Behavior Research Methods (2024) 57:34
https://doi.org/10.3758/s13428-024-02513-1

ORIGINAL MANUSCRIPT

A large-scale database of Mandarin Chinese word associations from the Small World of Words Project

Bing Li[1,2] · Ziyi Ding[1] · Simon De Deyne[3] · Qing Cai[1,4,5]

Accepted: 14 October 2024 / Published online: 30 December 2024

Abstract

Word associations are among the most direct ways to measure word meaning in human minds, capturing various relationships, even those formed by non-linguistic experiences. Although large-scale word associations exist for Dutch, English, and Spanish, there is a lack of data for Mandarin Chinese, the most widely spoken language from a distinct language family. Here we present the Small World of Words–Zhongwen (Chinese) (SWOW-ZH), a word association dataset of Mandarin Chinese derived from a three-response word association task. This dataset covers responses for over 10,000 cue words from more than 40,000 participants. We constructed a semantic network based on this dataset and evaluated concurrent validity of association-based measures by predicting human processing latencies and comparing them with text-based measures and word embeddings. Our results show that word centrality significantly predicts lexical decision and word naming speed. Furthermore, SWOW-ZH notably outperforms text-based embeddings and transformer-based large language models in predicting human-rated word relationships across varying sample sizes. We also highlight the unique characteristics of Chinese word associations, particularly focusing on word formation. Combined, our findings underscore the critical importance of large-scale human experimental data and its unique contribution to understanding the complexity and richness of language.

Keywords Word association · Chinese · Semantic network · Mental lexicon

Introduction

Humans store the meaning of tens of thousands of words, and this mental lexicon supports our daily activities, such as reading, writing, and communication. This vast repository of words is not just a static collection of terms – it exists in our brains as a dynamic and complex network (Hills & Kenett, 2022). Within this semantic network, each word is not isolated, but interconnected with others through their meanings. Thus, when we use language, these words and their interrelations assist us in efficiently understanding and expressing thoughts. While some semantic networks, such as WordNet (Fellbaum, 2010) or ConceptNet (Speer et al., 2017), have been built primarily based on linguistic annotation, empirical semantic networks can be more directly derived from a free word association task. In this task, participants list the first words that come to mind when presented with a cue word. This method is easy and remarkably efficient, covering a wide range of semantic relations without the need to spell out full propositions (Szalay & Deese, 1978). The information encoded in word association is unique because it encodes not only lexical relations but also includes experiential knowledge (De Deyne et al., 2021; Ufimtseva, 2014).

Scaling psychological measures of meaning to include the most common words requires an immense data collection effort, which has become feasible through online data collection tools (Chen et al., 2023). The Small World of Words

Bing Li, Ziyi Ding and Simon De Deyne contributed equally to this work.

✉ Qing Cai
qcai@psy.ecnu.edu.cn

1 Key Laboratory of Brain Functional Genomics (MOE & STCSM), Affiliated Mental Health Center (ECNU), Institute of Brain and Education Innovation, School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

2 Univ. Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, 59000 Lille, France

3 School of Psychological Sciences, University of Melbourne, Melbourne, VIC, Australia

4 Shanghai Changning Mental Health Center, Shanghai, China

5 NYU-ECNU Institute of Brain and Cognitive Science, New York University, Shanghai, China

(SWOW) project has used online crowd-sourcing to recruit volunteers since 2010, which has led to the most extensive word association norms to date in Dutch (De Deyne et al., 2013), English (De Deyne et al., 2019), and Rioplatense Spanish (Cabana et al., 2024). These norms have been adopted widely to study a range of phenomena related to early word learning (Cox & Haebig, 2023), creativity (Johnson & Hass, 2022), the prediction of primary semantic dimensions like concreteness or emotional valence (Meersmans et al., 2022; Van Rensbergen et al., 2016; Vankrunkelsven et al., 2015), the interaction between semantics and perceptual modality (De Deyne et al., 2021), and studies of risk (Meersmans et al., 2020; Wong et al., 2022; Wulff & Mata, 2022), etc.

Although large-scale word association norms have been developed for several alphabetic languages, such studies are currently lacking for Mandarin Chinese, one of the most spoken logographic languages in the world.[1]The current study aims to build the first large-scale dataset of Mandarin Chinese associates. As a logographic language, Chinese exhibits unique characteristics compared to alphabetic languages, particularly in its semantic structure and the formation of words. In Chinese, most words are compound, created by combining single characters that also function as individual words. These characters are interconnected through a non-inflectional morphological system, displaying intricate relational, semantic, and syntactic associations (Packard, 2000). Most characters represent morphemes, but they are also words themselves. For instance, "推翻 (topple)" features a cause-effect dynamic between "推 (push)" and "翻 (turn over)." Similarly, "电脑 (computer)" reflects a semantic link between "电 (electricity)" and "脑 (brain)." While Chinese words can be made up of any number of morphemes in theory, they typically contain fewer than four characters. It is therefore difficult to separate the effects of morphemes/characters from those of words. As for cultural differences, previous studies suggested that Chinese speakers are also more holistic and focus more on context and relationships between things, while alphabetic speakers are more analytical in general and pay more attention to details, rules, and logic (Ji et al., 2000; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005). Reflected in language, many Chinese words have a nondiscriminatory role, which in turn often requires contextual information for grammatical processing and analysis (Vigliocco et al., 2011). Unlike Indo-European languages, which use inflection or suffixes to signal changes in part of speech, Chinese words frequently can function as different parts of speech in the same form. For example, "花" can mean "spend/cost [verb]," "flower [noun]," or "colorful [adjective]." These words are nondiscriminatory depending on their context. Therefore, it is unclear whether findings from Indo-European languages so far would generalize to Mandarin Chinese. Given the aforementioned factors, a systematic comparison between Chinese and alphabetic languages necessitates comprehensive and fundamental linguistic analyses, taking into account characteristics unique to Chinese word formation and production.

To evaluate the norms, we referred to prior reports in Dutch, English and Spanish, examining how semantic networks, built from word associations, predict word processing speeds as measured by reaction times in lexical decision and naming tasks (Barber et al., 2013; Cañas, 1990; Pexman et al., 2008; Rodd et al., 2002; Yap et al., 2011), as well as human ratings of relatedness and similarity of word pairs. These validations involve using text-based external measures, including word frequency and context diversity (Cai & Brysbaert, 2010; Liu et al., 2010), as well as distributional word relation estimates derived from word embeddings (Li et al., 2018), which are relatively broadly used and are assumed to capture meaning as well. Previous studies have suggested that incorporating indirect associations into the estimation of word pair relations can most effectively approximate human perceptions of relatedness and similarity. To be specific, broad relatedness aligned more closely with associative relationships (Cabana et al., 2024; De Deyne et al., 2019), while strict similarity corresponded to shared semantic features (Vulić et al., 2020). For instance, while "lion" and "zoo" are highly related, they are not similar. On the other hand, "lion" and "tiger," both members of the Felidae family and similar in appearance, exhibit a high degree of similarity. In addition to differentiating between similarity and relatedness, our analysis also considers the contextual reliance characteristic of Chinese. Besides the complexity of Chinese word relationships mentioned earlier, the inherent ambiguity of words makes it unclear whether estimates derived from Chinese word associations might require a larger sample size compared to other languages. While association networks were reported to outperform word2vec in some languages (Cabana et al., 2024; De Deyne et al., 2013, 2019), the consistency of this advantage in Chinese and with smaller-sized association norms remains unclear. To address the sample size issue, noting that previous results utilized larger participant groups, we also aimed to determine how sample size might influence alignment with human ratings.

Furthermore, with the rapid advancement and increased application of computer models in language processing, particularly large language models (LLMs) such as GPT and other transformer-based models like BERT, the current study also addresses a critical question: Can human big data-based association norms like SWOW-ZH provide a more nuanced understanding of human behavior in language tasks compared to these advanced language models, which primarily leverage vast amounts of text data? This inquiry is pivotal, as it evaluates whether the inherently different nature of data sources—human-generated associations versus machine-learned patterns

[1] Over 1.3 billion people by 2018, see https://www.ethnologue.com/statistics

from text—leads to distinct or complementary insights in linguistic research. Considering that generative transformers demonstrated near-human capabilities in next-word prediction (Goldstein et al., 2022; Schrimpf et al., 2021), we sought to explore whether generative models like GPT could effectively capture word relationships and establish association norms in the same way as SWOW. Across these comparisons, the aim is not primarily to establish what measure performs best, but to determine whether good performance can be achieved with a relatively parsimonious model in terms of data.

The remainder of the article is structured as follows. The next section provides information about the method and procedure, and how this project in Mandarin Chinese differs with related projects in other languages.

Methods

Participants

Participants have been continually engaged in the project through social media, emails, and lectures since 2016. The project primarily utilized the SWOW platform, where most participants contributed without compensation. They took part in a word association task consisting of 18 cue words. Up to April 2023, 40,903 individuals had participated in the word association task, which comprised native speakers of Chinese, without distinguishing between Mandarin and Cantonese speakers. At a later stage, to accelerate the construction of the database, our data collection was switched to NAODAO[2] (Chen et al., 2023) for a small section of the data. Participants responded to 80 to 90 cues, and each participant was compensated for 15 RMB (2.2 USD).

In data processing, we excluded data from three types of participants: Cantonese speakers, due to the focus on Mandarin language in the present study; occasional participants under 16 years old; and participants that do not fulfill the quality control criteria (see the Preprocessing section). The resulting norm contains the data and demographic information including age, gender, education level, native dialect, and optional geographic location.

Procedures

The main task is a three-response continued free association task, a variant of the single-response free association task that addresses the issue of response dominance, where most participants give the same response, and captures weaker associations between words, as well (De Deyne et al., 2013; Nelson et al., 2000). In a three-response free association task, participants were required to give three words related to a cue word. The instructions were the same as used in the English SWOW (De Deyne et al., 2019) for consistency. The details are described in Appendix 1.

Stimuli

The cues were continuously expanded using a snowballing sampling procedure. In the seeding set of cues, 983 commonly used words, including the most high-frequency words from Subtlex_CH (Cai & Brysbaert, 2010) and highly frequent words from previous semantic studies, were included. After presenting the complete set of cues to at least 60 participants, a new set of cues was obtained by selecting frequent responses that were not cues. In each iteration, a batch of around 1000 new cues was selected from the association responses. As of July 25, 2022, the study had collected a total of ten cue sets with 10,192 cues, and 8241 of those were also presented on NAODAO by April 15, 2023.

A separate group of participants was included to validate the impact of sample size on relation estimates (for more details, see Appendix 1). 164 cue words were derived from the relatedness judgment task of De Deyne et al. (2020), with the number of trials/participants expanding to 80 for concrete words and 120 for abstract words.

Preprocessing

Preprocessing was performed to ensure the quality of the responses, and resulted in two datasets that will be publicly released: raw and balanced. The raw dataset accommodates diverse research needs by preserving the original responses. The balanced dataset, prepared through cleaning and standardizing with exactly 55 trials for each cue word, facilitates comparison across cue words by controlling the cue presentation rate. This specific trial count, which is lower than similar projects in Dutch and English, will be further validated in subsequent stages of the study to confirm its effectiveness for the comparative analyses highlighted.

