ABB, a salient prototype of collocate-ideophone constructions in Mandarin Chinese

Thomas Van Hoey[1]

Abstract

ABB words in Chinese, e.g., *hēi-qīqī* 'pitch black', have been studied for a long time. Most traditional studies analyze these words through derivational rules involving empty suffixes. However, this is problematic, as they are better seen as compounds involving a prosaic A and an ideophonic BB part. By treating ABB as a schema sanctioned by collocate-ideophonic constructions, it is possible to investigate other similar patterns. A corpus study (more than 5000 tokens) revealed that on the level of schemas, ABB truly acts as a prototype of such constructions, but that it is far from the only pattern to be identified. A second corpus-based study on the level of exemplars showed there are different pockets of salience and non-uniformity in the data from four angles: cue validity, frequency, dispersion, and constructional preference. This paper provides evidence that the traditional ABB narrative needs to be complemented with usage-based data, and grapple with the lexical salience effects this brings along for words involving iconicity.

Keywords: ABB, ideophones, iconicity, prototype, tupleization

Declaration of interest: none.

1. Introduction

Chinese linguistics has devoted particular attention to so-called ABB adjectives. A first major reason is the very recognizable templatic structure of ABB adjectives. These adjectives typically contain a noun, adjective or verb in the A slot, followed by a reduplicated element in the BB slot. On the whole, Chinese speakers find that ABB often make the element in the A slot more vivid. A second reason is that this construction has been identified in a number of Sinitic languages, such as, but not exclusive to, Mandarin[2], Cantonese, Taiwanese Southern Min, and Hakka, as illustrated in (1).

[1] Department of Linguistics, KU Leuven, Leuven, Belgium, ORCID: 0000-0002-5226-9752, E-mail: thomas_van_hoey@outlook.com

[2] Glottocodes for languages mentioned in this paper: Mandarin [mand1415], Cantonese [yuec1235], Southern Min [coas1318], Hakka [sanh1239].

(1) a. Mandarin *hēi-qīqī* 黑漆漆[3] ‘pitch dark’
b. Cantonese *hak1-maa4maa4* 黑麻麻 ‘pitch dark’
c. Southern Min *o̍-sôsô* 烏趖趖 ‘pitch dark’
d. Hakka *vú-sǒsǒ* 烏趖趖 ‘charcoal black; dark; dirty looking’

A third reason could be that “ABB is the (a) *basic form* in which (b) *sound symbolism* could be fruitfully explored [authors’ highlights]” (T’sou 1978: 67). T’sou views this predominantly in terms of valence, with (Cantonese) ABB often carrying a negative connotation. T’sou’s seminal study has two important implications. On the one hand, the mentioning of a recognizable template that involves (b) sound symbolism with often vivid meanings is reminiscent of the generally accepted cross-linguistic definition of ideophones as “ideophones are marked words that depict sensory imagery and belong to an open-lexical class” (Dingemanse 2012; 2019). ABB is definitely marked as a template, and often depictive of vivid sensory images. We will return to this idea below. On the other hand, noting that ABB is the (a) basic form implies that there are other, less basic forms or templates that might have similar traits as the ‘basic’ prototype of such constructions.

Despite this early observation dating back to the 1970s (T’sou 1974; 1978) and a continued line of thinking in terms of ideophones in English studies (Mok 2001; Kwok, Chin and Tsou 2016), traditional Chinese approaches have not taken these ideas up, focusing instead on the more structural properties of ABB types. That is, these studies are mostly interested in questions like: which BB patterns co-occur with A; what combinations of parts-of-speech are observed; or what different derivational patterns can be found” Studies have focused on Cantonese and other Yue dialects (Bodomo 2006; Sew 2008; Kwok 2012; Lai 2015; Kwok, et al. 2016), Southern Min (Chang 2009) and Mandarin (Shào 1990; Cáo 1995; Zhāng 2005; Yáo 2006; Lǐ 2008; Sūn 2012; Wang 2010; 2014; Lǐ 2019; Wáng 2020; Zhào 2021). It needs to be stressed that this list is by no means exhaustive. On the contrary, it is indicative of just how much the ABB forms stand out in linguistics and how much research has been devoted to them.

This study aims to reframe the traditional Chinese perspective by operationalizing ABB as constructions involving ideophones (BB)[4], similar to the interpretation originally advanced by T’sou (T’sou 1978; Mok 2001; Kwok et al. 2016). It explores what it means to take seriously the statement that ABB constructions are the “basic form for sound symbolism” by exploring

[3] Chinese characters used in this study are generally Traditional, unless noted otherwise.
[4] We would like to thank an anonymous reviewer for pointing out that not all ABB forms are best seen as constructions involving ideophones, but that other well-known cognitive mechanisms may be at play, e.g., metaphorical and metonymical extensions. In Section 3, this idea is revisited under the term of ideophonizing coercion.

other similar constructions. This will be done in three steps. First a survey of the main relevant literature on ABB adjectives, then a shift in perspective involving ideophone and collocate constructions, and finally a corpus-based study that probes collocate-ideophone constructions found in the data from four single sources of information. These single sources of information are then put together in a "tuple", presenting a picture of collocate-ideophone constructions in Mandarin.

