	0.245	1.544
zuò	作做坐座（凿唑酢祚柞胙怍阼蓙）	0.504	1.540
dì	的地第帝弟递缔（蒂谛棣娣睇碲遞禘菂締釱腣逓諦祶）	0.470	1.540
shǐ	使始史驶屎（矢豕駛鉂）	0.232	1.535
píng	平评凭瓶屏苹（萍坪鲆枰評聠蛢蓱缾）	0.136	1.534
nán	难南男喃（楠難）	0.109	1.529
gōng	公工功供攻宫弓躬（蚣恭龚觥肱魟碽龔）	0.455	1.528
qì	气器弃汽泣砌（妻契迄亟憩讫碛槭葺碶汔磜磩蟿）	0.224	1.525
xì	系细戏隙（夕阋翕饩細禊舄蠹绤钑闟餼）	0.108	1.514
xi	西系息	0.116	1.508
wǔ	五武午舞侮捂（伍鹉妩庑忤迕怃仵牾膴碔）	0.163	1.481
lián	联连怜廉帘镰（莲涟濂臁鲢裢蠊奁連鐮蓮聯簾鎌鬑聫）	0.110	1.474
jié	结节杰洁截捷竭睫（桔劫诘颉桀偈拮孑碣婕羯結讦疖絜蛣鲒蓵蜐詰）	0.130	1.470
chéng	成程城承盛诚乘呈惩（澄丞橙裎枨铖塍酲埕郕脭誠絾碀）	0.377	1.466
jué	决觉绝角掘嚼（脚爵厥诀崛倔抉攫獗蕨蹶谲橛珏噱矍镢桷劂孓絕爝钁觖觼蕝蹷玃芵訣蟨穱絶躩）	0.172	1.464
wù	物务恶误悟雾（勿晤兀坞戊鋈骛鹜婺寤焐芴杌誤軏霚痦霧阢鶩）	0.212	1.454
cháng	长常场偿尝肠（裳嫦長苌腸徜鲿）	0.192	1.443
xiào	笑效校肖啸（孝詨）	0.131	1.442
bàn	办半伴扮瓣拌（绊辧跘絆）	0.151	1.434
zhòng	中种重众（仲眾諥茽）	0.199	1.421
jīng	经精惊睛晶鲸茎腈（京荆兢菁經旌泾粳鶁莖鼱麖秔荊鯨）	0.246	1.408
yuè	月越乐阅跃悦（钥岳粤樾刖钺閲龠瀹躍篗趯軏粵礿躒）	0.173	1.407
shēn	身深参申伸绅呻（娠莘砷诜糁鲹蓡詵蔘鯵籸紳葠鯓）	0.198	1.376
ma	吗妈麻嘛蟆（么）	0.146	1.371
shén	什神甚（鰰）	0.496	1.369
liàng	量亮辆谅晾（踉靓諒）	0.115	1.367
jiè	界介借届戒诫（解藉芥疥蚧骱褯誡衸蛶）	0.107	1.344
dàn	但弹担淡旦蛋诞氮（石惮澹啖萏瘅霮禫蜑饏駳誕贉髧）	0.185	1.338
guān	关观官棺（冠倌莞關鰥觀蒄窤）	0.184	1.327
hé	和合何河核荷盒（颌禾劾涸阂阖龢纥菏曷貉盍翮饸龁盇鞨粭鶡麧覈鹖礉頜盉）	0.776	1.326
xí	习席袭媳（锡褶檄習隰觋郋襲鳛騽霫蓆薂鎴）	0.104	1.313
xíng	行形型刑（邢硎饧荥陉鉶）	0.263	1.287
jìn	进近尽禁劲浸（晋烬靳噤荩觐缙妗進盡赆賮藎祲）	0.319	1.282
zhàn	战站占蘸（颤绽湛栈菚）	0.232	1.272
jǐ	己给几挤（脊戟麂虮鱾掎）	0.321	1.269