The preprocessing (see Fig. 1) involved merging, cleaning, and balancing the data similar to what has been done before in Dutch, English and Spanish. In the following sections, if not specified, we used the term "types" to refer to words as uniquely encoded Chinese words and "tokens" to refer to the occurrences of a type. Figure 1 shows an overview of the preprocessing, with detailed steps described in Appendix 2. We will outline the steps specifically adjusted to accommodate the unique characteristics of Chinese, which differ from the prior SWOW-EN preprocessing protocols.

The raw data were cleaned for responses and participants one by one. With regard to the data cleaning for responses,

[2] Due to the unreliable access to the main SWOW website experienced in mainland China over the past few years, additional data were collected through the NAODAO platform beginning November 29, 2022.

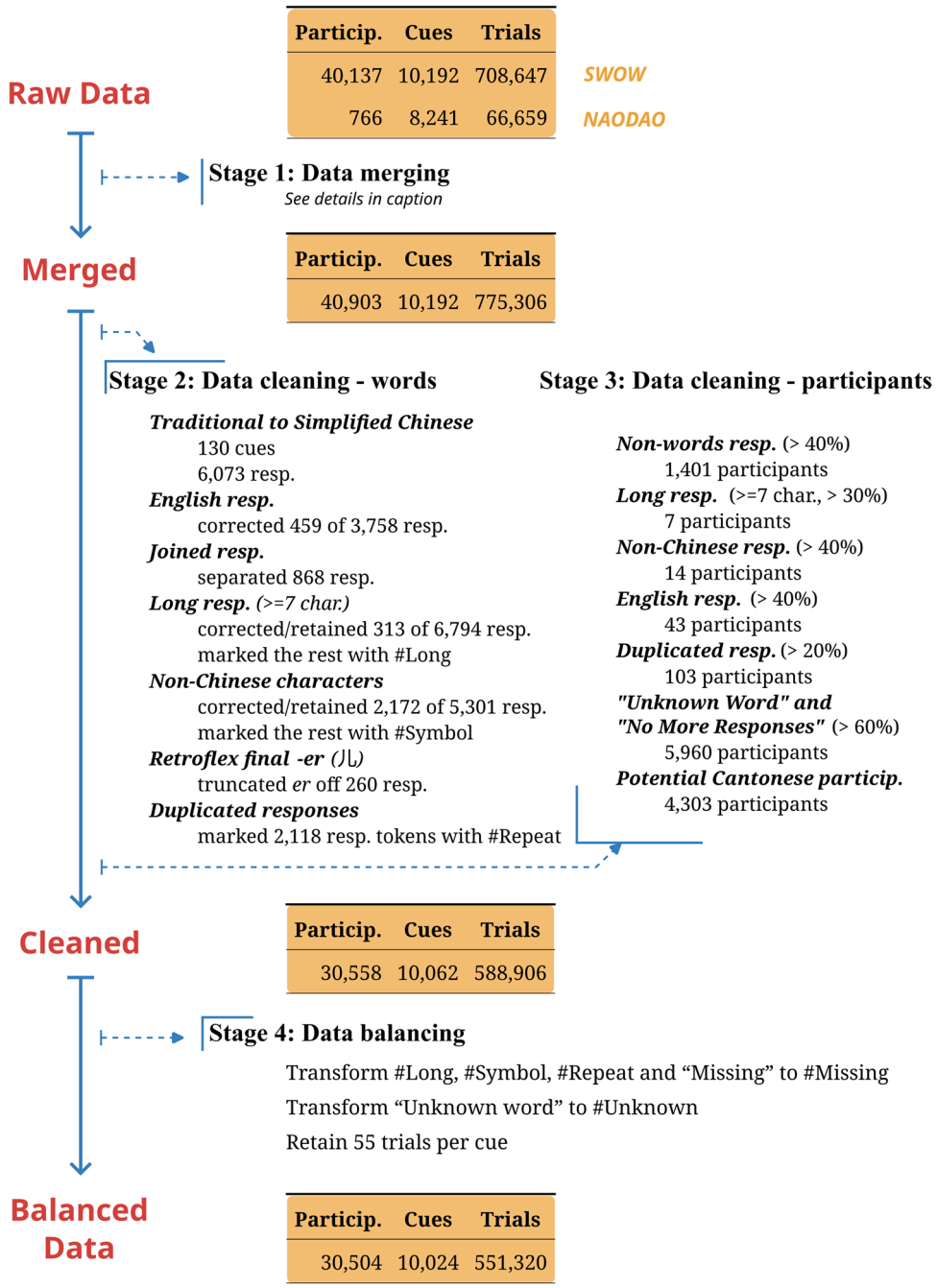

Fig. 1 Data preprocessing procedure. The procedure consists of three stages: merging, cleaning, and balancing. During pre-screening, 85 taboo responses were masked with hexadecimal codes, and 19 taboo cues were eliminated. The count of responses refers to the number of unique response types, excluding duplicates. Abbreviations: Particip. for participant; resp. for response; char. for character

since Chinese words are not separated by white space and have indefinite lengths, they have a blurry boundary with multi-word phrases. To eliminate lengthy phrases and sentences, responses exceeding six characters—unless they were proper nouns—and those containing non-Chinese characters or considered meaningless were tagged as “Missing” (tags were coded in the main data file with a preceding # sign). Three other types of responses underwent the following transformations: Traditional Chinese responses were converted to their Simplified Chinese form, English responses were retained and corrected for capitalization, and retroflex *er* finals (erhua or erization) were deleted. In the data cleaning for participants, Cantonese participants and those who left more than 60% blank were excluded, resulting

in the exclusion of over 10,000 participants. Preprocessing codes are available in MATLAB and R.

Constructing the SWOW-ZH network and measures

We constructed the SWOW-ZH network and corresponding measures based on the aforementioned data. The nodes in the network represent cues, while the associative strengths between these cues form the weighted edges. Notably, we only used the strongly connected component in this directed and weighted cue-to-cue graph, to ensure that each node, or cue, was part of a bidirectional relationship with both inbound and outbound connections. Responses that were not part of the cue sets, and cue words that were not associated with other cue words, were not included. The associative strength was calculated as the conditional probability between 0 and 1 of a response given a cue (Abbott et al., 2015; De Deyne et al., 2019). Two versions of the network, R1 and R123, were validated and compared in the following sections. Connections in the R1 network were based on the first response only and included 9814 nodes (i.e., 97.91% of the original cues), while the R123 network was based on all three responses, including 9899 nodes (98.75% of the original cues).

External validation

To evaluate the effectiveness of SWOW-ZH, we investigated the extent to which metrics based on SWOW-ZH associations could explain human performance in word processing. Following SWOW-EN (English), SWOW-RP (Rioplatense Spanish) and SWOW-NL (Dutch), we employed external norms of lexical decision response times (RTs), naming latencies, and word pair ratings.

In the context of lexical processing, we paid special attention to the unique aspects of Chinese, particularly the role of individual characters in composing words during word naming.

In terms of word relationships, we focused on examining the differences in explanatory power relative to sample size between measures derived from SWOW and those based on large-scale text models in interpreting human rating data. Finally, we explored the potential of large language models (GPT-4o) to establish SWOW-LLM association norms by employing the same free association paradigm used in SWOW.

Lexical decision task

Centrality metrics, such as degree and betweenness, are expected to impact the prominence of a word in the network and its accessibility in language understanding. Words more central in the network are reached quicker and easier during cognitive processes, such as recall or recognition. Hence, effective centrality measures should be good predictors of lexical processing (Barber et al., 2013; Pexman et al., 2008; Yap et al., 2011). In this study, we aligned SWOW-ZH metrics with those from other sources to evaluate their effectiveness in predicting lexical decision latencies from the MELD-SCH megastudy (Tsang et al., 2018), aiming to discern which measures are most effective and their unique contributions. Two alternative sources, the Unigram subset of Chinese Web 5-g Version 1 (Liu et al., 2010) and SUBTLEX-CH (Cai & Brysbaert, 2010), were employed for comparisons.

To avoid multicollinearity and assess the unique contributions of each measure, we initially employed backward stepwise regression to identify significant predictors of lexical decision RTs, thereby establishing a base model. Subsequently, we applied a leave-one-out strategy, systematically removing each independent variable to form a restricted model, which was then compared with the base model to determine the impact of the removed variable by likelihood ratio test (LRT, *lmtest* R package, see Hothorn et al., 2022). Since the base model was identical and ΔR^2 was adjusted by the number of predictors, we can weigh the effects of each measure by inspecting the ΔR^2 values.

Word naming task

Naming latencies and lexical decision times are often used alternatively as measures of word recognition (Katz et al., 2012). While the lexical decision task has sometimes been questioned because it incorporates a decision-making component, word naming latencies could be considered more direct measures, although some studies suggest that monosyllabic words benefit from the predictable associations between graphemes and phonemes, which could affect the directness of word naming latencies as a measure (Balota et al., 2004). We then investigated whether word centrality, derived from SWOW-ZH, can effectively predict word naming latencies.

It is also important to note that, unlike alphabetic languages, Chinese is characterized by its character-based and analytic structure. The process of retrieving multi-character words is more transparent and parsimonious, yet more influenced by within-word structures and complexity (Tse et al., 2017). Although "single-character word" naming provides a simple and informative metric and has been widely used in previous studies, it is challenging to dissociate it from "character" naming. Given that SWOW-ZH also encompasses a large number of single-character words, it is feasible to differentiate characters and single-character words in our analysis. Therefore, we used two datasets of word naming as the dependent variables: 2423 single-character words from Liu et al. (2007), and 1922 two-character words from Zhang et al. (2023). We conducted separate analyses on the character- and word-related measures in both datasets, focusing on

the influence of the characters that constitute words, as well as their properties as single-character words, on word naming. In each analysis, both the character- and word-related measures were included as candidate independent variables and then filtered using backward stepwise regression, followed by LRT to evaluate each selected predictor.

Relatedness and similarity rating tasks

Human ratings of relatedness and similarity are widely employed in studies of semantic cognition and as benchmarks for assessing distributional models (Bhatia, 2017; Mandera et al., 2017). Likewise, previous versions of SWOW in Dutch, English, and Spanish found that word associations provide a reasonable representation of the mental lexicon, which captures the relationships between words, even those not directly linked. However, it is not clear to what extent these effects will replicate in a language with different distributional properties, such as Chinese. To evaluate SWOW-ZH association norms, we used two behavioral datasets: relatedness judgment from De Deyne et al. (2020) and semantic similarity ratings (SimLex) from Vulić et al. (2020). The main difference between the two datasets is that SimLex tasks focus on similarity, with participants instructed to judge strict similarity (e.g. *coffee–tea*), whereas relatedness judgement tasks assess a broader range of associations (e.g. *coffee–cake*) and include distinct sets of abstract and concrete word pairs.

In addition to the SWOW-ZH association, our analysis incorporated word embeddings like word2vec and advanced language models such as BERT (Turc et al., 2019), GPT-2 (Radford et al., 2019). Turbo, the Chinese word2vec, originates from a publicly available model trained on Baidu Encyclopedia using skipgram and negative sampling (Bojanowski et al., 2017; Li et al., 2018). We also accessed the pre-trained GPT-2-Medium (trained on Chinese Wikipedia)[3] and WoBERT (training corpus had been processed with word segmentation).[4]

We also explored the impact of sample size on behavioral alignment. To investigate this, we increased the number of SWOW-ZH participants for the words used in the relatedness judgment task (De Deyne et al., 2020), which included 82 concrete words and 82 abstract words. To assess how sample size affects estimating relatedness between the 164 words, we gradually increased the number of participants in intervals of five, from 55 to 120 for abstract words and from 55 to 80 for concrete words. With each increment in participant numbers, the new associations contributed by the extra participants led to a significant reweighting of edges surrounding these 164 words.

