Language and Cognition (2024), 16: 4, 1674–1696
doi:10.1017/langcog.2024.22

ARTICLE

What ratings and corpus data reveal about the vividness of Mandarin ABB words

Thomas Van Hoey[1], Xiaoyu Yu[2], Tung-Le Pan[3] and Youngah Do[2]

[1]Department of Linguistics, KU Leuven, Leuven, Belgium; [2]Department of Linguistics, The University of Hong Kong, Hong Kong, SAR of China and [3]Graduate Institute of Linguistics, National Taiwan University, Taiwan
Corresponding author: Youngah Do; Email: youngah@hku.hk

(Received 27 July 2023; Revised 08 February 2024; Accepted 17 March 2024)

Abstract
A well-known method of studying iconic words is through the collection of subjective ratings. We collected such ratings regarding familiarity, iconicity, imagery/imageability, concreteness, sensory experience rating (SER), valence and arousal for Mandarin ABB words. This is a type of phrasal compound consisting of a prosaic syllable A and a reduplicated BB part, resulting in a vivid phrasal compound, for example, *wù-mángmáng* 雾茫茫 'completely foggy'. The correlations between the newly collected ABB ratings are contrasted with two other sets of prosaic word ratings, demonstrating that variables that characterize ABB words in an absolute sense may not play a distinctive role when contrasted with other types of words. Next, we provide another angle for looking at ABB words, by investigating to what degree rating data converges with corpus data. By far, the variable that characterizes ABB items consistently throughout these case studies is their high score for imageability, showing that they are indeed rightfully characterized as vivid. Methodologically, we show that it pays off to not take rating data at face value but to contrast it with other comparable datasets of a different phenomenon or data about the same phenomenon compiled in an ontologically different manner.

Keywords: ABB; corpus; ideophone; iconicity; imagery; norms; ratings; vividness

1. Introduction

Linguistic iconicity can be understood as the phenomenon whereby the meanings of words, signs and constructions are perceived directly from their formal aspects (Dingemanse et al., 2020; Haiman, 1985). On the lexical level, languages typically have onomatopoeias, such as English *crunch* 'eat food with a crackling sound' or Dutch *knor knor* 'sound of a pig' (Körtvélyessy & Štekauer, 2024). However, iconic depictions can extend beyond the sensory modality of sound, including modalities like textures, feelings, states, movements and smells (Dingemanse, 2012; Van Hoey,

2024), for example, Korean *polleok polleok* 'wind fluttering', or Basque *xiri miri* 'soft little drops of rain'. Such lexical items are usually called ideophones, which can be cross-linguistically defined as 'marked words that depict sensory imagery and which belong to an open lexical class' (Dingemanse, 2011, 2019).

This definition has led to a surge in iconicity ratings for iconic words as well as prosaic words (Hinojosa et al., 2021; Perry et al., 2015; Thompson et al., 2020; Winter et al., 2023; Winter & Perlman, 2021 among many others). These ratings are typically decontextualized, providing *an* insight into the way language users use and think of these words, but of course without presenting the full picture (see Winter, 2019, pp. 138–139 for discussion). One way of arriving at a more well-rounded picture would be to use a second source of information. Corpus data, containing words in context, are a plausible candidate, as has been shown in the domain of lexical and syntactic synonymy (Klavan & Divjak, 2016).

The methodological goals of this paper are:

(Goal 1) Collecting decontextualized ratings – not only in terms of iconicity, but also for other subjective variables such as familiarity, imageability, concreteness, valence, arousal and sensory experience ratings (SERs). After all, it has been established that many of these variables often correlate, for example, iconicity with sensory experience, abstractness and age of acquisition in Spanish (Hinojosa et al., 2021). Another common positive correlation is that between concreteness and imageability (Brysbaert et al., 2014; Della Rosa et al., 2010), though it is not always found (Song & Li, 2021). The different ratings we collected and present here provide a first descriptive look at how items within a category can be characterized.

(Goal 2) Comparing the decontextualized ratings for iconic words that we collected to prosaic words. This comparison aims to answer for which variables we see significant differences between these two groups. The rationale behind such a comparison is that the characterizing findings for the first goal in an absolute sense may turn out to be less useful in a relative manner, that is, when we want to characterize iconic words in opposition to prosaic words. Concretely, we will use binary logistic regression here to investigate which variables can aid in the discrimination between the two groups.

(Goal 3) Observing to what degree decontextualized ratings and contextualized corpora tap into similar underlying distributions within the data. This step approaches the absolute findings of the first goal from yet another angle, namely that of converging evidence based on an ontologically different data source. The method employed in this step will consist of principal component analysis (PCA), an exploratory statistical technique for the reduction of data dimensionality and the relative contributions of the variables.

In terms of data, we turn to a phrasal construction of Mandarin Chinese called ABB. The template of this construction consists of one prosaic syllable (A) followed by a reduplicated syllable (BB). Some examples include the following: *hēi-qīqī* 黑漆漆 'lacquer black', *wù-mángmáng* 雾茫茫 'completely foggy', *chì-luǒluǒ* 赤裸裸 'stark naked'. In these examples, the first term is prosaic, respectively, 'black', 'fog' and 'red (the color of skin)'. These ABB words are of special interest to the field of iconicity because they are situated somewhere along the cline between prosaic description and iconic depiction (Dingemanse, 2017). That is, while reference grammars of Chinese do not categorize these words as ideophones or iconic words (the term has not widely entered Chinese linguistics), they will describe them as particularly vivid in meaning (Huang et al., 2016). A recent study has argued that ABB items are prototypes of a

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

more general collocate–ideophone construction (Van Hoey, 2023). Additionally, seminal work has identified ABB items as a locus of fruitfully observing sound symbolism in Chinese (T'sou, 1978). Finally, they also occur in other Sinitic languages, such as Cantonese, Hakka and Southern Min (Chang, 2009; Mok, 2001). In other words, this paper takes the starting position that Chinese ABB items consist of a prosaic (A) and an iconic or iconized (BB) part.

Against this backdrop, ABB items make excellent candidates for the collection of ratings, the comparison with fully prosaic words and the investigation regarding the degree of convergence between rating and corpus data. The paper is structured as follows: in Section 2, we present the collection of subjective ratings for ABB words (goal 1). In Section 3, the comparison with two sets of prosaic words is made (goal 2). Section 4 introduces the corpus and corpus measures. Section 5 shows to what degree the corpus material and the ratings we collected tap into the same underlying structure within the data. Finally, we conclude in Section 6.

2. Rating ABB words

Many ratings have been collected for all sorts of words in the literature. We chose the following seven ratings for their well-researched nature and apply them here to ABB items. Familiarity ratings measure whether a stimulus is known from everyday life (Brown & Watson, 1987; Noble, 1953). Valence measures whether a stimulus is pleasant or unpleasant (Bradley & Lang, 1999). Arousal probes the intensity of feeling a stimulus evokes (Bradley & Lang, 1999). Imageability or imagery measures how well a stimulus gives rise to a mental image (Paivio et al., 1968; Rofes et al., 2018). Concreteness measures whether a stimulus is concrete or abstract (Brysbaert et al., 2014). SERs reflect the extent to which a stimulus evokes a sensory or perceptual experience in one's mind (Juhasz & Yap, 2013). Finally, iconicity ratings measure the degree to which a stimulus is perceived as resembling its meaning (Perry et al., 2015; Thompson et al., 2020; Winter & Perlman, 2021).

The materials are largely based on existing ABB wordlists. In a few pretest rounds, we reduced the number of stimuli so that false positives were weeded out. To assess the enterprise of this study, we then collected familiarity ratings. After an initial exploration of the data, we conducted the six other types of subjective ratings of the data: valence, arousal, sensory experience, imageability, concreteness and iconicity. We followed the best practices for the collection as presented in Winter (2022). Here, we present the distributions of ABB items along the seven ratings and the correlations between them.