Pinyin	Character(s)	Percentage	Entropy
dǎo	导倒岛蹈捣（祷禱隝）	0.119	1.237
jiǔ	九久酒（灸韭玖紤韮）	0.131	1.234
tā	他她它踏塌（遢趿铊祂溻）	1.409	1.228
lù	路陆露录鹿碌（禄赂戮麓漉璐辘箓潞鹭渌逯蓼辂陸簏錄蕗菉盝録駱祿簬粶騼膟騄睩稑醁賂簶鷺）	0.156	1.209
shēng	生声升牲（胜甥笙聲陞苼鼪鉎）	0.446	1.204
xīn	心新辛欣薪芯锌（馨鑫忻歆莘昕鋅）	0.328	1.203
liú	流留榴硫（刘瘤浏琉遛馏镏鎏旒骝鎦飗騮飀鹠藰驑）	0.104	1.188
yīn	因音阴姻（殷荫茵湮氤喑陰洇堙铟骃禋秵絪闉駰裀霠銦蒑陻）	0.163	1.129
kē	科颗棵磕瞌蝌（柯苛珂轲窠嗑颏髁稞疴蚵簻顆钶窼趷薖軻頦）	0.113	1.108
cái	才财材裁（財）	0.170	1.095
què	却确雀（鹊阙榷阕確悫鵲闕）	0.127	1.092
kuài	会快块筷（脍侩狯哙浍鲙郐鄶駃）	0.151	1.091
tí	提题蹄啼（題缇绨鹈醍荑鳀虒遆稊趧綈鶗騠緹磃蕛）	0.194	1.069
diàn	电店殿垫奠淀佃惦（甸玷癜钿靛簟電阽坫蜔鈿磹）	0.133	1.061
rèn	任认韧（刃妊纫饪恁仞衽認轫葚訒韌袵讱靭餁纴絍靱飪紝軔）	0.149	1.036
yǐ	以已椅乙蚁倚（矣迤旖苡钇锜螘顗齮蟻饮艤舣陁肊釔礒）	0.505	1.025
jing	经静晴	0.124	1.020
dù	度杜渡肚镀（妒蠹芏詑鍍秺）	0.113	1.016
xiān	先鲜仙纤掀（酰暹锨跹籼氙祆莶鲜纖縿韱鱻）	0.145	1.013
xiàng	相像象项巷橡（向項蟓鱌）	0.221	1.011
míng	明名鸣（铭冥茗瞑溟螟暝鳴眳銘眀鄍）	0.163	1.011
ne	呢吶	0.276	1.000
dào	到道倒盗稻悼（焘纛帱盗稲衜翢軇）	0.799	0.999
jiā	家加佳夹茄（挟嘉迦枷袈痂浃珈跏笳葭镓筴麚猳貑服）	0.401	0.997
gēn	根跟	0.166	0.991
shǒu	手首守（艏）	0.199	0.938
zhēng	正争挣睁怔蒸（症征铮筝狰徵峥钲箏篜）	0.121	0.930
jūn	军均君菌（钧筠麇皲軍麕鲪袀碅麏覠莙）	0.135	0.929
bìng	并病（摒竝）	0.129	0.927
zhēn	真针珍侦（贞斟臻帧桢祯甄箴砧榛針胗椹溱蓁鉁禛鎭鱵貞禎眞）	0.154	0.906
měi	美每镁（浼羙）	0.103	0.906
bù	不部步布怖埠（簿钚瓿蔀篰踄）	1.676	0.891
gé	格革隔骼（蛤阁葛阖嗝镉搿騔膈鬲閣詥轕裓塥鎘）	0.156	0.883
yú	于鱼渔愚愉舆娱竽（与予余於逾瑜虞禺俞榆隅渝欤谀盂馀觎腴臾揄畲萸嵛窬顒蝓餘颙雩狳舁妤魚隃邘諛輿貗舒鱮苧釪羭）	0.250	0.864
lǐ	里理礼鲤（李哩蠡俚澧锂醴娌逦裡鳢悝鱧裏粴禮鋰）	0.627	0.851
bǐ	比笔彼鄙（匕俾吡妣筆秕舭聛貏柴粃）	0.129	0.821
quán	全权泉拳（痊蜷诠荃颧铨醛鬈筌鳈鰁詮譔絟辁硂葲銓顴）	0.204	0.804
tài	太态汰（泰钛肽酞鈦粏舦）	0.106	0.796