Finally, we experimentally compared human associations with large language model generated associations using GPT-4o-2024–08-06 (OpenAI). It received the same instructions as humans (see Appendix 1 for details) and provided three responses for each of the 164 cues. To enhance the diversity of associations, the temperature was set to 1.5 (which could range from 0 to 2), and the frequency penalty was set at 1.0. Raising the temperature beyond this level resulted in more nonsensical responses. To ensure comparability between human and language model associations, the same number of "participants" and identical preprocessing procedures were applied. To estimate word relations based on GPT-4o, we removed the edges from the 164 cues in the SWOW-ZH network and replaced them with responses from GPT-4o, using the same algorithm to assess relations among the cues.

Results

The preprocessing trimmed the association counts per cue to ensure uniformity. Each cue received 55 trials of three responses from 55 distinct participants, creating a balanced dataset that gave equal weight to all cues. This was crucial for preventing any potential bias that could arise if certain cues were overrepresented. The cutoff of 55 associations per cue was determined after observing that the influx of new association types plateaued at this number, which will be detailed in the section "Types, tokens, and word distribution." Our analysis, starting with an examination of the demographic and lexical characteristics of the data, is based on this balanced dataset. We then proceed to external validation, maintaining the use of the balanced dataset throughout. It is important to note that only the strongly connected component of the network was utilized for external validation.

Demographic information A total of 40,903 participants completed the free association task, with 40,137 participants coming from the SWOW platform and 766 participants coming from NAODAO. A total of 30,504 participants were retained after preprocessing, 80.73% of which were female, 17.72% male, and 1.55% non-binary. The average age was 23.04 years ($SD = 6.15$, min = 16, max = 99). Most participants had a college or university bachelor's degree or above (91.67%), and 8.09% of the participants were high school graduates. The majority of the participants (88.37%) reported Mandarin as their native language, followed by northern dialects (6.39%) and southeastern dialects (2.71%). It is worth noting that there was a minor oversight in the coding process, where Northwestern and Northern dialects were inadvertently assigned the same code. This made them appear indistinguishable (see Preprocessing—Stage 3 in Appendix 2, and the Native Language column in the main

[3] https://huggingface.co/mymusise/gpt2-medium-chinese

[4] https://github.com/ZhuiyiTechnology/WoBERT?tab=readme-ov-file

data file for the exact issue). Most participants were located in China (88.67%), the United States (4.65%), and Japan (1.83%). Among those from China, metropolitan areas had a significant representation of participants, e.g. Shanghai (16.81%), Beijing (9.65%), and Guangzhou (4.91%).

Missing and unknown responses Responses were marked as missing when participants skipped a trial by selecting "Unknown Word" or "No More Responses" if they were unable to provide a second or third associate. On average, 0.84% of all cues were marked as "Unknown Word," with a standard deviation of 2.00%. For R1, where the "No More Responses" option was unavailable, the missing response rate was $0.53\% \pm 1.32\%$. In this case, the missing R1 responses were composed of responses categorized as Long, Symbol, or Repeat during preprocessing. Additionally, following the methodology used in the English, Dutch, and Rioplatense Spanish norms, we combined the unknown cues with the missing R1 responses to calculate an "unknown" cue rate of 1.37%. For R2 and R3, the missing rates, which included both "No More Responses" and the Missing tags, were $14.62\% \pm 7.60\%$ and $25.60\% \pm 10.19\%$, respectively. Compared to the Dutch, English, and Rioplatense Spanish data, the rate of unknown cues was relatively low in Chinese (SWOW-ZH: 1.37%, SWOW-EN: 1.42%; SWOW-DU: 2.5%; SWOW-RP: 3.3%), but the rates of missing responses in R2 and R3 were relatively high in Chinese and Rioplatense compared to Dutch and English.

Types, tokens, and word distribution A total of 122,396 types (unique words) and 1,415,409 tokens (word occurrences) were retained after preprocessing. Among them, 65,148 types appeared only once. These hapax legomena responses covered 53.23% of types but only 4.60% of tokens. For the R1 responses, we retained 65,846 types and 543,775 tokens, of which 34,099 types (51.79%) appeared only once, covering 6.27% of the tokens. The most common responses in SWOW-ZH and SWOW-EN are listed in Table 1. The frequent response types refer to how many cues elicited those responses, with the denominator being the total number of cues; The frequent response tokens refer to how many times those responses were given across all cues, with the denominator being the total times of cues multiplied by the number of times each prompt was responded to (which is 55 in the balanced dataset). Notably, they exhibit a cross-language consistency, as indicated by the bold format, with most related to instinct and basic needs.

The vocabulary growth curves represent the main distributional characteristic of the balanced dataset (Fig. 2a). The number of types was introduced as a function of the number of response tokens. To examine whether SWOW-ZH is scale-free, we tested how well finite Zipf-Mandelbrot models could capture the word distribution in SWOW-ZH. The Zipf-Mandelbrot model, often used in linguistics, suggests that a few words (types) are used very frequently (high token count), while many others are rare (Baayen, 2001). The finite Zipf-Mandelbrot model extends the original concept by considering a finite number of words, which is more practical for real-world datasets

Table 1 The ten most frequent responses in SWOW-ZH and SWOW-EN norms. These metrics were calculated for the first responses only (R1) and for all three responses (R123). The bold items represent words that are common to both English and Chinese

	Types R1		Tokens R1		Types R123		Tokens R123	
	SWOW-ZH	SWOW-EN	SWOW-ZH	SWOW-EN	SWOW-ZH	SWOW-EN	SWOW-ZH	SWOW-EN
1	人 **(man)**	**money**	钱 **(money)**	**money**	人 **(man)**	**money**	人 (man)	**money**
2	我 **(me)**	food	人 (man)	food	钱 **(money)**	**water**	钱 **(money)**	**water**
3	钱 **(money)**	**water**	水 **(water)**	**water**	我 (me)	food	水 **(water)**	food
4	水 **(water)**	car	我 (me)	car	工作 **(work)**	**red**	工作 **(work)**	car
5	实验 (experiment)	love	工作 (work)	music	水 **(water)**	love	可爱 (cute)	music
6	工作 **(work)**	**work**	衣服 (cloth)	old	实验 (experiment)	**work**	红色 **(red)**	green
7	游戏 (game)	bad	数学 (math)	sex	游戏 (game)	bad	老师 (teacher)	**red**
8	红色 (red)	**good**	游戏 (game)	love	白色 (white)	fun	游戏 (game)	love
9	白色 (white)	**man**	考试 (exam)	dog	红色 **(red)**	good	时间 (time)	**work**
10	好 **(good)**	**me**	红色 (red)	bird	死亡 (death)	**man**	朋友 (friend)	old

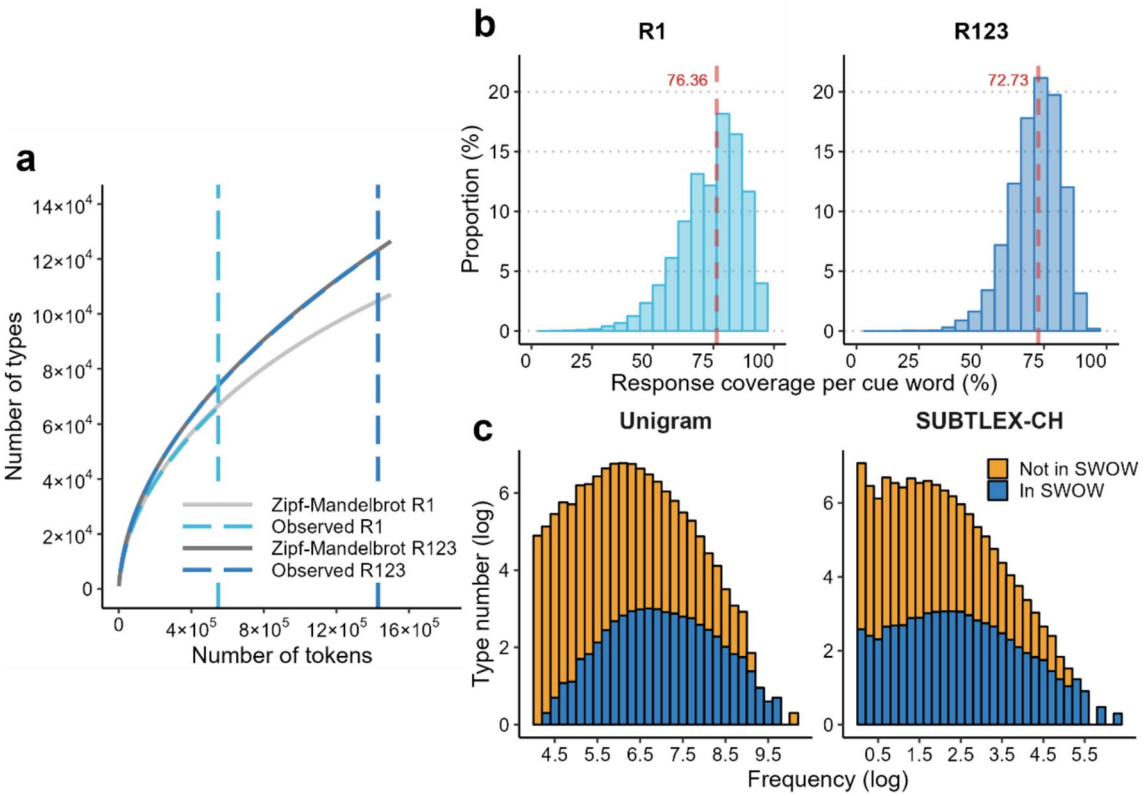

Fig. 2 The distributional characteristics of the balanced dataset. **a** The growth of types as a function of the number of tokens among responses. Both R1 and R123 curves were fitted to a finite Zipf-Mandelbrot model. **b** The response coverage of cues. The red dashed lines mark the median response coverage among cues. The y-axis represents the proportion of tokens summed from each column, with the denominator being the total tokens in SWOW-ZH networks. **c** The coverage of SWOW-ZH cues on Unigram and SUBTLEX-CH. The *x*-axis displays word frequency bands labeled with their midpoints based on corresponding lexicons, while the *y*-axis shows the number of words in each band

where the number of words is limited. The curves based on R1 and R123 were separately predicted by finite Zipf-Mandelbrot models using the zipfR package (Baroni & Evert, 2014).

As shown in Fig. 2a, for the balanced dataset with 55 participants, the growth rate decreases with the addition of tokens. It is anticipated that adding new participants beyond 55 per cue will result in fewer new types in the balanced dataset. Nevertheless, extracting the strongly connected component might have led to a loss of responses. Given that the SWOW-ZH networks were built based on the strongly connected component and thus discarded non-cue responses, we need to ensure that the SWOW-ZH networks cover the majority of tokens after the transformation. To this end, we calculated the response coverage for each cue—the ratio of its associated tokens in the network to those in the balanced dataset. The distribution of response coverage is shown in Fig. 2b. The median response coverage among cues was 76.36% in the R1 network and 72.73% in the R123 network, indicating that half of the cues retain most of their associated tokens after network transformation.