2.1. Stimuli

We first collected ABB words for familiarity. We combined Van Hoey's (2023) dataset (N = 571) with the extensive list provided in Lǔ (1980) (N = 303). Despite item overlap (N = 111), the combined set contained many unique items (N = 763). We weeded out false positives, which are items that look like ABB but are actually names or other non-pertinent expressions, such as *kǒng-gāogāo* 孔高高, *Hú Zhēnzhēn* 胡珍珍 or *chuāng hǎohǎo*窗好好 (see Van Hoey, 2023 for a discussion). Additionally, we took out items that were judged as unfamiliar by four native speakers with a linguistic background and knowledge of the project, resulting in a

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

Table 1. Stratified sample of mean familiarity ratings, sorted from most familiar to least familiar

ABB	Pinyin	Meaning	Mean familiarity (0–6 scale)
凶巴巴	*xiōng–bābā*	'tough'	5.94
转圈圈	*zhuàn–quānquān*	'turning circles'	5.62
臭烘烘	*chòu–hōnghōng*	'stinky'	5.33
怯生生	*qiè–shēngshēng*	'timid'	5.00
白晃晃	*bái–huǎnghuǎng*	'bright'	4.41
赤条条	*chì–tiáotiáo*	'bare naked'	3.84
白乎乎	*bái–hūhū*	* 'white'	3.32
水淋淋	*shuǐ–línlín*	'dripping wet'	3.03
扑簌簌	*pū–sùsù*	'trickling down'	2.57
白苍苍	*bái–cāngcāng*	* 'white'	2.17
弯曲曲	*wān–qūqū*	'winding'	1.92
黄乎乎	*huáng–hūhū*	* 'yellow'	1.61
黑淋淋	*hēi–línlín*	* 'black'	1.22
冰僵僵	*bīng–jiāngjiāng*	* 'icy'	0.84
帷斜斜	*wéi–xiéxié*	* 'bent'	0.22

Note: Meanings with an asterisk (*) indicate meanings missing from dictionaries.

smaller set (N = 596). Because we decided to collect only the ratings from Mainland Chinese participants, we further merged items with two different orthographic variants in traditional Chinese but only one in simplified Chinese. For example, *xīng-chōngchōng* 'excited-*chongchong*' written as both <興沖沖> and <興衝衝> in traditional Chinese characters merged into <兴冲冲> in simplified characters. This resulted in a set of N = 564, which was used to collect all other ratings. Finally, we set a threshold of at least 30 ratings per item, and so the final set of words that were taken into account for the analysis is at N = 561. Many rating studies are often based on a much lower threshold (often situated between 10 and 15 ratings per word), which means that the data for our study are more generalizable. Table 1 contains a sample of items that are sorted from most familiar to least familiar.

2.2. Participants

We recruited native speakers of Mandarin Chinese. Most of them were students from the authors' institute and were offered an option to earn course credits as a reward. Additionally, nonstudent participants had a chance to earn up to 500 HKD in exchange for their contribution. For all ratings, the majority consisted of college-age females (ranging from 54% to 75% depending on the rating). In terms of geography, they originally came from all over Mainland China and Taiwan. They were also asked about other Chinese varieties they are fluent in besides Mandarin Chinese. Detailed demographic information is shown in the document '1_ratings' in the materials in the OSF repository (https://osf.io/tv34b/).

For data trimming, we followed Engelthaler and Hills (2018) by inspecting the standard deviation for potential low variability ($SD < 0.2$) per participant but we did not find any such 'categorical participants'. The number of participants per rating is presented in Table 2. The number of participants rating familiarity is relatively high due to shorter lists of items. There was a high inter-rater reliability across the rating tests, as their two-way averaged intra-class correlation (ICC(A,k), shows (see McGraw & Wong, 1996) implemented with the irrNA package (Bruekl & Heuer, 2022) in R.

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

Table 2. Number of participants, their intra-class correlation coefficient (ICC) and the mean times a stimulus was rated

Rating	Participants (*N*)	ICC (A, k)	Mean ratings per item
Familiarity	166	0.97 ***	37.23
Valence	85	0.96 ***	40.12
Arousal	83	0.87 ***	39.15
Ser	85	0.91 ***	40.27
Iconicity	78	0.95 ***	37.67
Imagery	71	0.95 ***	35.77
Concreteness	76	0.93 ***	34.33

Table 3. Descriptive statistics for mean item ratings

Rating	Mean	SD	Minimum	Maximum	Skewness	Most items are
Arousal	2.40	0.62	1.03	4.37	0.64	not very arousing
Concreteness	3.50	1.00	0.69	5.31	−0.70	quite concrete
Familiarity	3.15	1.65	0.22	5.95	0.13	relatively familiar
Iconicity	3.56	1.15	0.55	5.64	−0.46	quite iconic
Imageability	4.27	1.09	0.34	5.71	−1.20	very imageable
SER	3.30	0.85	0.60	4.91	−0.69	quite sensorial
Valence	2.76	0.96	0.69	5.10	0.42	not very positive

2.3. Procedure

We presented the questionnaires with ratings through Qualtrics (2022). Each rating was conducted on a 7-point scale, ranging from 0 to 6. The extremes of the Likert items were presented with the following text: 0 'not at all [RATING]' and 6 'completely [RATING]'. Only for the familiarity rating, we added 'relatively familiar' to the midway point (3). We chose to present these ordinal items as a small but positive scale, rather than a negative ranging to a positive, because previous studies (Winter, 2022) have shown that this may cause problems for interpretation. By using a clear 7-point scale ranging from 0 to 6, we ventured that the distances between each point can be treated as metric rather than as purely ordinal. This statistical practice is widespread for norming studies, and even though there are several reservations in doing so (Baayen & Divjak, 2017; Liddell & Kruschke, 2018), ratings tend to approximate normal distributions with fine-grained scales, and are thus amenable to standard linear models (Winter et al., 2023).

In each rating experiment, participants were first introduced to what the rating was about, and what the scale points stood for with examples (e.g., valence rating: 'rate 6 if the words make you feel extremely pleasant and 0 if it makes you feel extremely unpleasant'). Then, they were presented with a randomized set of ABB items for which they had to assign a score. Because we deployed our experiment online and also made it accessible on smartphones, we did not consider response time to be accurate across all participants, as there may have been differences in download speed. The instructions can be consulted in full in the OSF repository.

2.4. Results

It can be gleaned from Table 3 that most ratings have a mean around the halfway point of the scale, that is, 3, with minimums and maximums that span the edges of the

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

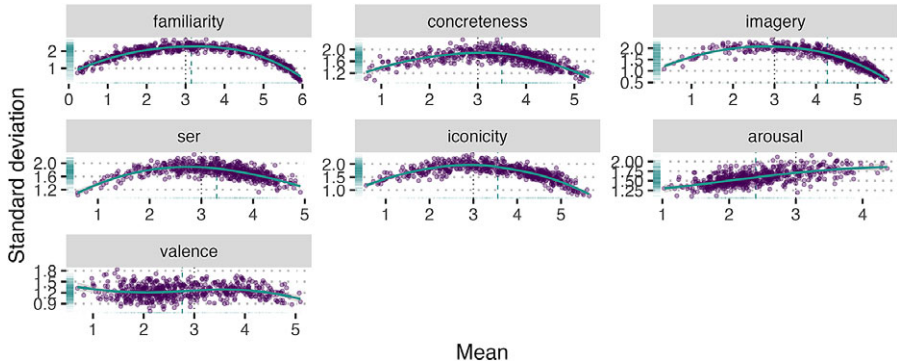

Figure 1. Plotting the mean and standard deviation of each rating against each other. The dashed vertical line indicates the mean of each rating set. The dotted line at (x = 3) indicates the midway point of rating. Shoe horse curves appear for familiarity, concreteness, imagery, SER, and iconicity, but not for arousal or valence.

scale, except for arousal. This suggests that the ABB inventory we used was reasonably exhaustive in terms of range. However, the mean ratings were not symmetrically distributed. Most items were slightly less familiar (skewness = 0.13). Overall, most items were judged to be less arousing and not positive. As expected, most items were judged to be quite concrete, sensorial, iconic and very imageable. This is the first indication that we are able to operationalize the vivid nature of ABB items (Huang et al., 2016; Wang, 2014) and provides further evidence that it makes sense to see them as constructions involving ideophones (Mok, 2001; T'sou, 1978; Van Hoey, 2023), that is, the iconic lexicon of Mandarin. In Section 3, we will show that vividness can be mostly reduced to high imagery ratings.

When we plotted the mean and standard deviation of each mean item rating against one another (Figure 1), we saw for most ratings a shoe-horse curve that has also been observed in other rating studies (Pollock, 2018; Winter et al., 2023; Xu & Li, 2020). Pollock (2018) argues that it is problematic to use values for items that have a high standard deviation in the design of further experiments, and we agree with that assessment. However, as this pattern is widespread in rating studies, it is also emblematic of the structures inherent in datasets. It makes sense that items at the polar ends have a lower standard deviation: all of our participants agree on the familiarity level of *xīong-bābā* 凶巴巴 versus *wèi-xiéxié* 帷斜斜 (see Table 1), while items in the middle are sometimes rated higher, sometimes lower. Like Winter concluded, if there is a follow-up task, it is better to use ratings cautiously (Winter, 2022). However, for charting the structure in a particular domain, all items should be taken into consideration.

Note that this shoe-horse curve (Figure 1) does not appear for valence and arousal. The trends of these items have also been found for English (Warriner et al., 2013) and Spanish (Stadthagen-Gonzalez et al., 2017). It appears that, decontextualized, it is difficult to assign a valence or a level of arousal to most items. Contrast this to ratings like imageability, SER, concreteness and iconicity: even decontextualized, most items score high in this regard and with a reasonable standard deviation range between the raters.