Pinyin	Character(s)	Percentage	Entropy
wán	完玩顽丸（烷芄纨蚖貦）	0.107	0.795
qián	前钱潜钳（乾虔黔荨掮钤箝錢葴鍼軡鰬）	0.269	0.787
yǎn	眼演掩衍（奄俨偃魇兖鼹琰厣郾罨剡龂顩蝘菴鄢黡齗鶠黶）	0.158	0.780
duì	对队（兑怼碓隊憝鐓錞镦祋鋭陮）	0.388	0.773
zhōng	中终钟忠（衷盅锺忪螽舯終鐘蝩蹱鈡）	0.341	0.765
zhǎng	长掌涨（鏱仉鞝）	0.134	0.749
sì	四似饲（食伺寺肆嗣祀巳俟泗笥姒驷汜耜兕覗飼蕼竢禩）	0.152	0.733
de	的地得（底）	5.273	0.727
xiǎng	想响享（饷飨響鲞餉饟鯗）	0.330	0.698
bǎi	百摆柏（佰捭襬）	0.111	0.686
hái	还孩（骸還）	0.427	0.665
ge	个格哥歌搁	0.129	0.655
biān	边编鞭蝙（砭笾鳊煸邊編邉箯）	0.139	0.641
wén	文闻纹蚊（雯阌玟聞闅閺鼤紋閿螡魰）	0.126	0.599
dōng	东冬（咚氡鸫岽鶇蝀笗鶫菄）	0.112	0.598
ér	而儿（鸸粫鲕輀胹）	0.243	0.592
méi	没梅眉煤枚霉黴酶（媒玫湄嵋楣莓镅鹛郿猸䁋蘪鋂禖鎇）	0.365	0.588
yòu	又右幼诱佑（釉祐柚囿宥蚴鼬侑誘褎褏）	0.328	0.570
gè	个各（铬虼硌箇）	0.808	0.568
zì	自字（渍恣眦眥胔胾）	0.320	0.548
fēn	分纷氛吩（芬酚玢雰紛翂鈖）	0.157	0.526
zuì	最罪醉（蕞）	0.139	0.525
cì	次刺伺（赐莿賜）	0.114	0.516
suǒ	所索锁琐（唢鎖鏁）	0.165	0.515
tiáo	条调（迢笤龆苕蓨髫鲦蜩鞗鰷蓧儵）	0.111	0.514

4.4 Implementing entropy-based learning with Pleco flashcards

To translate the theoretical insights of our entropy-based analysis into practical learning tools, we developed a set of flashcards compatible with the Pleco Chinese dictionary app—a widely used platform among Chinese language learners on iOS and Android devices. By integrating our findings into interactive flashcards, we provide learners and educators with tangible resources to directly apply the concept of entropy in language instruction.

4.4.1 Zero-entropy flashcards

The first set of flashcards focuses on zero-entropy pinyin sounds, as identified in Table 2. These pronunciations uniquely map to single characters, reducing ambiguity and facilitating easier character recognition. The flashcards are organized according to the new HSK levels 1 through 6, extended levels 7–9, and an additional level 10 that includes all characters not listed in the standard HSK levels. This organization allows learners to select decks that match their proficiency, providing a structured pathway from basic to advanced characters.

To enhance the learning experience, we designed three types of flashcard tests within the Pleco app:

1. **Ear Training Exercise:** This test aids students in practicing the transcription of spoken characters into pinyin. The app plays the audio pronunciation of a character, and learners input the corresponding pinyin. This exercise sharpens listening skills and reinforces the association between sounds and their written representations.
2. **Stroke Order Writing Practice:** In this exercise, learners hear the pronunciation of a character and are prompted to write it using the correct stroke order. The Pleco app offers immediate feedback, providing hints after a few incorrect attempts and allowing students to practice writing the character multiple times if needed. This reinforces orthographic knowledge and enhances writing proficiency.
3. **Free Writing Without Stroke Order:** This test allows learners to write any character they believe matches the given pronunciation, without restrictions on stroke order. After submission, the app verifies the correctness of the character. This exercise encourages active recall and tests the learner's ability to produce characters based solely on auditory input.

4.4.2 High-entropy flashcards

The second set of flashcards targets high-entropy pinyin sounds, as detailed in Table 3. These pronunciations correspond to multiple common characters, presenting a higher level of ambiguity. The flashcards are organized by the number of associated characters:

- **High-Entropy 2: Pinyin associated with exactly two common characters (e.g., "bìng" for** 并 [and] and 病 [illness]).
- **High-Entropy 3: Pinyin associated with exactly three common characters (e.g., "bǎi" for** 百 [hundred], 摆 [place], and 柏 [cypress]).
- …
- **High-Entropy 9+: Pinyin associated with nine or more common characters (e.g. "shì" for** 是 [be], 事 [matter], 世 [world], 市 [market], and 16 others).

These flashcards are intended for self-review, enabling learners to focus on differentiating between characters that share the same pronunciation. By studying these high-entropy sounds, learners engage with the inherent ambiguity in Chinese phonology, improving their ability to disambiguate meanings based on context—a skill crucial for advanced language proficiency.

4.4.3 Integration into teaching practices

The implementation of entropy-based flashcards in the Pleco app exemplifies how theoretical concepts can be seamlessly integrated into practical teaching tools without requiring learners to have explicit knowledge of entropy or predictability. Educators can incorporate these flashcards into their curriculum to implicitly guide students through phonetic complexities, tailoring instruction to address specific learning difficulties associated with sound-character mappings.