We further compared the SWOW-ZH distribution to the frequency spectra from text-based corpora, such as SUBTLEX and Unigram norms, to gauge its representativeness of natural language usage. The results showed a similar distribution between the frequency spectrum from word associations and those from the corpora (Fig. 2c). This similarity indicates that SWOW-ZH can comprehensively represent the lexical distribution of Chinese natural language.

External validation

Lexical decision task

To evaluate the effectiveness of centrality measures derived from SWOW-ZH, we compared the node unweighted in-degree[5] from SWOW-ZH with an alternative centrality indicator from SUBTLEX-CH, which offers word frequency and

[5] We also explored several other centrality measures from SWOW-ZH, such as in-strength (weighted in-degree), PageRank, betweenness and in-closeness, but found that in-degree performed on par or better.

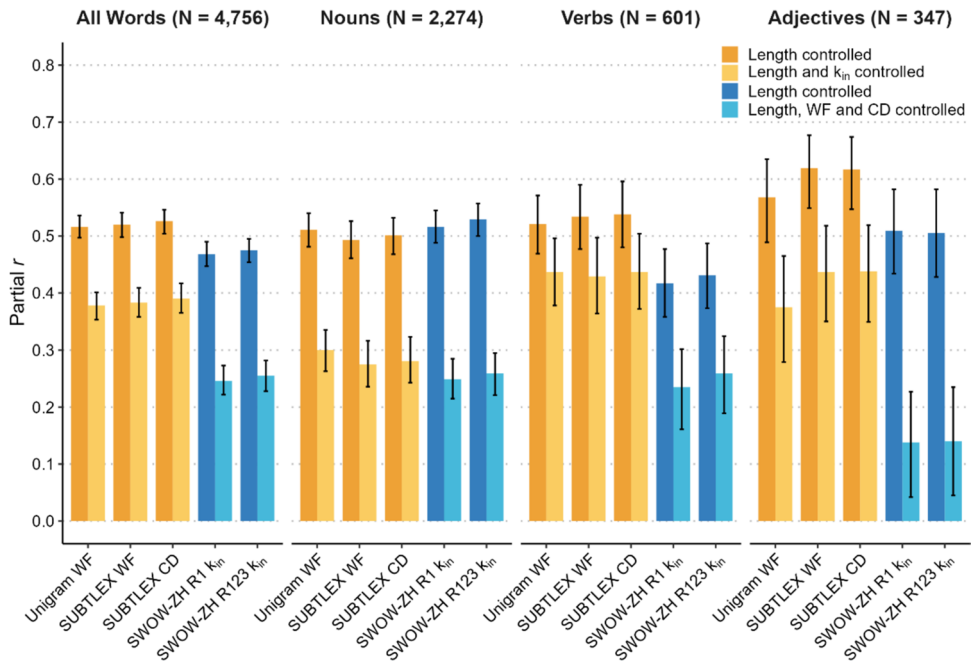

Fig. 3 SWOW-ZH word centrality showed a significant correlation with lexical decision RTs. The absolute values of the partial correlation coefficient *r* are displayed on the *y*-axis. The indicators on the *x*-axis, from left to right, include three text-based indicators (in orange: Unigram word frequency, SUBTLEX-CH word frequency, and word contextual diversity) and two association-based indicators (in blue: unweighted in-degree of SWOW-ZH R1 and R123). The dark bars represent results controlling for word length (ranging from one to four characters). The light orange bars show results for word length alongside two association-based indicators, while the light blue bars display results for word length and three text-based indicators. The error bars depict 95% confidence intervals based on 1000 bootstraps

contextual diversity based on movie subtitles (Cai & Brysbaert, 2010). Furthermore, we incorporated word frequency from the Unigram subset of Chinese Web 5-g Version 1, an N-gram database developed from web content (Liu et al., 2010). In upcoming analyses, we accounted for the variability contributed by these three measures of accessibility in lexical decision latencies, thereby isolating the unique impact of word association. These lexical decision latencies were sourced from Tsang et al. (2018), with word length spanning one to four characters. All metrics underwent log transformation, consistent with the methodology applied to previous SWOW norms (e.g. SWOW-EN, De Deyne et al., 2019).

We analyzed the data of Tsang et al. (2018) for the words that were covered in all the three norms (SWOW-ZH, SUBTLEX-CH, Unigram). This resulted in a total of 4756 words. Each word was annotated with its dominant part-of-speech (PoS), which was identified as having a higher than 95% PoS frequency in SUBTLEX-CH database. Figure 3 and Supplementary Table 1 show the partial correlations between lexical decision latency and word accessibility measurements for all words combined and split across PoS (nouns, verbs, and adjectives). The results showed that overall, text-based word frequencies and contextual diversity had slightly higher correlations with lexical decision task (LDT) performance than word associations. However, word association, despite the relatively small number of observations, demonstrated the strongest performance specifically within the category of nouns, outperforming all other measures.

Previous studies suggested that contextual diversity and word frequency tap into distinct aspects of lexical processing (Adelman & Brown, 2008; Brysbaert & New, 2009). Similarly, association-based metrics have shown unique contributions in studies involving different languages. Here we further examined the unique contributions of word associations within the context of Chinese lexical processing. To tackle multicollinearity among the mentioned metrics, a backward stepwise regression was initially performed to discern significant predictors of LDT performance from five indicators. This process led to the selection of Unigram WF, contextual diversity (SUBTLEX CD), and the unweighted in-degree from SWOW-ZH (SWOW-ZH R123 k_{in}) as primary predictors. These formed our base model (as shown on the left side of Table 2). Further model comparisons were carried out to ascertain the individual contributions of each predictor within this base model. We used a likelihood ratio test (LRT) to evaluate the impact of excluding each predictor (details on the right side of Table 3), applying the *lmtest* package (Hothorn et al., 2022).

Based on the model comparisons, we found both text-based and association-based indicators uniquely contribute

to predicting LDT reaction times. Among them, excluding the in-degree (k_{in}) resulted in the largest change in explained variance, at 4.7%, compared to 2.9% for word frequency (WF) and 3.6% for contextual diversity (CD). The independent contribution of association-based network centrality is consistent with what was found in SWOW norms in other languages (Cabana et al., 2024; De Deyne et al., 2013, 2019). The results indicate that in-degree in Chinese plays a comparably distinctive role compared to text-based word frequency and contextual diversity. It also indicates that these results are robust, despite the smaller size of the Chinese dataset, and the larger lexicon size and ambiguity of its language system.

Word naming task

Similar to the approach used with the LDT, we evaluated the contribution of various frequency and contextual diversity measures in word naming. This consisted of two independent analyses for single-character word naming datasets and two-character word naming datasets. Specially, considering the impact of word formation, we additionally included the character word frequency (CF) and character contextual diversity (CCD) soured from SUBTLEX-CH as potential predictors for single-character naming tasks, besides the five predictors evaluated in the LDT above, resulting in seven independent variables for backward stepwise regression (see the left part of Table 3).

For single-character words, text-based word frequency (unigram WF), contextual diversity both for characters and single-character words (SUBTLEX CCD and CD), and association-based unweighted in-degree of the three-response network (SWOW-ZH R123 k_{in}) emerged as significant, unique predictors of naming speed. Notably, besides word indicators selected to predict naming latency similarly to those in the LDT, the largest explained variance, at 3.4%, was attributed to a character indicator, SUBTLEX CCD (see the right part of Table 3). This suggests that the speed of single-character word naming was influenced by contextual diversities arising from its use both as a character and a single-character word.

For two-character words, word formation was taken into account in addition to the predictors of single-character-word naming. Specifically, we investigated whether and how the lexical processing of two-character words was influenced by the features of the characters that compose them. Therefore, in addition to the aforementioned five measures for the two-character words, the measures for the head (first) and tail (second) characters composing the words were also included, resulting in 15 variables: five each for two-character words, head-character words and tail-character words. Since not all characters are single-character words, we also included the corresponding SUBTLEX CF and CCD measures for both head and tail characters. Therefore, 19 independent variables were entered into the backward stepwise regression.

Four significant factors were identified for two-character naming speed in backward regression, including text- and association-based indicators, as well as character- and word-level indicators. The base model predicting the naming

Table 2 Results of base model predicting LDT and likelihood ratio test for model comparisons

Predictors	Base model					Likelihood ratio test		
	β	*SE*	*t*	*p*	VIF	χ^2	*p*	ΔR^2
Unigram WF	−.23	.001	−14.75	<.001	1.80	212.86	<.001	.029
SUBTLEX CD	−.26	.001	−16.55	<.001	1.80	266.35	<.001	.036
R123 k_{in}	−.25	.002	−18.84	<.001	1.35	342.48	<.001	.047

Base model: $F(3, 4752) = 921.10$, $p < .001$, $R^2 = .37$

Table 3 Results of the base model predicting the naming RT of single-character words (left) and likelihood ratio test results for model comparisons

Predictors	Base model					Likelihood ratio test		
	β	*SE*	*t*	*p*	VIF	χ^2	*p*	ΔR^2
SUBTLEX CCD	−0.35	.016	−6.17	<.001	3.59	37.40	<.001	.034
Unigram WF	−0.24	.011	−5.00	<.001	2.55	24.74	<.001	.022
SUBTLEX CD	0.13	.015	2.16	.032	4.39	4.66	.031	.004
R123 k_{in}	−0.13	.010	−3.93	<.001	1.32	15.40	<.001	.014

Base model: $F(4, 846) = 68.69$, $p < .001$, $R^2 = .25$

speed of two-character words is outlined in the left part of Table 4, while the unique contributions of each predictor are shown on the right side. Notably, the character frequency of the first character of words contributed the largest explained variance, at 3.8%.

Our findings primarily highlighted the unique contributions of each predictor. Both text-based and association-based approaches significantly contributed to naming tasks across different word lengths, consistent with results observed in Chinese LDT and in other languages (Cabana et al., 2024; De Deyne et al., 2013, 2019). Notably, the naming speed of words is most accelerated by their first character (character CD for single-character words and character frequency of the first characters for two-character words), rather than by the words as a whole. For both single-character words and two-character words, text-based contextual diversity (SUBTLEX CD) and association-based unweighted in-degree of the three-response network (SWOW-ZH R123 k_{in}) were found to be significant predictors of naming speed. This finding highlighted that both textual and associative contexts are important and distinct.

The impact of word formation is evident in the lexical preprocessing of both single- and two-character words. Specifically, the naming speed of single-character words is significantly influenced not only by their own contextual diversity (as a word) but also by that of the underlying character (as a character). Similarly, the naming speed of two-character words is markedly affected by both their contextual diversity and the frequency of the individual characters that compose them.

To further explore word formation in the SWOW-ZH network, we analyzed the proportions of edges based on word formation relationships. Among the edges derived from 1029 single-character words in the SWOW-ZH network, two types relate to word formation: those that connect to multi-character words composed of their predecessors, and those that link to single-character words which, combined with their predecessors, form meaningful two-character words. We found that the former represents 17.62% of all edges in the R1 network and 11.30% in the R123 network. For the latter, the figures are 31.15% in R1 and 21.51% in R123, respectively. Taken together, these findings highlight the significant role of word formation within SWOW-ZH, underscoring their contribution to the structural and semantic complexity of Chinese.

Relatedness and similarity rating

To evaluate the SWOW-ZH association norms, we derived relationship estimates from SWOW-ZH and three sources of open-source word embeddings (word2vec, GPT-2-Medium and WoBERT), and then compared their correlations with human ratings of relatedness (data from De Deyne et al., 2020) and similarity (data from SimLex, from Vulić et al., 2020). To align with previous norms, we utilized the same algorithms, including associative strength, positive pointwise mutual information (PPMI), random walk (RW, cf. De Deyne et al., 2019), and compressed random walk (RW Embedding, see Cabana et al., 2024).