The different ratings are highly correlated with each other. In Figure 2, we present all pairwise Pearson correlations between the ratings. All of them were significant but

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

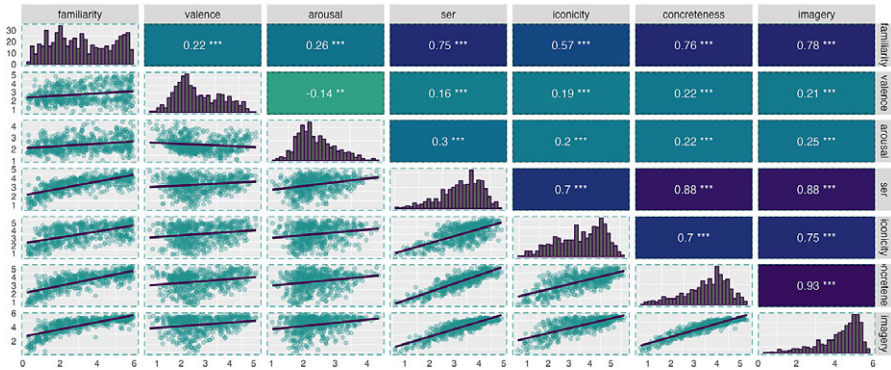

Figure 2. Pairwise correlation plot of all ratings. The lower end shows scatterplots and a fitted linear model. The diagonal shows histograms, which display the distribution. The upper end shows the result of Pearson correlation tests, with asterisks indicating the level of significance.

not equally strong. For example, concreteness and imageability (r = .93, p < 0.001) were highly positively correlated. This is in line with most studies (Paivio et al., 1968; Yao et al., 2017; Yee, 2017), but contrary to what was found for Mandarin compounds (Song & Li, 2021). Because ABB words occupy a place between compound or phrase (Van Hoey, 2023), we did not have a specific expectation, but a strong correlation intuitively made sense to us and was borne out in the ratings. Relatedly, SER also strongly correlated with concreteness (r = .88, p < 0.001) and imagery (r = .88, p < 0.001). This also holds up; most evocative items likely tend to be concrete.

The only ratings that deviate from these positive and significant correlations in Figure 2 are those that involve arousal and valence, which is also apparent if one looks at the visualizations of mean and standard deviation (Figure 1). While significant, the effect sizes of the Pearson correlations between arousal and valence versus the other ratings were quite minor, ranging from −.14 (arousal ~ valence) to .3 (SER ~ arousal). ABB words tend to have a relatively low valence, which is not entirely surprising since it has been observed before that ABB words tend to have bad connotations (T'sou, 1978). More surprising is that the majority of these words were judged as not very or mildly arousing. Perhaps this is because ABB words have a qualificative nature, providing commentary on phenomena in the world without necessarily evoking strong reactions regarding those phenomena.

2.5. Interim discussion

We successfully obtained subjective ratings for 561 ABB words for the following variables: familiarity, valence, arousal, SER, iconicity, concreteness and imagery. Except for valence and arousal, these are all highly and positively correlated with each other. These ratings also make clear that there is a large non-equality of values among the items. That is, we cannot just assume that every ABB item is equally iconic, abstract, or likely to induce imagery in speakers' heads. These variables are all scalar and suggest that even though ABB words can be characterized as imageable, not every item is so to the same degree. One takeaway from the ratings is that the bulk of ABB items are judged as very imageable, quite concrete, iconic and sensory-evoking, relatively familiar but not very arousing or positive. But those judgments are perhaps

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

best taken not by themselves, but in comparison to other, non-iconic, words. We explore this issue in Section 3.

3. Comparison between ABB and prosaic words

In the preceding section, we presented the results of subjective norms for familiarity, valence, arousal, SER, iconicity, concreteness and imageability for ABB items. Given that ABB items are characterized as vivid and are situated along a cline between fully iconic words (such as ideophones) and more prosaic items, it would make sense to compare them with either ideophones or 'normal' prosaic words in terms of the same set of ratings. Unfortunately, such ratings for Chinese ideophones (Van Hoey & Thompson, 2020) are not yet available. We can, however, compare ABB items with prosaic words (Chen et al., 2019; Song & Li, 2021; Xu & Li, 2020; Xu et al., 2021; Yao et al., 2017; Yee, 2017).

To identify which rating types can help characterize ABB items, we compare the data we obtained with two different datasets. In Section 3.1, we will compare our ratings to those of Yao et al. (2017), who collected ratings for valence, arousal, concreteness, familiarity and imageability of 1,100 Chinese two-character words. In Section 3.2, we compare our ratings to the concreteness, familiarity, imageability and SER ratings of 3,783 Chinese two-character words (Song & Li, 2021). Of course, only the ratings that were in common will enter into the analysis. We first conducted an exploratory analysis that involved PCA – the technique we will present in detail in Section 5.1, where we compare the decontextualized ratings with corpus-based measures. We present the results of the PCA in full in the supplementary materials on the OSF repository ('2_boundary' documents) and report here the results of two binary logistic regression models. The exploratory analysis indicated that ABB items would be characterized by at least low arousal and high imagery. The logistic regression models will show that there are also other distinctive significant variables between the two sets of data.

3.1. ABB versus prosaic words (Yao et al., 2017)

Yao et al. (2017) collected the following ratings for 1,100 two-character Chinese prosaic words: valence, arousal, concreteness, familiarity, imageability and context availability. We only consider the ratings we have in common, that is, valence, arousal, concreteness, familiarity and imageability. Yao et al. adopted a 9-point scale, while we adopted a 7-point scale. This means that to make the data comparable, we first normalized the data so that both sets would fall in the interval [0,1]. Then, following Neumann and Evert (2021), we further standardized the data to z-scores with zero mean $(\mu = 0)$ and unit variance $(\sigma^2 = 1)$, and then sign transformed it. Contrary to $\log_{10}$ transformation, signed log[1] can also handle negative values, which are the result of standardizing.

In the next step, we conducted a binary logistic regression, with word type (ABB or prosaic) as its dependent variable and valence, arousal, concreteness, familiarity and imagery as the predictors. We did not fit any random effects because the datasets contained only single observations per item and per variable. The model with the best

[1]The formula for the signed logarithmic transformation is $f(x) = sgn(x) \cdot \log(|x| + 1)$.

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

Table 4. ABB words versus prosaic words (set of Yao et al., 2017)

Term	*β*	*SE*	*Z*	*p*
Intercept (− indicates prosaic, + indicates ABB)	−3.31	0.32	−10.41	<0.001
Valence	−0.15	0.46	−0.33	0.742
Arousal	−3.76	0.52	−7.23	<0.001
Concreteness	−2.66	0.66	−4.03	<0.001
Familiarity	−2.16	0.51	−4.23	<0.001
Imagery	5.67	0.65	8.7	<0.001
Valence:concreteness	−2.73	0.96	−2.83	0.005
Concreteness:imagery	−2.17	0.93	−2.34	0.019
Familiarity:imagery	6.19	0.98	6.35	<0.001
Arousal:concreteness:imagery	−3.66	1.38	−2.66	0.008
Valence:concreteness:familiarity:imagery	3.82	1.33	2.88	0.004

Note: Only significant interactions are shown here; the full model is presented in the supplementary materials.

metrics (*Tjur's* R^2 = .73, *AIC* 757.1, *VIF* < 5.90, *AUC* = .97) contained interactions between all predictors. Table 4 shows the results of this model. All other variables being equal, ABB words are significantly characterized by high imagery ($\beta = 5.67$), especially when there are interactions between imagery and familiarity ($\beta = 6.19$) or valence, concreteness, familiarity and imagery ($\beta = 3.82$). On the other hand, they are comparatively low in arousal ($\beta = -3.76$), concreteness ($\beta = -2.66$), familiarity ($\beta = -2.16$) and variable coefficients that all point toward prosaic words. Given the model's high concordance index C (*AUC* = .97), it discriminates excellently; therefore, we can take these coefficients as a first set of typicality features of ABB items.

3.2. ABB versus prosaic words (Song & Li, 2021)

In the second comparison, we compare our ABB words with the prosaic word ratings studied by Song and Li (2021). They report a wide range of psycholinguistic ratings for an even larger set of Chinese two-character words (n = 3,783): subtitle frequency, number of strokes, number of meanings, familiarity, concreteness, imageability, age of acquisition, subjective frequency, subjective number of meanings, compositionality, emotional experience rating, SER and semantic transparency. The variables that our dataset has in common are familiarity, concreteness, imageability and SER. Like our study, Song and Li also adopted a 7-point scale. Nevertheless, we also normalized, standardized and sign transformed the data before conducting binary logistic regression.

Again, we performed a binary logistic regression, with word type (ABB or prosaic) as its dependent variable and concreteness, familiarity, imagery and SER as the predictors. Like before, we did not fit any random effects, but retained the model with interactions between all predictors, as it had the best metrics (*Tjur's* R^2 = .61, *AIC* 1356.0, *VIF* < 1.88, *AUC* = .96). Table 5 shows the coefficients of the model. ABB words are significantly discriminated by imagery ($\beta = 3.9$), an effect that is boosted in combination with higher unit values for familiarity and imagery ($\beta = 2.57$), imagery and SER ($\beta = 2.36$) and familiarity, imagery and SER ($\beta = 2.19$), despite higher familiarity values ($\beta = -4.78$) or concreteness values ($\beta = -1.19$) pointing more toward prosaic words.