For instance, starting with zero-entropy flashcards allows beginners to build confidence through unambiguous sound-character associations. As learners progress, introducing high-entropy flashcards challenges them to utilize contextual cues and deepen their understanding of character usage. This graduated approach aligns with pedagogical strategies that emphasize scaffolded learning and supports findings by Liu and Wiener (2020) on leveraging homophones to facilitate lexical development.

4.4.4 Accessibility and demonstration

To ensure ease of access, a flashcard text files has been prepared (https://tinyurl.com/3zec568r) for direct import into the Pleco app, which features built-in dictionary, audio, and handwriting functionalities conducive to interactive learning. A demonstration video posted on YouTube (https://youtu.be/LLTm2bo_pDA) accompanies this paper to help guide users through the process of importing and utilizing the flashcards.

5 Conclusion

This study introduces an entropy-based approach to analyzing sound-character mappings in Chinese and demonstrates its practical application through the development of specialized flashcards for the Pleco app. By quantifying the uncertainty associated with mapping sounds to characters, we provide a systematic way to identify and categorize characters based on their phonetic uniqueness. This data-driven method offers a unique perspective on the relationship between phonology and orthography in Chinese, potentially informing both pedagogical approaches and linguistic research.

Our analysis highlights the complexities of the Chinese writing system while offering a structured framework for understanding character-sound relationships. The educational implications of this entropy-based approach are significant. By providing a quantitative measure of character-sound relationships and integrating these insights into practical learning tools, this study offers educators and learners new resources for curriculum development and self-study. Lessons and materials can be structured to progressively introduce characters based on their entropy values, potentially leading to more efficient and effective Chinese language instruction. The use of Pleco flashcards enables an interactive and accessible means of applying these concepts, enhancing learner engagement and reinforcing key skills in listening, writing, and pronunciation.

Several limitations of this study should be acknowledged:

1. **Empirical Validation:** While our approach shows promise, the efficiency and effectiveness of this method for enhancing Chinese learning have not been directly tested. Future research should include empirical studies to evaluate how the entropy-based mapping, implemented through tools like the Pleco flashcards, impacts learning outcomes.
2. **Data Sources:** The analysis relies on character frequencies from the Chinese Character Wiki, which may not perfectly reflect spoken language frequencies or regional variations. Future research could utilize alternative spoken corpus data to refine entropy calculations and improve the generalizability of the findings.
3. **Focus on Individual Characters:** The study primarily focuses on individual characters rather than multi-character words, which are prevalent in modern Chinese. Contextual cues in multi-character words can significantly modify the underlying probabilities of possible characters. Extending the analysis to include words could provide a more comprehensive understanding of language use.

Future research directions could include:

1. **Developing and Testing Learning Strategies:** Creating and evaluating specific instructional strategies based on the entropy of character-sound relationships, assessing their effectiveness relative traditional teaching methods.

2. **Extending to Multi-Character Words:** Analyzing entropy at the word level and exploring how character-level entropy relates to word-level comprehension, potentially leading to the development of additional learning modules or flashcard sets.
3. **Integration with Other Language Learning Aspects:** Investigating how an entropy-based approach might be integrated with reading comprehension or other aspects of language learning, and how educational technology platforms like Pleco can facilitate this integration.

In conclusion, this study offers an innovative entropy-based approach for analyzing sound-character mappings in Chinese, providing a quantitative framework to assess the ambiguity or predictability of these relationships. By leveraging entropy calculations, we quantify the uncertainty associated with mapping pinyin to characters, offering new insights into the complexities of the Chinese writing system and tangible resources for learners and educators. Unlike traditional binary classifications of orthographic transparency, our approach captures a continuum of predictability, which better reflecting the nuanced challenges learners face. This entropy-based perspective allows educators to design more effective curricula by focusing on characters with lower entropy to build foundational knowledge, while progressively incorporating more ambiguous characters as students' proficiency develops. Future research should further explore how this method can be applied and evaluated in practical educational settings and its effectiveness in enhancing Chinese language acquisition.

Acknowledgments

We extend our appreciation to the anonymous reviewers and the journal editor for providing thoughtful and detailed comments that substantially improved this manuscript.

References

Cao, F., & Perfetti, C. A. (2016). Neural signatures of the reading-writing connection: Greater involvement of writing in Chinese reading than English reading. *PLOS ONE, 11*(12), e0168414. https://doi.org/10.1371/journal.pone.0168414

Chai, X., & Ma, M. (2022). Exploring relationships between L2 Chinese character writing and reading acquisition from embodied cognitive perspectives: Evidence from HSK big data. *Frontiers in Psychology, 12.* https://doi.org/10.3389/fpsyg.2021.779190

Chao, Y. R. (1968). *Language and symbolic systems.* Cambridge University Press.