The associative strength and its modified version, where biases between associating and being associated were adjusted using PPMI, performed less effectively than the RW-based algorithms (see Supplementary Table 2 for associative strength and PPMI results), which is in line with previous studies. Importantly, we focused on evaluating relationship estimates through the random walk approach. Unlike associative strength and PPMI, which only capture local connectivity, RW depicts word similarity through broader, global connectivity by incorporating indirect links with decayed strength into direct associations. We anticipated that this distinction would become more pronounced due to data sparsity, which is especially pronounced in SWOW-ZH. For a comprehensive comparison with word embeddings, we also included a compressed representation of the dense random walk graph (RW Embedding). In SWOW-RP, this measure slightly outperformed random walk for similarity ratings, where participants judged strict semantic similarity between words (Vulić et al., 2020).

Figure 4 shows the results for SWOW-ZH word relation estimates (See also Supplementary Table 2 for the full results). As can be seen in the figure, R123 consistently outperformed R1 estimates, highlighting the importance of the multiple associations derived from a continued procedure.

Table 4 Results of the base model predicting the naming latency of two-character words (left) and likelihood ratio test results of model comparisons (right)

Predictors	Base model					Likelihood ratio test		
	β	*SE*	*t*	*p*	VIF	χ^2	*p*	ΔR^2
SUBTLEX CF head	−0.22	.003	−4.27	< .001	1.31	18.04	< .001	.038
SUBTLEX CF tail	−0.11	.003	−2.22	.027	1.18	4.97	.026	.010
SUBTLEX CD	−0.14	.003	−2.24	.026	1.80	5.03	.025	.010
R123 k_{in}	−0.12	.005	−2.31	.021	1.26	5.39	.020	.011

Base model: $F(4, 401) = 20.00$, $p < .001$, $R^2 = .17$

Moreover, in line with previous studies, SWOW-ZH RW outperformed other estimates in predicting relatedness ratings (Fig. 4b), whereas SWOW-ZH RW embedding and WoBERT both exhibited top performance in predicting similarity ratings (Fig. 4a). When examining the results for similarity and relatedness together, it is evident that SWOW-ZH and word2vec performed optimally for the relatedness of concrete word pairs, followed by the relatedness of abstract word pairs, and performed the poorest on the similarity measures of SimLex. Conversely, GPT-2 and WoBERT excelled in strict similarity assessments but lagged in measures of broader relatedness.

For abstract word pairs, the correlations between the model-based estimates of random walk and the relatedness judgments increased with sample size, a trend that persisted up to 120 participants (see Fig. 5 and Supplementary Table 3). Conversely, a smaller sample size of about 50 participants was found sufficient for concrete concepts. Regardless of the sample size variation, the performance of random walk for abstract word pairs consistently remained lower than that of concrete word pairs. However, it was unlikely that sample size was the only explanation for this discrepancy: word2vec and GPT-4o, both trained on a much larger corpus, also exhibited relatively poor performance for abstract concepts, even though abstract words were presumably mostly learned from texts (but see De Deyne et al., 2021 for an alternative perspective).

For the performance of associations generated through GPT-4o, increasing the number of response failed to improve relation estimates for both abstract and concrete word pairs, resulting in significantly lower performance in predicting human relatedness ratings compared to human associations. Even with a sample of just 20 participants (yielding a maximum of 60 words), the results from SWOW-ZH R123 outperformed word2vec and GPT-4o, both of which were trained on billions of tokens. These findings suggest that semantic networks derived from word associations may offer a valuable alternative to other measures, with implications for a variety of tasks where accurate relation estimates between word pairs are crucial.

General discussion

This article introduces the first large-scale Mandarin Chinese word association database, SWOW-ZH, which encompasses over 10,000 cue words and involves over 40,000 participants. Similar to earlier norms in Dutch, English, and Rioplatense Spanish, SWOW-ZH used a tailored pipeline to adapt to Chinese linguistic features for data preprocessing and achieving balanced data. The database demonstrated strong explanatory abilities in behavioral tasks, confirming its effectiveness.

Distributional characteristics of SWOW-ZH

The vocabulary characteristics in SWOW-ZH align with patterns observed in natural language and other SWOW norms, fitting a finite Zipf-Mandelbrot model (Cabana et al., 2024; De Deyne et al., 2013, 2019). This model describes how the growth of types tends to plateau despite a continuous accumulation of tokens. The skewed distribution in SWOW-ZH

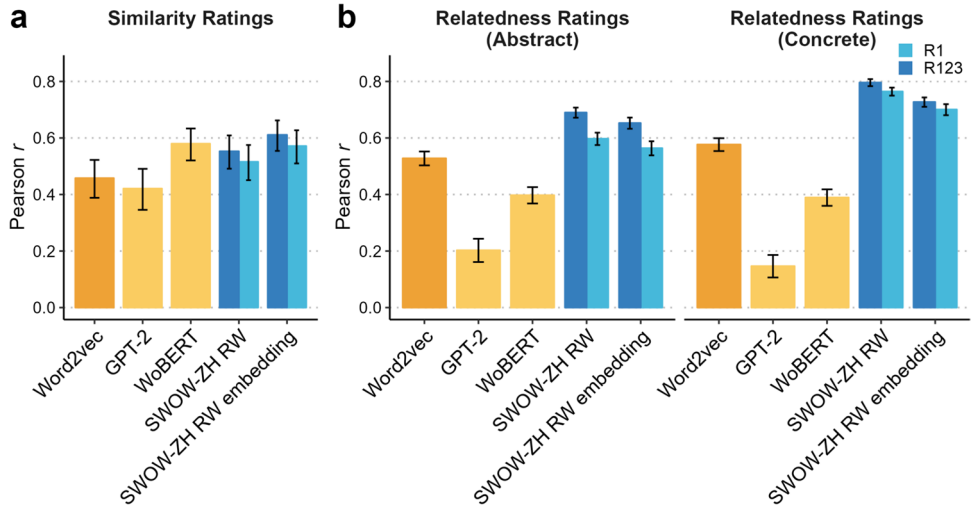

Fig. 4 SWOW-ZH word relation estimates significantly correlated with human similarity and relatedness ratings. The y-axis shows the correlation coefficient *r* between human ratings and word relation estimates driven from SWOW-ZH (blue) and word embeddings (orange). **a** Correlations with similarity ratings. **b** Correlations with relatedness ratings; left: abstract word pairs, right: concrete word pairs. The error bars show the 95% confidence intervals based on 1000 bootstraps

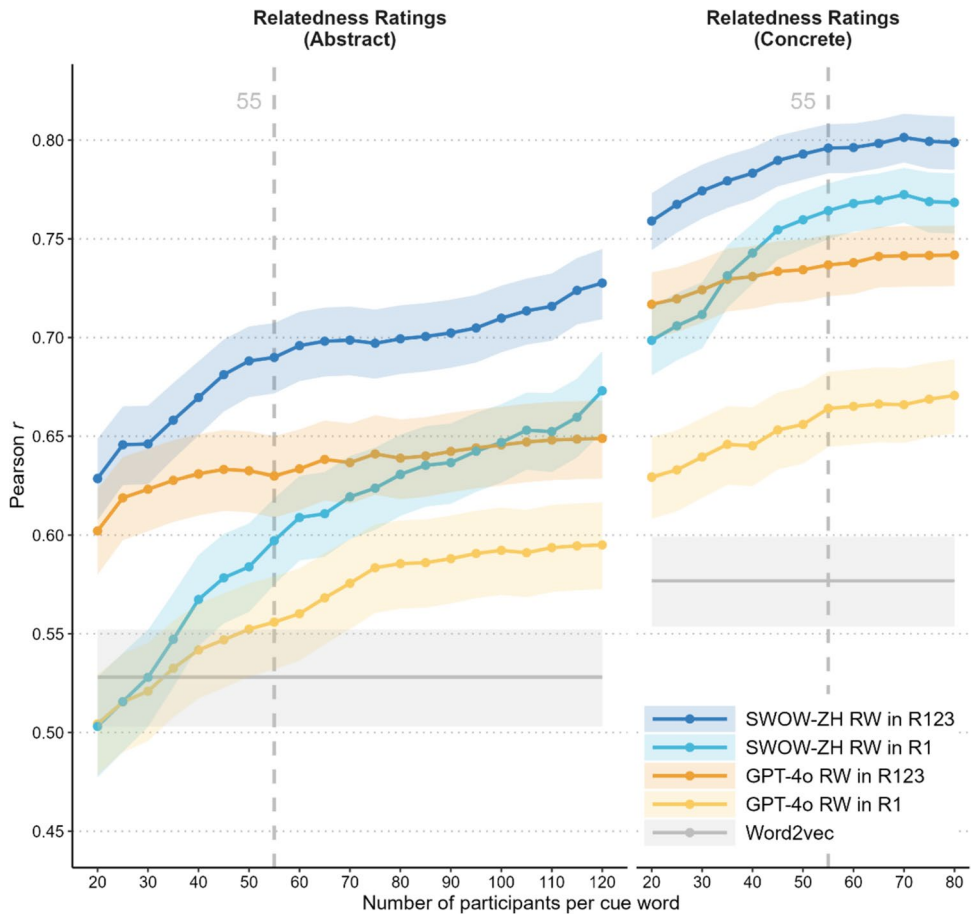

Fig. 5 The correlations between human relatedness ratings and word relation estimates derived from SWOW-ZH (blue) and GPT-4o (orange) associations. The correlations increase with the number of participants (or responses) per cue word and nearly stabilize until achieving a large sample size. Left: abstract word pairs; right: concrete word pairs. The vertical dashed lines mark the performance of the balanced dataset, i.e., 55 participants per cue. The shaded area represents the 95% confidence interval based on 1000 bootstraps for the curve

featured a significant number of hapax legomena (rare words occurring once) and a smaller number of hub words (frequently associated words). Hub words, which are connected to fundamental human concepts such as "man" and "money" although some task-related terms like "game," were also observed.

The results of fitting the finite Zipf-Mandelbrot model indicate that the SWOW-ZH (words and lexical relationships) follows an adjusted power-law distribution, reflecting that in the network, a minority of nodes (a few words) have very high connectivity, while the majority of nodes have low connectivity. This distribution characteristic corresponds to the typical features of scale-free networks, where the coexistence of a few hubs and a large number of low-degree nodes is a hallmark of scale-free networks. Traditionally, scale-free networks exhibit a power-law distribution (Lynn & Bassett, 2020). By adjusting the traditional Zipf model, the finite Zipf-Mandelbrot model is more applicable to finite systems or datasets like lexical networks (Newman, 2005), making it particularly valuable for understanding networks like SWOW. This enhances our ability to better predict and explain various phenomena in lexical networks.