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

Table 5. Prosaic words (set of Song & Li, 2021) versus ABB words

Term	*β*	*SE*	*Z*	*p*
(Intercept)	−5.18	0.3	−17.43	<0.001
(− indicates prosaic, + indicates ABB)				
Concreteness	−1.19	0.46	−2.58	0.010
Familiarity	−4.78	0.37	−12.95	<0.001
Imagery	3.9	0.37	10.41	<0.001
SER	0.61	0.46	1.33	0.183
Familiarity:imagery	2.57	0.5	5.13	<0.001
Imagery:SER	2.36	0.58	4.08	<0.001
Familiarity:imagery:SER	2.19	0.75	2.93	0.003

Note: Only significant interactions are shown here; the full model is presented in the supplementary materials.

3.3. Interim discussion

The comparison of ABB items with the two sets of prosaic words (Song & Li, 2021; respectively, Yao et al., 2017) indicates that the formal requirement of the trisyllabic structure of ABB versus disyllabic structure of the prosaic words aside, there are a number of subjective ratings in which they can be discriminated. The logistic models, both with high C values (respectively, C_{Yao} = .97 and C_{Song} = 0.96) strongly suggest that we can characterize ABB words as predominantly high in imageability or imagery. That is, we may have operationalized what it means to be vivid – the feature that the traditional literature usually attributes to ABB words. At the same time, it must be noted that ABB words in both comparisons were judged as significantly less familiar. However, this may be an effect caused by the size difference of the datasets (N_{ABB} = 561, N_{Yao} = 1100, N_{Song} = 3783). We know, for instance, that the ABB items span the entire range of familiar items. If we had only taken a subset of relatively familiar items, the results may have proven differently. Still, that spread across the range of items will prove useful to understand ABB items as a whole (see Section 5.2).

Interestingly, ABB items were significantly less concrete or arousing than prosaic words. The difference in concreteness may have to do with the qualificative nature of ABB items. When compared with very concrete items like *bēizi* 杯子 'cup', *gāngbǐ* 钢笔 'fountain pen' or *luòtuo* 骆驼 'camel', which all have the highest score in the Song and Li (2021), even very vivid ABB words may not stand much of a chance. We have already noted the relatively low values for arousal (Table 3), and this is borne out in the comparisons here. These low arousal values are potentially caused by the decontextualized nature of the task. After all, it is difficult to allow yourself the chance to be aroused or excited when the task is set up to judge items in rapid succession, and the mind may need some time for the items that are high in imagery to take their arousing effect. In contrast, it may not be surprising that the following items in Yao et al. (2017), for example, *cùsǐ* 猝死 'sudden death', *chēliè* 车裂 'dismemberment by chariots pulling in different directions' or *sānglǐ* 丧礼 'funeral' have very high values for arousal. Lastly, we note that by itself, valence and SERs were not significantly associated with ABB or prosaic words. However, interacting with the other values, they may boost those effects: ABB items are typically characterized then by lower valence items and higher SER items.

4. Corpus data

In Section 3, we identified in what ways ABB items differ from prosaic items when looking at decontextualized ratings. A follow-up question we take here is, can we

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

identify similar features when empirically investigating data from a different nature, that is, corpus data? In Section 4, we introduce the corpus data and relevant measures taken from them. In Section 5, we juxtapose rating and corpus data, and reduce their dimensionality, so that we can observe which variables contribute most to the variance within the data.

It is useful to discern three types of corpora. First, the traditional corpus consists of informative and imaginative prose, predominantly containing written data. This can be contrasted to a second type of corpus data, namely social media data. Social media occupies a space between spoken and traditional written data, a kind of 'digital orality' (Cutler et al., 2022). Finally, we also include a form of institutionalized data, which comprises lists and dictionary material. After all, dictionaries are often emblematic of consensus views regarding relevant words. Below, we provide details on these three types of data (traditional, social, and institutional), the corpus-based measures that can be derived from them and what they reveal about ABB words.

4.1. Corpus information

Traditional corpus material comes from the Academia Sinica Balanced Corpus of Chinese (ASBC) (CKIP Group & Academia Sinica, 2013) and the Chinese National Corpus (CNC) (National Language Committee of China, 2012). The ASBC contains about 17 million characters, and the CNC about 95 million. These corpora are accessible through an online interface and can be acquired in full through institutional license access. Van Hoey's (2023) study on ideophone–collocate constructions was based solely on the ASBC. Because the current study only focuses on the ideophone–collocate prototype of ABB words, we decided to augment the data with CNC data, which is about four times bigger than the ASBC. The 561 ABB items included in the analysis of the ratings (Section 2.1) were retrieved from the corpora through regular expressions in R. They are provided in the folder 'corpus_data' in the OSF repository, and the analysis in the '3_corpus' document. Note that the ASBC contains data in traditional characters, but we are presenting the simplified counterparts throughout this study, as the bulk of the data comes from simplified sources.

Social media data come from the Leiden Weibo Corpus (LWC; van Esch, 2012). Sina Weibo (*Xīnlàng Wēibó* 新浪微博) is a Chinese microblogging website on a par with Twitter. The LWC contains messages that were collected during January 2012, resulting in 5 million messages. Because this dataset is huge yet uncurated, we find the largest amount of ABB tokens here (see Table 6).

Institutionalized data are a cover term for lists and dictionary data, the typical material investigated in most studies discussing ABB words (Cáo, 1995; Lǐ, 2007, 2008; Wáng, 2020, 2014; Zhào, 2021; Zhāng, 2005). While these are not corpus material *per se*, they provide an insight into which items are deemed most stable from a consensus point of view. The institutionalized data consulted include the *800 words of Modern Chinese* (Lǚ, 1980), the *Comprehensive Dictionary of Chinese* (Luó, 1993), the *Dictionary of adjective usage in Chinese* (Zhèng & Mèng, 2003), the *Standard dictionary of Modern Chinese* (Lǐ, 2013) and the *Contemporary Chinese Dictionary* (Chinese Academy of Social Sciences, 2016). Respectively, these are abbreviated as

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

Table 6. Summary of corpus material

Data type	Dataset	Number of ABB tokens
Social	LWC	22,815
Traditional	ASBC	1,617
Traditional	CNC	7,484
Institutional	HYDCD	127
Institutional	BBC	281
Institutional	Xianhan	175
Institutional	Xingrong	21
Institutional	Guifan	182

Table 7. ABB words occurring across all types of corpus data, with their raw token frequencies

ABB	Pinyin	Meaning	Social	Traditional	Institutional
乱哄哄	*luàn–hōnghōng*	'noisy'	51	65	5
冷冰冰	*lěng–bīngbīng*	'cold'	303	91	5
水汪汪	*shuǐ–wāngwāng*	'watery'	55	41	5
白茫茫	*bái–mángmáng*	'white'	789	100	5
胖乎乎	*pàng–hūhū*	'chubby'	48	31	5
赤裸裸	*chì–luǒluǒ*	'naked'	800	161	5
闹哄哄	*nào–hōnghōng*	'noisy'	58	27	5
黑洞洞	*hēi–dòngdòng*	'black'	19	66	5

BBC, HYDCD, Xingrong, Guifan and Xianhan based on their Chinese titles and presented as such in Table 6.

4.2. Measures

The bird's-eye perspective of the corpora measures associations that are observable on a general level, and are comparable with the ratings presented in Section 2 as well as with the findings presented in Van Hoey (2023). We take into account (1) the token frequencies of different ABB types, (2) their dispersion across the three general corpus data types and their cue validity (3) from A to BB and (4) from BB to A.

Token frequencies involve counting how many times a particular ABB word appears in the three corpus data types. Examples are shown in Table 7. Intuitively, words with a higher frequency will be more familiar. In Section 5, we show to what degree token frequency and familiarity contribute to the same latent dimension of the variance.

Dispersion (Gries, 2008, 2020, 2021) measures how equally widespread a given ABB word is across all three corpus data types, and can be viewed as a sort of correction on raw frequencies. For instance, do some words only appear in the institutionalized dictionary data or are they used relatively equally in social media data and traditional corpus material as well? We operationalize dispersion as the Kullback–Leibler divergence (see Gries, 2021; Van Hoey, 2023). For example, hypothetically speaking, *xiōng-bābā* 凶巴巴 is highly familiar in our ratings (Table 1). This high rating may be related to a high token frequency in social media data while it is absent from traditional data and institutionalized data. In that case, its dispersion will be low, because it is skewed toward one type of data only, albeit very frequent in

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

that type of data. Realistically speaking, *xiōng-bābā* occurs 32 times in social data, 9 times in traditional data and 2 times in institutionalized data. Given the size of the corpus types, this results in a normalized dispersion value of 0.97, that is, *xiōng-bābā* is well-dispersed across the three kinds of corpus data and occurs as one would expect. In this case, we can expect that its high mean familiarity rating correlates with both token frequency and dispersion. Of course, given that we already know that familiarity highly correlates with SER, iconicity, concreteness and imageability, we could expect a high correlation with token frequency and/or dispersion as well.