Ho, C. S.-H., Yao, P. W.-Y., & Au, A. (2003). Development of orthographic knowledge and its relationship with reading and spelling among Chinese kindergarten and primary school children. In C. McBride-Chang & H.-C. Chen (Eds.), *Reading Development in Chinese Children* (pp. 55–71). Greenwood.

Hsuan, C.-H., Tsai, H. J., & Stainthorp, R. (2018). The role of phonological and orthographic awareness in learning to read among Grade 1 and 2 students in Taiwan. *Applied Psycholinguistics, 39*(1), 117–143. https://doi.org/10.1017/S0142716417000194

Lee, J.-R., & Huang, C.-R. (2022). Phonological awareness, orthography, and learning to read Chinese. In C.-R. Huang, I.-H. Chen, Y.-H. Lin, & Y.-Y. Hsu (Eds.), *The Cambridge handbook of Chinese linguistics* (pp. 3–22). Cambridge University Press. https://doi.org/10.1017/9781108329019.002

Lightbown, P. M., & Spada, N. (2013). *How languages are learned 4th edition—Oxford handbooks for language teachers.* Oxford University Press.

Lin, D., Mo, J., Liu, Y., & Li, H. (2019). Developmental changes in the relationship between character reading ability and orthographic awareness in Chinese. *Frontiers in Psychology, 10,* 2397. https://doi.org/10.3389/fpsyg.2019.02397

Liu, J., & Wiener, S. (2020). *Homophones facilitate lexical development in a second language.* System, 91, 102249. https://doi.org/10.1016/j.system.2020.102249

Liu, J., & Wiener, S. (2021). CFL learners' Mandarin syllable-tone word production: Effects of task and prior phonological and lexical learning. *Chinese as a Second Language Research, 10*(1), 31–52. https://doi.org/10.1515/caslar-2021-0002

Liu, J., & Xiao, C. (2021). Tone category learning should serve tone word learning: An experiment of integrating pronunciation teaching in the L2 Chinese curriculum. In C. Yang (Ed.), *The acquisition of Chinese as a second language pronunciation: Segments and prosody* (pp. 141–162). Springer. https://doi.org/10.1007/978-981-15-3809-4_6

Olsen. (n.d.). *Chinese character Wiki—懂中文 Dong Chinese—Learn Mandarin Chinese.* Chinese Character Wiki. Retrieved June 12, 2024, from https://www.dong-chinese.com/wiki

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27,* 379–423, 623–656.

Siok, W. T., & Fletcher, P. (2001). The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition. *Developmental Psychology, 37*(6), 886–899.

Tseng, C.-C., Hu, J.-F., Chang, L.-Y., & Chen, H.-C. (2023). Learning to read Chinese: The roles of phonological awareness, paired–associate learning, and phonetic radical awareness. *Reading and Writing, 36*(7), 1769–1795. https://doi.org/10.1007/s11145-022-10352-9

Wiener, S., Lee, C.-Y., & Tao, L. (2019). Statistical regularities affect the perception of second language speech: Evidence from adult classroom learners of Mandarin Chinese. *Language Learning, 69*(3), 527–558. https://doi.org/10.1111/lang.12342

笪骏. (2004). 中文文库计算：现代汉语单字字频 *[Dataset]*. https://lingua.mtsu.edu/chinese-computing/

Arthur Berg (董愉), Professor at Penn State University, has research interests that include the application of data science across various disciplines including Chinese language learning.

International Journal of Chinese Language Teaching (2025)
Vol. 6 (2) 63-89 https://doi.org/10.46451/ijclt.20250205

基于熵的汉字音字映射学习

董愉
宾夕法尼亚州立大学，美国

摘要

本研究介绍了一种利用独特音字关系的创新汉语学习方法。通过在音字映射中应用熵的概念，我们提供了一种基于语音独特性来识别和分类汉字的系统方法。我们的方法专注于听力和写作技能，着重通过区分对应于唯一汉字的声音和与多个汉字相关的声音来提高听写能力。这种方法不仅有助于准确书写汉字，还能强化正确的发音，从而全面提高汉语水平。通过熵计算提供发音和汉字之间关系的定量指标，并将这些发现整合到实际的学习工具中，本研究为更深入地理解汉语学习做出贡献，并为教育者和学习者提供实际应用，可能提高教学效果和学习成果。

关键词

声字映射，语音意识，音调识别，熵，教育技术

董愉（Arthur Berg），宾夕法尼亚州立大学教授，研究兴趣包括数据科学在汉语学习等各个学科领域的应用。