The finding that SWOW networks possess both small-world and scale-free properties aligns with established research. This network structure, with a high degree of clustering and short paths, significantly enhances the efficiency and flexibility of language processing, facilitating effective management of complex language information and swift adaptation to linguistic changes. Notably, Steyvers and Tenenbaum (2005) emphasize that the integration of small-world and scale-free network features not only boosts the efficiency of language transmission, but also heightens sensitivity to the loss of key vocabulary. It demonstrates how natural languages evolve in an organized manner, prioritizing efficiency and robust connectivity over random growth. The concept of preferential attachment in vocabulary growth, discussed by Perc (2014), reflects an inherent aspect of language evolution where linguistic structures are continuously refined over time to improve communicative effectiveness. Future research

could further utilize scale-free network theory to analyze semantic or vocabulary networks, thereby enhancing our understanding of language structure and evolution. Furthermore, incorporating SWOW data from various age groups, along with additional annotations on SWOW edge attributes, and integrating data on the oral and reading language environments of children at different stages (see for example CCLOWW and CCLOOW databases, Li et al., 2023a, b), provide valuable avenues to determine how vocabulary grows.

Advantages of the three-response association paradigm

The three-response continued association paradigm used in SWOW-ZH has been validated across other language SWOW datasets, demonstrating its efficiency in eliciting a broader spectrum of language processing. The higher growth rate of R123, compared to R1, suggested that heterogeneous types can be obtained by using the multiple response free association task in Chinese, which is consistent with the results of previous studies in other languages (Cabana et al., 2024; De Deyne et al., 2013, 2019). The paradigm's effectiveness is also reflected in its ability to consistently predict human-perceived relations, with R123 maintaining a substantial advantage over R1, indicating a closer alignment with natural word retrieval processes.

The inclusion of multiple associations in this paradigm not only diversifies the responses but also enhances the task's capacity to accurately estimate relationships between words, as suggested by De Deyne et al. (2013). This is especially important for cue words that have a very strong first associate (e.g. *umbrella–rain*), providing a better approximation of the associative distribution that might be overlooked in a single-response setup.

Response chaining effects

While the continued word associations ask participants to only respond to the cue word (as opposed to the *continuous* version of the task), it is likely that the responses for the second and third associate are affected by previous responses. While such chaining effects increase response variability, the prevalence of this effect across responses is generally low to moderate, as shown by Bayes factors. In SWOW-ZH, we calculated the corresponding Bayes factor for all possible cue–R1–R2 triples (see also De Deyne et al., 2019) and found that 66.3% of the Bayes factors of the R1–R2 responses fell into the low range between 0 and 3, and 32.9% fell into a moderate chaining effect range between 3 and 10. The latter is slightly higher than SWOW-EN. This effect most likely reflected the unique relationships between characters and words in Chinese, as well as the word formation effect, which will be further discussed in the subsequent section about "Chinese specificity." The effect may also be partly due to the task's heterogeneity, as many R2 responses did not replicate R1, resulting in moderate Bayes Factor values. Take the cue word "bamboo," for instance, where "panda" appears in 79% of R1 responses. In such cases, the dominance of a response like "panda" tends to obscure other strong associations, yet the continued three-response procedure enables a comprehensive assessment of the entire range of responses. The diversity of R2 responses, often not mirroring R1, serves to diminish the effects of these response chains, thereby preserving the significant advantages of the multiple-response approach in linguistic research (Kumar et al., 2021; Maxwell & Buchanan, 2020). Furthermore, De Deyne & Storms, who coded the semantic relations between cues and responses, found that later R2 and R3 responses are also qualitatively different from earlier ones, suggesting a time-course where certain types of information (e.g. taxonomic) are accessed earlier than others (e.g. featural, thematic). As such, the continued procedure offers several pathways to reveal complex patterns in word associations, including cross-linguistic and cross-cultural studies.

Multimodal gains brought by associations

Multiple-response free association not only enhances word representation by aligning closely with the spreading activation mechanism, but also incorporates a broader range of psychological processes, including visual and affective knowledge (De Deyne et al., 2021). This enrichment is evident in the improved estimation of emotional word ratings such as valence and arousal (Van Rensbergen et al., 2016; Vankrunkelsven et al., 2015). In the current study, word associations achieved better performance in human relatedness and similarity ratings compared to text-based models, indicating a significant multimodal gain. The spreading activation in semantic retrieval for R123 involves a wider array of words, including those distantly connected. The results are consistent with De Deyne et al. (2016), who found that spreading activation in SWOW-EN aligns well with human behavior in triads of weakly connected words. Recent fMRI studies using representational similarity analysis (RSA) also provided evidence for the multimodal gains of SWOW-ZH. Yang et al. (2024) showed that, compared to text-based language models, the semantic relations computed with SWOW-ZH more closely approximate the patterns of neural activation, likely encoding multimodal representations that incorporate both distributed and experiential information. Compared to single words *per se* or the representative instances of the word's category, community/category-level RSA demonstrated the broadest involvement of brain regions, engaging areas critical for semantic processing, including the angular gyrus, superior frontal gyrus, and a large portion of the anterior temporal lobe.

Word association patterns and word frequency

The relationship between word associations and word frequency is complicated. When mapping word frequencies from

SWOW-ZH, Web 5-g, and SUBTLEX-CH onto a frequency spectrum, similar patterns emerge across these measures (Fig. 2c). However, this similarity does not necessarily imply that word associations are primarily influenced by real-life usage frequency. Notably, while high-frequency words are prevalent, emergent words at the low-frequency end of the distribution, though less common, highlight the diverse contexts in which these rarer words are used (Cai & Brysbaert, 2010). Although speculative at this time, one possibility is that rare words require more context to be retrieved as response, which could be investigated in follow-up studies. More generally, the consistency between association and text-based frequency suggests some degree of underlying similarity mechanism operates in both word retrieval during association tasks and in language production in both writing and conversation.

Chinese specificity: Comparisons with SWOW in other languages

Chinese, as a logographic language, differs significantly from the alphabetic languages of previous SWOW norms (Dutch, English, and Rioplatense Spanish). In exploring the effects of language-specific features on word association, the SWOW-ZH data showed notable differences from earlier datasets. Chinese participants encountered fewer unfamiliar cues, which might reflect the fact that new words are easily formed by recombining existing characters. Conversely, Chinese participants found it more challenging to provide multiple associations compared to their English and Dutch counterparts, likely due to the lack of inflectional morphology in Chinese, which limits response variations through changes in gender, number, or conjugation. These types of responses are more common in alphabetic systems.

Chinese is distinguished by its frequent use of compounds that are formed without obvious word boundaries and the absence of morphological markers that identify parts of speech. This structural feature impacts lexical access, as evidenced by the analysis of lexical decision times, which reveal distinct retrieval patterns for different parts of speech. Our results also showed that the naming speed of single-character words is influenced not only by the word's CD (as a whole word, SUBTLEX CD) but also by the CD of the underlying character (as a character that serves as a component of words, SUBTLEX CCD), while the naming speed of two-character words is affected by the CD of the word and the frequency of the constituent characters. Analysis of the SWOW-ZH network also revealed approximately 30% of associations might involve word formation, underscoring the important role of word formation effects in Chinese. The multi-character composition of Chinese words suggests that the semantic and phonological processing of individual characters may influence the associative relationships formed by multi-character words, highlighting a unique aspect of Chinese lexical retrieval. This complex interplay of characters within words provides a rich area for further exploration in understanding how word associations reflect cognitive-linguistic processes.

Furthermore, cultural imprints in languages, particularly the emphasis on the relationships between objects and their environments, might influence these findings. This cultural aspect suggests a stronger contextual dependence in Chinese, which is likely reflected in the word association norms where word meanings are more influenced by associated words. Such reliance on context in Chinese is supported by studies showing that children's lexical growth heavily depends on contextual factors (Cox & Haebig, 2023). Despite these differences, the variations in lexical growth curves between Chinese and alphabetic languages suggests that the current sample is adequate, depending on the use case. Especially in studies focusing on similarity and relatedness, the use of the random-walk-based approach, which accounts for both direct and indirect associative links, addresses the sparsity problem by providing more robust relation estimates even with fewer direct associates.

Extending the SWOW family to include Mandarin Chinese allows for broader investigations into the mechanisms of language processing. Incorporating a large-scale Chinese word association norms, along with other psycholinguistic databases, could enhance our understanding of the cognitive mechanisms involved in language generation and comprehension. Echoing the aforementioned linguistic annotation across languages, one approach to addressing the cultural differences implicit in languages involves aligning words with identical meanings across different languages and comparing variations in the semantic dimensions (Wang et al., 2023) or affective connotations (De Deyne et al., 2020) of the same concept across cultures. The ongoing SWOW project will continue to facilitate research into both the specific linguistic features of Chinese and broader linguistic phenomena, offering valuable insights into the nature of language and cognition. On the other hand, it is noteworthy that Chinese features a large number of homophones, and the mapping of thousands of Chinese characters to just over 400 syllables results in a high density of homophones, introducing a particularly close phonological relationship between words, especially among single-character words. In other words, by adopting a multi-layer network perspective, the high density of homophones might enhance the layer of phonological associations among words (see also Hsiao & Shillcock, 2006; Siew & Vitevitch, 2020). These insights into the structure and function of the SWOW-ZH database not only enhance our understanding of word associations in Mandarin, but also provide unique perspectives for cross-linguistic and cross-cultural studies, emphasizing the specificity and complexity of Chinese language processing.

SWOW-ZH and language models

Our research substantiates the independent and significant contribution of association norms in predicting lexical processing and understanding subjective perceptions of word relationships, overall outperforming both text-based models and reference LLMs. It is important to note that these comparative analyses do not aim to position SWOW-ZH as an alternative to language models, including LLMs, due to their different objectives and distinctive characteristics. Theoretically, word associations cannot provide a full account, since they do not address questions about how meaning is acquired through language. Others have pointed out that method overlap could also contribute to the effect sizes when comparing associations with strongly related tasks such as relatedness judgements (see Kumar et al., 2021 for a critical discussion). However, our view is that performance is much more driven by overlap in representations than in overlap of procedures, given that versions of the task that presumably match the underlying processes more often result in poor performance (e.g. using direct measures of associative strength) compared to measures that actively address biases inherent to the task (response sparsity, response frequency biases).

Taking a more representational view, we believe SWOW-ZH captures nuanced aspects of human mental representations that are often overlooked or unrealistically portrayed in LLMs.

A second attribute that sets it apart from text-based approaches is the role of detailed demographic information in shedding light on variation within and between languages. As such, word association responses not only provide information about semantic and linguistic structures, but also embody cultural insights, unconstrained by typical norms of communication and revealing how people understand and connect different words within their cultural contexts. In other words, these responses are not merely automatic outputs of language, but are manifestations of cultural experiences and values expressed without the full set of communicative constraints of natural language.

This aspect underscores the broader applicability and relevance of our findings in capturing cultural nuances that typical LLMs might miss.

On the other hand, while SWOW outperforms traditional language models, we must acknowledge that most models studied here predominantly use characters rather than words as tokens in Chinese. This study's focus on words bypasses the usual "attention" mechanisms, potentially affecting the performance of these models. Moreover, direct comparisons of LLMs to human cognitive processes in tasks like relatedness prediction may not always be fair, due to the inherent similarities between association and relatedness judgment tasks (but see above). Notably, WoBERT, trained with word segmentation, displays performance comparable to SWOW-ZH in predicting similarity ratings. This performance may stem from both its architectural design and the advantages of word segmentation. Despite GPT-4o's extensive training, it did not show enhanced performance with an expanded sample size. However, the potential utility of LLMs remains significant. By integrating human association norms with simulated three-response association tasks, GPT-4o presents a viable alternative to traditional word embeddings, especially when large-scale human associations are challenging to gather. Studies such as those by Hansen and Hebart (2022) also suggest that LLMs can excel in specific aspects of semantic processing like semantic feature generation, achieving results comparable to human performance.