Two other measures that we examined involve the morphological construction of ABB words. As is known in the literature, some A's attract many different BB's, while others only occur with one BB (see Section 1). Likewise, some BB parts co-occur with many A parts. We can calculate the so-called cue validity (Stefanowitsch & Gries, 2003) between A and BB through unidirectional association measures. In this case, we also adopt the Kullback–Leibler divergence (D_{KL}) or relative entropy (Baayen, 2011; Gries & Durrant, 2020). The rationale behind using this measure is as follows: the collected iconicity ratings may tell us if the compositionality of a given ABB word is clear, but it is unclear whether the salience of A or BB plays a role in this. Keeping with the example of *xiōng-bābā*, we find that its A *xiōng* is a very strong cue for the BB *bābā* ($D_{KL:A \rightarrow BB} = .99$). This means that when people see *xiōng* and are asked to come up with an ABB form, they are expected to reply *xiōng-bābā*. However, it is well-known that *bābā* is a very common component of ABB words (Cáo, 1995), so its cue validity for *xiōng* is much lower ($D_{KL:BB \rightarrow A} = .44$). The prediction then is that iconicity ratings are expected to correlate with either one of these Kullback–Leibler divergence measures in terms of their contributions to the underlying dimensions in Section 5.

4.3. Results

In Figure 3, the token frequencies in the traditional and social corpora follow a Zipfian distribution (diagonally), with an expected large number of hapaxes. Overall, though, ABB words have a relatively low frequency in corpus material, as has been

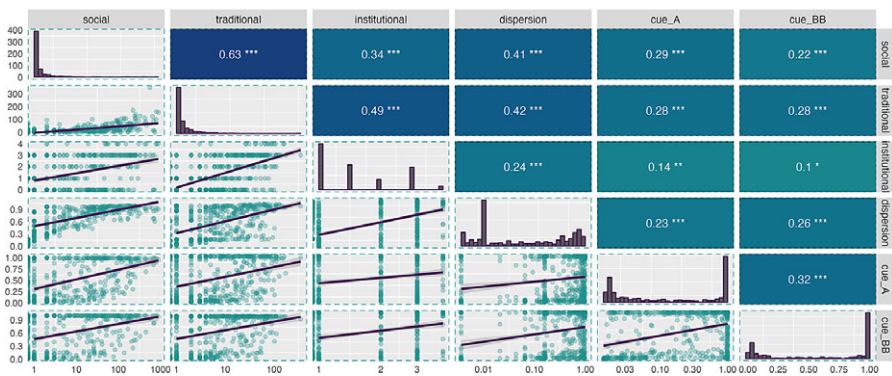

Figure 3. Pairwise correlation plot of all corpus-based measures. The lower triangle shows scatterplots and a fitted linear model. The diagonal shows histograms, which display the distribution. The upper triangle shows the result of Pearson correlation tests, with asterisks indicating the level of significance.

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

shown for other iconic words, see Van Hoey (2020) for frequencies of Chinese ideophones, Van Hoey (2023) for frequencies of ABB words and the work of Perry et al. (2015) or Winter et al. (2023) for studies on English data. Institutionalized data further displays a similar distribution: many items only occur in one data source. Furthermore, only a few items (N = 8) appear across all of them, as shown in Table 7. The dispersion value of most ABB items approaches 0, that is, distribution across the three corpus types is skewed heavily. The cue validity measures both show that many ABB items have extreme values near the 1 mark, but not all of them.

The correlations between the measures are all significant yet of differing effect sizes. It is no surprise that the token frequencies of traditional and social corpus types are highly correlated ($r = .63$, $p < 0.001$) and that institutional data are somewhat correlated with traditional data ($r = .49$, $p < 0.001$) and social data ($r = .34$, $p < 0.001$). We take this to mean that actual corpus usage is relatively comparable and its relation to occurrence across different lists is as expected as well. Perhaps somewhat surprising is the positive and significant correlation between dispersion and corpus data type ($r = .41$, $p < 0.001$ for social; $r = .42$, $p < 0.001$ for traditional; $r = .24$, $p < 0.001$ for institutional). This indicates that ABB words occurring with a higher token frequency in one corpus data type will also be somewhat equally frequent[2] across the other data types. Finally, the two compositional measures of cue validity are significantly correlated with the other measures, but their effect sizes are quite weak, ranging from .1 to .32.

4.4. Interim discussion

Two notions can be summarized from the corpus study. First, we find that ABB items follow typical word effects in terms of their frequency and distribution patterns. Their differences regarding membership salience and typicality shines even more through the subjective ratings in Section 2. Second, given the results we find, we might need to revise the predictions concerning correlations between rating data and corpus data. The prediction for a high correlation between token frequency and/or dispersion and familiarity ratings still stands, as well as the correlated SER, iconicity, concreteness and imageability (Section 2). The prediction for a high correlation between cue validity measures and iconicity ratings is a lot less sure. The dimensionality reduction in Section 5 explores to what degree the ratings and corpus measures tap into the same underlying structure to explain the greatest amount of variance within the data.

5. Bridging corpus and norm

Thus far, we have seen that ABB item exemplars vary widely with regard to the average values of the subjective ratings and corpus-based measures. We have also argued that there are good reasons for predicting that there are correlations not only within ratings or corpus measures, but also across these two ontologically different data sources. We can rephrase the question as follows: which variables highlight the

[2]We wish to point out that this is not necessarily true on a more fine-grained level. In a preliminary phase of this study, we only investigated the ASBC corpus and looked at dispersion across genres there and actually found a significant and strongly negative correlation ($r = -.76$). We are currently exploring more fine-grained analyses such as register analysis (Biber & Finegan, 1994) to investigate this.

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

same underlying mechanisms responsible for the structural variance within the set of ABB words? To answer this question, we adopt a technique that involves dimensionality reduction, namely PCA. After all, we have information regarding 561 ABB words in 13 dimensions. These dimensions include familiarity, valence, arousal, imageability, concreteness, sensory experience, iconicity; token frequency in traditional, social and institutional data types; dispersion and cue validity from A to BB and from BB to A. By reducing the 13 dimensions to a few latent dimensions that contain the most explanatory variables, we can observe whether a bridge between corpus and rating data are possible to characterize iconic phrases like ABB words. The analysis can be found in '4_pca' on the OSF repository.

5.1. Principal component analysis

The data (561 rows × 13 columns, that is, the ratings and corpus measures) were standardized to z-scores with zero mean ($\mu = 0$) and unit variance ($\sigma^2 = 1$). Like in Sections 3.1 and 3.2, we used a signed log transformation to handle the negative values that appeared after standardization (see Neumann & Evert, 2021). The scree plot (Figure 4) shows that Dimension 1 (eigenvalue = 5.58) captures 42.9%, Dimension 2 (eigenvalue = 1.76) 13.5%, Dimension 3 (eigenvalue 1.17) 9% and so on. The elbow criterion indicates that Dimensions 1 and 2 (an acceptable 56.4% of the variance) suffice for the analysis.

Next, we look at what drives Dimensions 1 and 2. For Dimension 1, the variables that contribute more than expected include familiarity, imageability, concreteness, SER and iconicity of the ratings – which we already know are highly correlated (see Section 2.4), but also the token frequencies of the traditional and social (barely) corpus data types and also the dispersion values. For Dimension 2, both cue validity values, the token frequencies in the social and traditional corpora, and imagery (barely) from the ratings contribute more than the baseline. If we had presented Dimension 3 here, we would have found it to be completely driven by valence and arousal. However, since the scree plot (Figure 4) indicates that two dimensions suffice to capture most of the variance, we leave this plot to the supplementary materials.

The PCA is successful in reducing the complexity of the 13 dimensions. As Figure 5 shows, Dimension 1 is driven mostly by familiarity, imagery, concreteness,

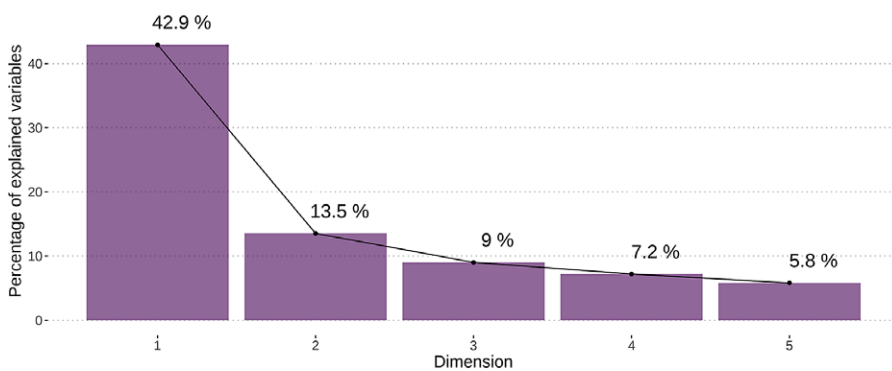

Figure 4. Scree plot of the principal components analysis based on the 13 transformed variables (ratings and corpus).

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

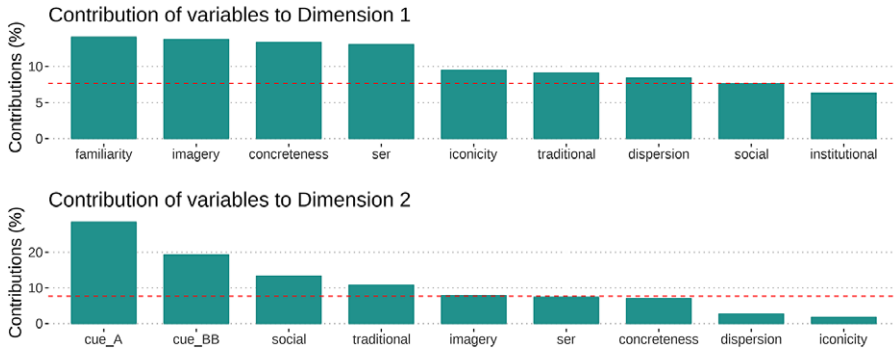

Figure 5. Contributions of variables to the first and second PCA dimensions. The red line indicates the expected contribution if the variables' contributions were equal.

SER, iconicity, traditional, dispersion and (borderline) social. Dimension 2 is driven by cue validity measures, social, traditional and (borderline) imagery. This is exciting because it shows that the rated data and the corpus data tap into the same latent variables, that is, they capture the same structure within the group of ABB words. Furthermore, we have a more fine-grained idea of which measures weigh stronger (Figure 5). This kind of prototypicality effect is relevant for identifying the features one needs to consider when one wants to define or characterize ABB words.

5.2. Exemplars in the PCA

The PCA not only provides insight into the convergence of rating and corpus data on the feature level, but also shows how the ABB items relate to each other. Figure 6 displays all ABB items based on their respective coordinates in Dimensions 1 and 2. For reference, we have highlighted the items discussed in Tables 1 and 7. As expected, these cover the entire range of Dimension 1 (the x-axis) since this dimension is driven by familiarity ratings, among other things. Additionally, Dimension 2 also shows a good range of the spread, confirming again that both dimensions indeed are informative. The centroid of this plot (purple square partly covered by *làng-gǔngǔn* 浪滾滾) lies at (0,0), which is a result of the scaling we performed. We have additionally added density lines to indicate where concentrations of similar points are situated.

To ease the discussion of the visualization in Figure 6, we have marked three heuristic clusters (cluster 1 in yellow; cluster 2 in blue cluster 3 in green). These clusters serve a didactic function, as the ABB items are spread in a continuous manner across the conceptual space resulting from the PCA.

The first cluster in (Figure 6 yellow rectangle) shows that most ABB are concentrated near 0 on the first dimension. This means that they have a mean familiarity, imageability, concreteness, SER and iconicity as well as token frequency and dispersion to a certain degree. They are negative on the second dimension, which means lower values for cue validity and token frequency in the social and traditional corpus data types. In sum, these items do not occur very often in the corpus and are not overly familiar, imageable and so forth. Yet, they are typical in the sense that most ABB items behave like them; they are average, and in that way are good

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

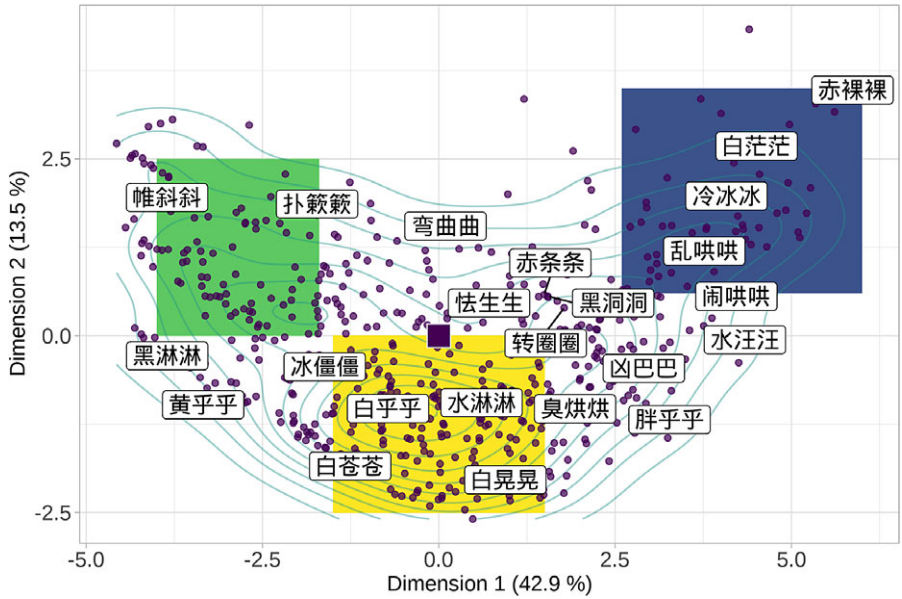

Figure 6. ABB items plotted on the PCA space for Dimension 1 and Dimension 2. A density plot (green lines) shows the concentration of items in different bands. The colored rectangles indicate three clusters of interest. Highlighted items are those from Table 1 and Table 7. Versions with transcriptions and English translations are provided in the OSF repository.

representatives of ABB words. Examples of this cluster include *shuǐ-línlín* 水淋淋, *bái-hūhū* 白乎乎 and *bái-huǎnghuǎng*白晃晃.

The second cluster (Figure 6 blue rectangle) contains items that score high in both dimensions. These items are very familiar and occur often in the corpus. It is not unthinkable that a prototypicality experiment in the tradition of Rosch (Mervis & Rosch, 1981) would yield items like *bái-mángmáng* 白茫茫, *lěng-bīngbīng* 冷冰冰, *luàn-hōnghōng* 乱哄哄 and *chì-luǒluǒ* 赤裸裸. That is, such ABB items contain prototypes in the psycholinguistic sense, with the typical benefits of faster production, faster recognition, less processing time for their semantics and so forth (Geeraerts, 1989 for seminal overviews; see Medin & Smith, 1984).

The third cluster (Figure 6 green rectangle) concerns items that score negative on Dimension 1 but positive on Dimension 2. These are items that occur in the corpus and hence have corpus statistics, but which may be novel or foils detected by the retrieval method. If the prototypicality experiment conjectured above would yield items from the blue rectangle, it would certainly not yield those of the green one, for example, *wèi-xiéxié* 帷斜斜 and *hēi-línlín* 黑淋淋. We informally asked native speakers about these words, and they were either completely incomprehensible ('mistakes') or compositionally interpretable, but highly unfamiliar. In other words, these items are situated on the boundary of the ABB category.

Finally, we note that there are not many ABB items with mean scores of the first dimension and very positive second dimensions, although they occur, for example, *chì-tiáotiáo* 赤条条 or *wān-qūqū* 弯曲曲.

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

5.3. Interim discussion

While we were unsure whether findings from the ratings and the corpus would yield complementary insights in the ABB lexicon or would overlap, we are now confident that the two types of data can tap into similar latent variables yet are also distinct. The 'major players' that drive the structure of the ABB lexicon are the ratings of familiarity, imageability, concreteness, SER, iconicity and the corpus-based measures of token frequency in the social and traditional corpora, cue validity measures and dispersion. Valence and arousal, on the other hand, can help structure the data in a third dimension, but this dimension turns out to be not very relevant here. Whether items occur across many dictionaries (institutionalized data) turns out to not play a significant role. The picture that emerges then is that pleas to study data in observed contexts (corpus data) and with the help of many heads (rating data) remain important.

6. Discussion and conclusion

This study began with the methodological observation that decontextualized ratings are currently (and successfully so) part and parcel of the field of linguistic iconicity. At the same time, we wanted to know whether it would be possible to attain a more well-rounded picture of iconic words if we supplemented ratings with corpus-based information. As the object of the study, we chose Mandarin Chinese ABB words, for example, *wù-mángmáng* 雾茫茫 'completely foggy', since these words are situated along the cline of completely iconic words such as onomatopoeias and non-iconic (prosaic) words (Dingemanse, 2017). Formally, ABB items are composed of a prosaic A syllable and a reduplicated BB syllable. This has resulted in characterizations of ABB, ranging from a careful 'vivid' in the literature (Huang et al., 2016; Wang, 2014) to stronger statements that BB's are ideophones or at least ideophonized (Van Hoey, 2023), that is, iconic.