To sum up, our results emphasize that "big data" from large-scale human experiments could offer unique contributions in multiple aspects, significantly benefiting natural language processing and cognitive research. Models based on human-generated data can serve as complementary resources to LLMs and other advanced algorithms, helping us understand the complexity and richness of language more effectively.

The SWOW-ZH dataset is now publicly accessible on the SWOW website, which hosts the official and current release. For further details, please refer to the "Availability of data and materials" section. We provide access to both the raw data and the preprocessed data on the website. Users can either adapt the raw data to fit their own needs or use the balanced data preprocessed through our pipeline (Fig. 1). Note that the results presented in this article are based on the preprocessed balanced data, where the number of participants per cue is set to 55. Additional data are included in the raw dataset, where the number of participants per cue averages about 76, varying depending on the snowball iteration. Since the SWOW-ZH project is continuously expanding, users can optionally subscribe to a newsletter to receive news about updates and new releases via the SWOW website.

SWOW-ZH illustrates the importance of large-scale word associations in understanding language, particularly highlighting its utility for Mandarin Chinese. By addressing notable gaps in linguistic data and offering insights that could enhance language processing models, this study underscores the potential benefits of large-scale human experimental data in exploring language's complexity. It suggests that human-generated big data can be a valuable resource in advancing research in linguistic and cognitive science.

Appendix 1

Details of the procedures

On the SWOW website, after reading a brief introduction (see Appendix Fig. 6a), participants were asked for consent and demographic information (Appendix Fig. 6b). The instructions are shown in Appendix Fig. 6c. Two rules were highlighted: The response word has to be the first word that comes to mind spontaneously; the response word has to be associated

with the cue rather than previous responses. The exact instructions in Mandarin Chinese are included in the data repository.

A trial consisted of a cue and three response boxes. Participants were required to enter the first (R1), second (R2), and third (R3) responses for each cue in order. Pressing "Enter" allowed participants to move to the next response box. After completing all three responses, they could click the "Finish" button to proceed to the next cue. If participants did not know the target cue, they could skip the trial by clicking "Unknown Word"; or if they were unable to come up with a word in a response box, they could skip the remaining response boxes by clicking "No More Responses" (see Appendix Fig. 6d).

A full session included 18 trials, semi-randomly selected based on the number of observations in the database. At the end of the session, participants could optionally provide their email addresses to receive information about the study's progress. They could also choose to proceed to another session or learn more about the SWOW project. Participants were able to participate in the experiment as many times as they wished.

Due to the unreliable access to the main SWOW website experienced in mainland China over the past few years, our subsequent data collection has been shifted to the NAODAO platform since November 29, 2022. On NAODAO, each participant completed a three-response association for 80 to 90 cues. Unlike trials on the main SWOW website, where participants clicked a button to skip a trial, NAODAO participants manually entered "000" to substitute "No More Responses" or "Unknown Word" to skip a trial. Responses containing "000" were classified as "Unknown Word" if "000" was entered in the first box; otherwise, they were considered to include "No More Responses". Participants could take breaks as desired (procedures outlined in Appendix Fig. 6e-h). Completing the task took approximately 30 min, with participants receiving payment of 15 RMB (approximately 2.2 USD). Other procedures on NAODAO mirrored those on SWOW.

To evaluate reliability across different sample sizes, we collected an additional set of responses for 328 cue words on NAODAO (see Relatedness and similarity rating tasks in the Methods—External Validation section). These cues comprised 82 abstract and 82 concrete words initially used in De Deyne et al. (2020) 's relatedness judgment task. We combined these with 164 other words randomly selected from SWOW (as 'filling words') to ensure a diverse range and variability in concreteness.

Fig. 6 Data collecting procedure. On the SWOW website: (**a**) A brief introduction to the SWOW-ZH project and informed consent process; (**b**) Participants' information; (**c**) Instructions; and (**d**) An example trial, where the cue was presented at the top, followed by three response boxes, an "Unknown Word" button at the bottom-left, and a "No More Responses" button, which replaces the "Unknown Word" button after the first response is entered. On NAODAO: The procedures were identical to those on the SWOW website

Instead of the 55 participants for each cue in the 'balanced' version, we expanded the number of trials/participants to 80 for concrete words and 120 for abstract words. Each participant completed either 80 or 82 cues, with half of them originating from De Deyne et al.'s (2020) task. The number of abstract and concrete words was balanced within participants.

To compare associations between humans and Large Language Models, we presented the three-response free association tasks for 164 cue words to GPT-4o-2024–08-06. Each cue word was presented 120 times. The prompt was adapted from the human version, highlighting its effortlessness. The full instruction reads, "I'll give you a word or character. Enter the first three words or characters that come to mind without thinking. Use commas to separate them. Note: only provide the words or characters, do not explain or add any extra information. (我会给你一个词或者字。看到我给你的词或者字以后, 输入你的脑海里没有经过思考就浮现出来的三个词语或者字,用半角逗号隔开。注意, 只给出词或字即可, 不要解释, 不要添加多余信息。)" We accessed GPT-4o-2024–08-06 via its API, adjusting the fine-tuning parameters as detailed in the Methods—External Validation section. The same preprocessing steps were applied to ensure comparability, resulting in 0.86% missing R1, 1.57% missing R2, and 2.42% missing R3. All the codes are available on request.

Appendix 2

Details of the preprocessing

Stage 1: Data Merging. The cues comprised words from 10 snowball iterations (Set 1 to Set 10), supplemented with data collected from NAODAO. Cues from the ongoing data collection (Set 11) on the SWOW platform were excluded. Furthermore, participants under 16 were excluded from the raw data. Eighty-five taboo words in response types were masked with hexadecimal codes, and 19 taboo words in the cues were removed.

Stage 2: Data Cleaning—Words. The cues and responses that met the following exclusion criteria were either masked by their problem categories or corrected to their proper form. These adjustments reduced 16,884 response types, accounting for 10.14% of the response types before Stage 2.

The following seven types of problematic words were processed sequentially, without replacement, ensuring that each word was processed no more than once:

1. Traditional Chinese cues and responses were transformed into simplified equivalents based on the Open Chinese Convert library.
2. English words, commonly used by Chinese people and integral to Chinese contexts, were retained and processed following the procedures described in De Deyne et al. (2019).
3. Joined responses, where participants typed two or more responses in a single box, separated by punctuation or symbols, were sequentially divided into individual responses. Only the first three were processed. In exceptional cases, if the separated responses contained long responses or symbols, they were passed to the next suitable case as a whole, marked as either a long response or symbol.
4. Responses exceeding six characters were labeled #Long, except for meaningful long words appearing at least twice. Meaningless long responses were defined as character strings requiring the addition or deletion of at least one character to form a coherent phrase. Common expressions integral to the Chinese lexicon were manually defined as long words and treated as regular words, such as titles of masterpieces, professional terms, colloquialisms, and sentences from classical poetry (e.g. "第二次世界大战" [World War II], "平面直角坐标系" [Plane Rectangular Coordinate System]).
5. Responses containing non-Chinese characters (letters, symbols, numbers, or punctuation) were modified and retained if they were meaningful and appeared more than once, otherwise labeled as #Symbol.
6. Retroflex final (erhua or erization) was deleted from responses, as it pertains to pronunciation features and is often omitted in typing (e.g., "一点儿" to "一点" [a little] and "女孩儿" to "女孩" [girl]).
7. Duplicated responses within trials were marked as #Repeat.

Stage 3: Data Cleaning—Participants. Participants who met at least one of seven exclusion criteria were excluded from the final analysis. This led to the removal of 10,345 participants (25.29% of the total before Stage 2), 24,498 response types (16.37%), and 186,400 trials (24.04%). The seven exclusion criteria, applied in strict order, were:

1. Participants with more than 40% of responses not found in a lexicon of unigram characters (Liu et al., 2010) and the word frequency database (SUBTLEX-CH) (Cai & Brysbaert, 2010). In this case, many responses could be word pairs, sentences, or meaningless character strings.
2. Participants with more than 30% of responses marked #Long.
3. Participants with more than 40% of responses marked #Symbol.
4. Participants with more than 40% of responses in English, indicating potential non-native or non-proficient Chinese speakers.
5. Participants with more than 20% of responses marked as #Repeat, including duplicated responses under the same cue and across cues.

6. Participants with more than 60% of responses as "Unknown Word" or "No More Responses".
7. Cantonese participants identified by their dialect. Due to a coding error, southwestern dialects, southern dialects, and Wu dialects were grouped as southern dialects, complicating the identification of Cantonese speakers. To address this, affected participants (grouped as speakers of southern dialects) were excluded from the analyses. To maintain balance in the dataset following these exclusions, additional data were collected on NAODAO.

Stage 4: Data Balancing. The number of observations per cue was balanced to calculate word centrality unbiasedly, ensuring that each cue had an equal chance to capture an association and avoiding any skewing of the results due to some cues being responded to more frequently than others. This same principle applies when computing word relations. We marked "Unknown Word" as #Unknown, and "No More Responses", #Long, #Symbol, and #Repeat as #Missing. We retained 55 responses for each cue, prioritizing participants with fewer missing values. In the case of having an equivalent number of responses labeled as #Missing, we prioritized participants with fewer #Missing responses who reported Mandarin as their native language. This removed 2,791(2.23%) response types and 37,586 (6.38%) trials, leading to the complete removal of 54 (0.18%) participants.

Additionally, a file titled Supplementary_Tables.xlsx file was provided, including the results of partial correlations in the Lexical Decision Task section and Pearson correlations in the Relatedness and Similarity Rating Tasks section.

Code availability The preprocessing codes and codes of secondary measures derived from SWOW-ZH (word centrality, word relation) in both MATLAB and R are available on GitHub (https://github.com/lib314a/SWOWZH). The MATLAB version used is R2021a, and the R version is 4.2.2. The network analysis in MATLAB was conducted using the Brain Connectivity Toolbox in v1.1.1.0 (http://www.brain-connectivity-toolbox.net). The *ropencc* library (https://github.com/Lchiffon/ropencc, based on v0.1 of the Open Chinese Convert library) was utilized to convert traditional Chinese into simplified Chinese. All codes are released under the Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Declarations

Ethics approval The study was approved by the KU Leuven Ethics Committee (G-201407017).

Consent to participate All participants provided online informed consent via checkbox.

Consent for publication All participants involved in this study provided online informed consent for the publication of the data and findings derived from this research. Participants were assured that their identities would remain confidential and stored anonymously.

Open practices statement The data and materials for all experiments are available at https://smallworldofwords.org/zh/project/research and none of the experiments was preregistered.

Competing interests The authors have declared no competing interests.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-024-02513-1.

Acknowledgements The authors extend their heartfelt gratitude to the participants who generously contributed to the SWOW-ZH. Their collective efforts make this dataset possible.

Authors' contributions Simon De Deyne and Qing Cai conceived the experiments. Bing Li, Ziyi Ding, Qing Cai and Simon De Deyne conducted the experiments, Bing Li and Ziyi Ding analyzed the results, Bing Li, Ziyi Ding, Simon De Deyne and Qing Cai drafted the manuscript. Qing Cai and Simon De Deyne revised the manuscripts.

Funding This work was funded by the National Natural Science Foundation of China (31970987 to Qing Cai), the Fundamental Research Funds for the Central Universities (to Qing Cai), and the Australian Research Council Early Career Grant (DE140101749 to Simon De Deyne).