Methodologically, we presented the first-ever set of subjective ratings for Mandarin ABB words. We collected norms about their familiarity, valence, arousal, imageability, concreteness, sensory experience and iconicity. In Section 2, we showed that the majority of ABB items had relatively high scores for imageability, familiarity, sensory experience, iconicity and concreteness, while they scored rather low for arousal and valence. Moreover, we predicted and corroborated that especially the high scoring values would correlate.

The findings of rating information by themselves already informed our understanding about the differences in typicality of ABB items within the set, but we further compared our data with two different sets of prosaic words (Section 3). We conducted two binary logistic regressions and found that the most salient feature that set apart ABB words from prosaic words is their high values for imagery. As we pointed out above, we may have thus operationalized the vague notion of 'vividness' that is found scattered throughout the literature in an empirical manner. Surprisingly, while ABB items by themselves scored high for familiarity and concreteness, high scores for these values were more associated with prosaic words. Less surprising was that arousal and valence were inversely correlated with ABB words. A first takeaway from the comparison is that, formal differences aside (trisyllabic ABB vs. disyllabic prosaic words), there is some overlap between the two sets of words. However, the variables contain enough distinctions to differentiate between them. The second

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

methodological takeaway is that values that structure a category internally may turn out not as relevant when stacked up against comparable sets of items.

However, this does not mean that category-internal structure should be thrown out. Instead, we argue to augment it with other kinds of data, which now brings us to the question that started this study. In Section 4, we presented corpus data and the measures derived from it. These measures (token frequency for three different of corpora, dispersion across these corpora and cue validities for A-BB internal structure based on the corpus data) provided a bird's-eye overview of the contextualized structure of ABB usage. We found there were correlations between the measures and that the distribution of data largely followed known (Zipfian) patterns. In Section 5, we then combined the subjective ratings and the corpus-based measures, in order to conduct a PCA. This dimensionality reduction technique showed that two latent dimensions explained a sizeable amount of variance within the data. More importantly, those two dimensions showed that the two empirically and ontologically different data sources tap into the same underlying structuring of the data. The first dimension was mostly driven by the values for familiarity, imagery, concreteness, SER, iconicity, but also the token frequencies of the three corpora; the second dimension mostly by the internally composition measures of cue validity (in both directions), but also token frequencies and imageability.

Additionally, when we observed the ABB loadings of the PCA (Section 5.2), we saw there were three areas of interest. We identified which ABB words were more central to the category, with high values for all variables participating in the two dimensions (high ratings, high corpus frequencies, etc.). We also found which ABB words shared average values. Unsurprisingly, these items were the largest in number. And we found items that were novel lexicalizations or false positives; in any case, they were rather unfamiliar and had a low occurrence in the data.

The different angles we considered in this study point toward a simple take-home message for the field of iconicity: take word effects into account when presenting or using decontextualized ratings. This can be done by placing a new set against a comparable set of words, showing how absolute findings for one set may turn out to be relatively less important or associated with that one set. It can also be done by turning to different kinds of data, highlighting which findings are common across the two kinds of data. We should be wary of inflated claims of iconicity that disregard what may be going on at other linguistic levels, such as the Pokémonastics paradigm, which tends to overlook the morphological level. It also makes conclusions based on the assumption that all Pokémon names display iconic qualities in equal manners, without considering familiarity or frequency effects at play (Kawahara et al., 2018). Fortunately, we also observe critical warnings about iconicity effects driven by associating iconic features with known words rather than grounded in perceived structural form-meaning mappings (McLean et al., 2023; Thompson et al., 2020, but see Winter & Perlman, 2021), as well as the body of work to which we wish to contribute: subjective ratings that are informed by other ratings (Dingemanse & Thompson, 2020; Winter et al., 2023), that are contrasted with other kinds of evidence such as the corpus (Klavan & Divjak, 2016), or that are compared to different sets (Perlman et al., 2018). In conclusion, we are hopeful that vivid words will continue to receive more multivariate treatments in the future.

Data availability statement. The data and code with the analyses can be found on the following OSF repository: https://osf.io/tv34b/ (doi: 10.17605/OSF.IO/TV34B).

Acknowledgments. The authors would like to thank the engaging questions raised at the 16th International Cognitive Linguistics Conference ('Bridging corpus and norm: Mandarin sensory adjectival phrases') and the online Iconicity Seminar. The authors appreciate the suggestions and feedback from the anonymous reviewers and Bodo Winter whose detailed comments improved the structure of the paper.

Funding statement. This work was supported by General Research Fund 17603120 awarded by the Research Grant Council of Hong Kong.

Competing interest. The authors declare none.

References

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *SciELO Brasil*, 11(2), 295–328.

Baayen, R. H., & Divjak, D. (2017). Ordinal GAMMs: A new window on human ratings. In L. A. Janda, A. Makarova, S. M. Dickey, & D. Divjak (Eds.), *Each venture a new beginning: Studies in honor of Laura A. Janda* (pp. 39–69). Slavica.

Biber, D., & Finegan, E. (Eds.) (1994). *Sociolinguistic perspectives on register*. Oxford University Press.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. In *Technical Report C-1*. Center for Research in Psychophysiology, University of Florida.

Brown, G. D., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity. *Memory & Cognition*, 15(3), 208–216.

Bruekl, M., & Heuer, F. (2022). *irrNA: Coefficients of interrater reliability – Generalized for randomly incomplete datasets (0.2.3)*. Computer Software.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cáo, R. (1995). Pǔtōnghuà ABB shì xíngróngcí de dìngliàng fēnxī [Quantitative analysis of Mandarin Chinese ABB adjectives]. *Yǔwén Yánjiū*, 3, 22–25.

Chang, Y. (2009). The phonology of ABB reduplication in Taiwanese. In Y. Xiao (Ed.), *Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21)* (Vol. 1, pp. 28–41). Bryant University.

Chen, I.-H., Zhao, Q., Long, Y., Lu, Q., & Huang, C.-R. (2019). Mandarin Chinese modality exclusivity norms. *PLOS ONE*, 14(2), e0211336. https://doi.org/10.1371/journal.pone.0211336

Chinese Academy of Social Sciences. (2016). *Xiàndài Hànyǔ Cídiǎn (Contemporary Chinese Dictionary)* (7th ed.). Commercial Press.

CKIP group & Academia Sinica. (2013). *Academia Sinica Balanced Corpus of Modern Chinese (ASBC 4.0)* [dataset]. https://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm

Cutler, C. A., Ahmar, M., & Bahri, S. (Eds.) (2022). *Digital orality: Vernacular writing in online spaces*. Palgrave Macmillan.

Della Rosa, P. A., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, 42(4), 1042–1048. https://doi.org/10.3758/BRM.42.4.1042

Dingemanse, M. (2011). The meaning and use of ideophones in Siwu *[Dissertation]*. Radboud University Nijmegen.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10), 654–672.

Dingemanse, M. (2017). Expressiveness and system integration: On the typology of ideophones, with special reference to Siwu. *STUF - Language Typology and Universals*, 70(2), 363–384. https://doi.org/10.1515/stuf-2017-0018

Dingemanse, M. (2019). Ideophone" as a comparative concept. In K. Akita & P. Pardeshi (Eds.), *Ideophones, mimetics and expressives* (pp. 13–33). John Benjamins. https://doi.org/10.1075/ill.16.02din

Dingemanse, M., Perlman, M., & Perniss, P. (2020). Construals of iconicity: Experimental approaches to form-meaning resemblances in language. *Language and Cognition*, 12, 1–14. https://doi.org/10.1017/langcog.2019.48

Dingemanse, M., & Thompson, B. (2020). Playful iconicity: Structural markedness underlies the relation between funniness and iconicity. *Language and Cognition*, 12, 1–22. https://doi.org/10.1017/langcog.2019.49

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods*, 50 (3), 1116–1124. https://doi.org/10.3758/s13428-017-0930-6

Geeraerts, D. (1989). Introduction: Prospects and problems of prototype theory. *Ling*, 27(4), 587–612. https://doi.org/10.1515/ling.1989.27.4.587

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. T. (2020). Analyzing dispersion. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 99–118). Springer.

Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2), 1–33. https://doi.org/10.32714/ricl.09.02.02

Gries, S. T., & Durrant, P. (2020). Analyzing co-occurrence data. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 141–159). Springer.

Haiman, J. (Ed.). (1985). Iconicity in syntax. In *Proceedings of a symposium on iconicity in syntax, Stanford, June 24-26, 1983*. Benjamins.

Hinojosa, J. A., Haro, J., Magallares, S., Duñabeitia, J. A., & Ferré, P. (2021). Iconicity ratings for 10,995 Spanish words and their relationship with psycholinguistic variables. *Behavior Research Methods*, 53, 1262–1275. https://doi.org/10.3758/s13428-020-01496-z

Huang, S.-Z., Jin, J., & Shi, D. (2016). Adjectives and adjective phrases. In C.-R. Huang & D. Shi (Eds.), *A reference grammar of Chinese* (pp. 276–296). Cambridge University Press.

Juhasz, B. J., & Yap, M. J. (2013). Sensory experience ratings for over 5,000 mono- and disyllabic words. *Behavior Research Methods*, 45(1), 160–168. https://doi.org/10.3758/s13428-012-0242-9

Kawahara, S., Noto, A., & Kumagai, G. (2018). Sound symbolic patterns in Pokémon names. *Phonetica*, 75(3), 219–244. https://doi.org/10.1159/000484938

Klavan, J., & Divjak, D. (2016). The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica*, 50(2), 355–384. https://doi.org/10.1515/flin-2016-0014

Körtvélyessy, L., & Štekauer, P. (Eds.) (2024). *Onomatopoeia in the world's languages*. Mouton de Gruyter.

Lǐ, J. (2007). *Xiàndài Hànyǔ níshēngcí yánjiǔ (Onomatopoeics in Modern Chinese)*. Xuélín chūbǎnshè.

Lǐ, J. (2008). ABB xíngshì róng cí de gòuchéng fāngshì (The structure type of adjectives in the form ABB). *Journal of Gannan Normal University*, 1, 87–91.

Lǐ, X. (2013). *Xiàndài Hànyǔ guīfàn cídiǎn [Standard dictionary of Modern Chinese]* (3rd ed.). Foreign Language Teaching and Research Press.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lǚ, S. (Ed.). (1980). *Xiàndài Hànyǔ bābǎi cí [800 words of Modern Chinese]* (5th ed.). Commercial Press.

Luó, Z. (Ed.) (1993). *Hànyǔ dà cídiǎn (Comprehensive Dictionary of Chinese)*. Shanghai Lexicographical Publishing House.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

McLean, B., Dunn, M., & Dingemanse, M. (2023). Two measures are better than one: Combining iconicity ratings and guessing experiments for a more nuanced picture of iconicity in the lexicon. *Language and Cognition*, 15, 1–24. https://doi.org/10.1017/langcog.2023.9

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35(1), 113–138. https://doi.org/10.1146/annurev.ps.35.020184.000553

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89–115.

Mok, W. E. (2001). Chinese sound symbolism: A phonological perspective *[PhD dissertation]*. University of Hawai'i.

National Language Committee of China. (2012). *China National Corpus (CNC)* [dataset]. http://corpus.zhonghuayuwen.org/index.aspx

Neumann, S., & Evert, S. (2021). A register variation perspective on varieties of English. In E. Seoane & D. Biber (Eds.), *Corpus-based approaches to register variation* (Vol. 103, pp. 143–178). John Benjamins Publishing Company. https://doi.org/10.1075/scl.103

Noble, C. E. (1953). The meaning-familiarity relationship. *Psychological Review*, 60(2), 89–98.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1, Pt.2), 1–25. https://doi.org/10.1037/h0025327

Perlman, M., Little, H., Thompson, B., & Thompson, R. L. (2018). Iconicity in signed and spoken vocabulary: A comparison between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in Psychology*, 9, 1433. https://doi.org/10.3389/fpsyg.2018.01433

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLOS ONE*, 10(9), e0137147. https://doi.org/10.1371/journal.pone.0137147

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, 50(3), 1198–1216. https://doi.org/10.3758/s13428-017-0938-y

Qualtrics. (2022). *Qualtrics* [Computer software]. https://www.qualtrics.com/

Rofes, A., Zakariás, L., Ceder, K., Lind, M., Johansson, M. B., de Aguiar, V., Bjekić, J., Fyndanis, V., Gavarró, A., Simonsen, H. G., Sacristán, C. H., Kambanaros, M., Kraljević, J. K., Martínez-Ferreiro, S., Mavis, İ., Orellana, C. M., Sör, I., Lukács, Á., Tunçer, M., … Howard, D. (2018). Imageability ratings across languages. *Behavior Research Methods*, 50(3), 1187–1197. https://doi.org/10.3758/s13428-017-0936-0.

Song, D., & Li, D. (2021). Psycholinguistic norms for 3,783 two-character words in Simplified Chinese. *SAGE Open*, 11(4), 215824402110544. https://doi.org/10.1177/21582440211054495

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1), 111–123. https://doi.org/10.3758/s13428-015-0700-2

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

T'sou, B. K. (1978). Sound symbolism and some socio- and historical linguistic implications of linguistic diversity in Sino-Tibetan languages. *Cahiers de linguistique - Asie orientale*, 3(1), 67–76. https://doi.org/10.3406/clao.1978.1039

Thompson, A. L., Akita, K., & Do, Y. (2020). Iconicity ratings across the Japanese lexicon: A comparative study with English. *Linguistics Vanguard*, 6(1), 20190088. https://doi.org/10.1515/lingvan-2019-0088

van Esch, D. (2012). *Leiden Weibo Corpus* [Database]. http://lwc.daanvanesch.nl/index.php

Van Hoey, T. (2020). Prototypicality and salience in Chinese ideophones: A cognitive and corpus linguistics approach *[PhD dissertation]*. National Taiwan University.

Van Hoey, T. (2023). ABB, a salient prototype of collocate–ideophone constructions in Mandarin Chinese. *Cognitive Linguistics*, 34(1), 133–163. https://doi.org/10.1515/cog-2022-0031

Van Hoey, T. (2024). A semantic map for ideophones. In T. F. Li (Ed.), *Handbook of cognitive semantics* (Vol. 2, pp. 129–175). Brill.

Van Hoey, T., & Thompson, A. L. (2020). The Chinese Ideophone Database (CHIDEOD). *Cahiers de Linguistique Asie Orientale*, 49(2), 136–167. https://doi.org/10.1163/19606028-bja10006.

Wáng, T. (2020). *Xiàndài Hànyǔ ABB shì zhuàngtài xíngróngcí rèn zhī yánjiū (A cognition study of ABB-type state adjectives in Mandarin Chinese)* [Master thesis]. China West Normal University.

Wang, Z. (2014). The head of the Chinese adjectives and ABB reduplication. *US-China Foreign Language*, 12 (5), 349–359.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Winter, B. (2019). *Sensory linguistics: Language, perception and metaphor*. John Benjamins Publishing Company. https://doi.org/10.1075/celcr.20

Winter, B. (2022). Managing semantic norms for cognitive linguistics, corpus linguistics and lexicon studies. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 489–497). The MIT Press. https://doi.org/10.7551/mitpress/12200.001.0001

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2023). Iconicity ratings for 14,000+ English words. *Behavior Research Methods*, 56. https://doi.org/10.3758/s13428-023-02112-6

Winter, B., & Perlman, M. (2021). Iconicity ratings really do measure iconicity, and they open a new window onto the nature of language. *Linguistics Vanguard*, 7(1), 20200135. https://doi.org/10.1515/lingvan-2020-0135

https://doi.org/10.1017/langcog.2024.22 Published online by Cambridge University Press

Xu, X., & Li, J. (2020). Concreteness/abstractness ratings for two-character Chinese words in MELD-SCH. *PLOS ONE*, 15(6), e0232133. https://doi.org/10.1371/journal.pone.0232133

Xu, X., Li, J., & Guo, S. (2021). Age of acquisition ratings for 19,716 simplified Chinese words. *Behavior Research Methods*, 53(2), 558–573. https://doi.org/10.3758/s13428-020-01455-8.

Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49(4), 1374–1385. https://doi.org/10.3758/s13428-016-0793-2

Yee, L. T. S. (2017). Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns in Cantonese speakers in Hong Kong. *PLOS ONE*, 12(3), e0174569. https://doi.org/10.1371/journal.pone.0174569

Zhāng, D. (2005). Hànyǔ zhōng ABB xíng zhuàngtài xíngróngcí de gòuchéng fènxī [An analysis of the formation of ABB type state adjectives in Mandarin Chinese]. *Journal of Shenyang Agricultural University (Social Sciences)*, 7(1), 124–126.

Zhào, Q. (2021). Tōnggǎn yīnyù shìjiǎo de Xiàndài Hànyǔ ABB shì zhuàngtài xíngróngcí (ABB-pattern state adjectives in Mandarin: A study from the perspective of synaesthetic metaphor). *Shìjiè Hànyǔ Jiāoxué*, 35 (2), 206–219. https://doi.org/10.13724/j.cnki.ctiw.2021.02.005

Zhèng, H., & Mèng, Q. (2003). *Hànyǔ xíngróngcí yòngfǎ cídiǎn [Dictionary of adjective usage in Chinese]*. Commercial Press.

Cite this article: Van Hoey, T., Yu, X., Pan, T.-L., & Do, Y. (2024). What ratings and corpus data reveal about the vividness of Mandarin ABB words, *Language and Cognition* 16: 1674–1696. https://doi.org/10.1017/langcog.2024.22