Data availability The SWOW-ZH dataset is publicly available on the SWOW website (https://smallworldofwords.org/zh/project/research), which provides the official and up-to-date release. All data are released under Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) License. The dataset is available for non-commercial use only. When referring to the dataset, please use the "SWOW-ZH" acronym and include the release date (2024).

A code-free option is also available on the SWOW site, including word centralities and word-pair relation estimates based on SWOW-ZH.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review, 122*(5), 558–569. https://doi.org/10.1037/a0038693

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115*(1), 214–227. https://doi.org/10.1037/0033-295X.115.1.214

Auguste, J., Rey, A., & Favre, B. (2017). Evaluation of word embeddings against cognitive processes: Primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP* (pp. 21–26). https://doi.org/10.18653/v1/W17-5304

Baayen, R. H. (2001). *Word frequency distributions*. Springer.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition ofsingle-syllable words. *Journal of Experimental Psychology: General, 133*(2), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language, 125*(1), 47–53. https://doi.org/10.1016/j.bandl.2013.01.005

Baroni, M., & Evert, S. (2014). The zipfR package for lexical statistics: A tutorial introduction [R package documentation]. Retrieved

from http://mirrors.nic.cz/R/web/packages/zipfR/vignettes/zipfr-tutorial.pdf. Accessed 2024.4.8.

Bever, T. G., Chomsky, N., Fong, S., & Piattelli-Palmarini, M. (2023). Even deeper problems with neural network models of language. *Behavioral and Brain Sciences, 46*, e387. https://doi.org/10.1017/S0140525X23001619

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review, 124*(1), 1–20. https://doi.org/10.1037/rev0000047

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments & Computers, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Cabana, Á., Zugarramurdi, C., Valle-Lisboa, J. C., & De Deyne, S. (2024). The "Small World of Words" free association norms for Rioplatense Spanish. *Behavior Research Methods, 56*(2), 968–985. https://doi.org/10.3758/s13428-023-02070-z

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

Cañs, J. J. (1990). Associative strength effects in the lexical decision task. *The Quarterly Journal of Experimental Psychology Section A, 42*(1), 121–145. https://doi.org/10.1080/14640749008401211

Chen, X., Gao, X., Yan, X., Du, M., Zang, Y., & Wang, Y. (2023). Online research in psychology and its future in China. *Journal of Psychological Science*, *46*(5), 1262–1271. https://doi.org/10.16719/j.cnki.1671-6981.20230529

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Cox, C. R., & Haebig, E. (2023). Child-oriented word associations improve models of early word learning. *Behavior Research Methods, 55*(1), 16–37. https://doi.org/10.3758/s13428-022-01790-y

De Deyne, S., Cabana, Á., Li, B., Cai, Q., & McKague, M. (2020). A cross-linguistic study into the contribution of affective connotation in the lexico-semantic representation of concrete and abstract concepts. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Humans, Animals, and Machines* (pp. 2776–2782). Cognitive Science Society.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science, 45*(1), e12922. https://doi.org/10.1111/cogs.12922

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods, 51*, 987–1006. https://doi.org/10.3758/s13428-018-1115-7

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General, 145*(9), 1228–1254. https://doi.org/10.1037/xge0000192

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods, 45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods, 40*(1), 213–231. https://doi.org/10.3758/BRM.40.1.213

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). Springer Netherlands. https://doi.org/10.1007/978-90-481-8847-5_10

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Hansen, H., & Hebart, M. N. (2022). Semantic features of object concepts generated with GPT. *arXiv*. https://doi.org/10.48550/arXiv.2202.03753

Hills, T. T., & Kenett, Y. N. (2022). Is the mind a network? Maps, vehicles, and skyhooks in cognitive network science. *Topics in Cognitive Science, 14*(1), 189–208. https://doi.org/10.1111/tops.12570

Hofmann, M. J., Müller, L., Rölke, A., Radach, R., & Biemann, C. (2020). Individual corpora predict fast memory retrieval during reading. *arXiv*. https://arxiv.org/abs/2010.10176

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, G., & Mitchell, D. (2022). lmtest: Testing linear regression models [R package documentation]. Retrieved from https://cran.r-project.org/web/packages/lmtest/index.html. Accessed 2024.4.3.

Houghton, C., Kazanina, N., & Sukumaran, P. (2023). Beyond the limitations of any imaginable mechanism: Large language models and psycholinguistics. *Behavioral and Brain Sciences, 46*, e395. https://doi.org/10.1017/S0140525X23001693

Hsiao, J. H., & Shillcock, R. (2006). Analysis of a Chinese phonetic compound database: Implications for orthographic processing. *Journal of Psycholinguistic Research, 35*(5), 405–426. https://doi.org/10.1007/s10936-006-9022-y

Ji, L.-J., Peng, K., & Nisbett, R. E. (2000). Culture, control, and perception of relationships in the environment. *Journal of Personality and Social Psychology, 78*(5), 943–955. https://doi.org/10.1037/0022-3514.78.5.943

Johnson, D. R., & Hass, R. W. (2022). Semantic context search in creative idea generation. *The Journal of Creative Behavior, 56*(3), 362–381. https://doi.org/10.1002/jocb.534

Katz, L., Brancazio, L., Irwin, J., Katz, S., Magnuson, J., & Whalen, D. H. (2012). What lexical decision and naming tell us about reading. *Reading and Writing*, 25(6), 1259–1282. https://doi.org/10.1007/s11145-011-9316-9

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). Semantic memory search and retrieval in a novel cooperative word game: A comparison of associative and distributional semantic models. *Cognitive Science, 45*(10), e13053. https://doi.org/10.1111/cogs.13053

Li, L., Zhao, W. T., Song, M., Wang, J., & Cai, Q. (2023a). CCLOOW: Chinese children's lexicon of oral words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02077-6

Li, L., Yang, Y., Song, M., Fang, S., Zhang, M., Chen, Q., & Cai, Q. (2023b). CCLOWW: A grade-level Chinese children's lexicon of written words. *Behavior Research Methods., 55*(4), 1874–1889. https://doi.org/10.3758/s13428-022-01890-9

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 138–143). https://doi.org/10.18653/v1/P18-2023

Liu, F., Yang, M., & Lin, D. (2010). *Chinese web 5-gram version 1* . Linguistic Data Consortium. https://doi.org/10.35111/647p-yt29

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39*(2), 192–198. https://doi.org/10.3758/BF03193147

Lynn, C. W., & Bassett, D. S. (2020). How humans learn and represent networks. *Proceedings of the National Academy of Sciences, 117*(47), 29407–29415. https://doi.org/10.1073/pnas.1912328117

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Maxwell, N. P., & Buchanan, E. M. (2020). Investigating the interaction of direct and indirect relation on memory judgments and retrieval. *Cognitive Processing, 21*(1), 41–53. https://doi.org/10.1007/s10339-019-00935-w

Meersmans, K., Bruffaerts, R., Jamoulle, T., Liuzzi, A. G., De Deyne, S., Storms, G., Dupont, P., & Vandenberghe, R. (2020). Representation of associative and affective semantic similarity of abstract words in the lateral temporal perisylvian language regions. *NeuroImage, 217*, 116892. https://doi.org/10.1016/j.neuroimage.2020.116892

Meersmans, K., Storms, G., De Deyne, S., Bruffaerts, R., Dupont, P., & Vandenberghe, R. (2022). Orienting to different dimensions of word meaning alters the representation of word meaning in early processing regions. *Cerebral Cortex, 32*(15), 3302–3317. https://doi.org/10.1093/cercor/bhab416

Nelson, D. L., Mcevoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition, 28*(6), 887–899. https://doi.org/10.3758/BF03209337

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Nisbett, R. E., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences, 100*(19), 11163–11170. https://doi.org/10.1073/pnas.1934527100

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences, 9*(10), 467–473. https://doi.org/10.1016/j.tics.2005.08.004

Packard, J. L. (2000). The morphology of Chinese: A linguistic and cognitive approach. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511486821

Perc, M. (2014). The Matthew effect in empirical data. *Journal of the Royal Society Interface, 11*(98), 20140378. https://doi.org/10.1098/rsif.2014.0378

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review, 15*(1), 161–167. https://doi.org/10.3758/PBR.15.1.161

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46*(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language, 134*, 104472. https://doi.org/10.1016/j.jml.2023.104472

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Siew, C. S. Q., & Vitevitch, M. S. (2020). An investigation of network growth principles in the phonological language network. *Journal of Experimental Psychology: General, 149*(12), 2376–2394. https://doi.org/10.1037/xge0000876

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-17)* (pp. 4444–4451). https://doi.org/10.1609/aaai.v31i1.11164

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

Szalay, L. B., & Deese, J. (1978). Subjective meaning and culture: An assessment through word associations. *Lawrence Erlbaum Associates*. https://doi.org/10.4324/9781003470236

Tsang, Y.-K., Huang, J., Lui, M., Xue, M., Chan, Y.-W.F., Wang, S., & Chen, H.-C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods, 50*(5), 1763–1777. https://doi.org/10.3758/s13428-017-0944-0

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods, 49*(4), 1503–1519. https://doi.org/10.3758/s13428-016-0810-5

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. arXiv. https://doi.org/10.48550/arXiv.1908.08962

Ufimtseva, N. V. (2014). The associative dictionary as a model of the linguistic picture of the world. *Procedia - Social and Behavioral Sciences, 154*, 36–43. https://doi.org/10.1016/j.sbspro.2014.10.108

Van Rensbergen, B., De Deyne, S., & Storms, G. (2016). Estimating affective word covariates using word association data. *Behavior Research Methods, 48*(4), 1644–1652. https://doi.org/10.3758/s13428-015-0680-2

Vankrunkelsven, H., Verheyen, S., De Deyne, S., & Storms, G. (2015). Predicting lexical norms using a word association corpus. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2463–2468). https://lirias.kuleuven.be/1786054

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews, 35*(3), 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., & Poibeau, T. (2020). Multi-Simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics, 46*(4), 847–897. https://doi.org/10.1162/coli_a_00391

Wang, S., Zhang, Y., Shi, W., Zhang, G., Zhang, J., Lin, N., & Zong, C. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data, 10*(1), 106. https://doi.org/10.1038/s41597-023-01995-6

Wong, T. Y., Fang, Z., Yu, Y. T., Cheung, C., Hui, C. L. M., Elvevåg, B., De Deyne, S., Sham, P. C., & Chen, E. Y. H. (2022). Discovering the structure and organization of a free Cantonese emotion-label word association graph to understand mental lexicons of emotions. *Scientific Reports, 12*(1), 19581. https://doi.org/10.1038/s41598-022-23995-z

Wulff, D. U., & Mata, R. (2022). On the semantic representation of risk. *Science Advances*, *8*(27), eabm1883. https://doi.org/10.1126/sciadv.abm1883

Yang, Y., Li, L., de Deyne, S., Li, B., Wang, J., & Cai, Q. (2024). Unraveling lexical semantics in the brain: Comparing internal,

external, and hybrid language models. *Human Brain Mapping, 45*(1), e26546. https://doi.org/10.1002/hbm.26546

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review, 18*(4), 742–750. https://doi.org/10.3758/s13423-011-0092-y

Zhang, M., Liu, Z., Botezatu, M. R., Dang, Q., Yuan, Q., Han, J., Liu, L., & Guo, T. (2023). A large-scale database of Chinese characters and words collected from elementary school textbooks. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02214-1