2. Background

In the vast literature on ABB items, there is a broad consensus regarding what kind of items are instances of the ABB type. ABB items typically include nouns, verbs or adjectives in the A slot, followed by a reduplicated element, e.g., *lǜ-yóuyóu* 綠油油 'glossy green'. However, superficial ABB combinations of one character or syllable followed by a different reduplicated character or syllable are barred from being treated as ABB items, e.g., the Mainland actress Fan Bingbing 范冰冰 or the Mainland linguists Zhao Qingqing 赵青青 or Zhang Weiwei 张炜炜. Nor are cute diminutives typically considered as ABB, e.g., *chī-fànfàn* 吃飯飯 'eat some rice', *shuì-jiàojiào* 睡覺覺 'get some sleep'. Edge cases that are fundamentally not ABB items constitute trisyllabic monomorphemic onomatopoeias, like *hōnglónglóng* 轟隆隆 'rumbling of thunder'. In other words, it is clear there is some sort of internal composition between A and BB.

Data in the literature typically comes from word lists provided in dictionaries or general works. Taking a recent study (Wang 2020) as emblematic, we can see that main data comes from three dictionaries: the 7th edition of the *Dictionary of Modern Chinese* (Institute of Linguistics, Chinese Academy of Social Sciences 2016), the 3rd edition of the *Standard Dictionary of Modern Chinese* (Lǐ 2013) and the *Dictionary of Adjective Usage in Chinese* (Zhèng and Mèng 2003). But older studies refer most often to the highly influential *800 words of Modern Chinese* (Lǚ 1980), for example Cáo (1995) or Wang (2014). Only in recent years are corpus data taken into account to study ABB items (see, for example, Lǐ 2019). The data pool most studies on ABB work with hovers around 300 types: Cáo (1995) counts 330 items, Wang 2014 (336), but Wáng (2020) 258 types.

Three main concerns have shaped the Chinese literature on ABB items: (a) the ways in which A and B can be combined to form different derivational patterns; (b) the analysis of BB as an empty suffix; (c) the item-level variability for different combinatory possibilities between A's and BB's. I will outline these concerns and raise some more for analyses that treat ABB as being formed through purely derivational rules involving empty suffixes. The third issue will be explored later in this paper (Section 6.3).

A first concern in ABB studies is related to the different ways A and B can be combined. Usually this is done by solely taking the orthographic representation (Chinese characters) into account, rather than including phonological aspects. After all, a pervasive folk model of Chinese is the notion that one character (orthographic form) equals one syllable (phonological form) and one meaning, in a convergence of shape, sound and meaning (*hànzì de xíng yīn yì* 漢字的形音義) (Hsieh 2006) or the one syllable/one meaning principle (Sun 1999). In other words, the A and B in ABB stand both for syllables as well as characters. This folk model is fundamental for understanding the typology laid out in ABB studies.

Shào (1990), for example, presents four main types of ABBs, based on whether patterns like AB, BA, BBA, AABB, ABAB and BABA occur on top of the ABB form. The classification (Table 1) includes: (type 1) an independent combination of A and BB, (type 2) a reduplication of B from an existing AB, (type 3) a reduplication of B based on an existing BA; (type 4) a subtraction of one A in an existing AABB. This basic categorization is largely followed in subsequent studies. For instance, in his discussion on headedness, Wang (2014) also recognizes the first three types, the fourth one being subtractive rather than additive, and thus standing out.

Table 1. Main typology of ABB types (adapted from Shào 1990)

Type	Morphological pattern	AB	BA	BBA	AABB	ABAB	BABA	Example numbers
1	A + BB							(2a-b)
2 a	(AB) + B	✓			✓			(3a-b)
2 b	(AB) + B	✓				✓		(3c-d)
2 c	(AB) + B	✓						(3e-f)
3	BA + B		✓	✓				(4a-b)
4	AABB - A				✓		✓	(5a-b)

(2) a. *gān-bābā* 乾巴巴 ‘dull and dry’
b. *pàng-dūndūn* 胖墩墩 ‘flabby, stout’

(3) a. *tián-mìmì* 甜蜜蜜 ‘very sweet’
b. *péng-sōngsōng* 蓬松松 ‘fluffy, puffy’
c. *huǒ-làlà* 火辣辣 ‘spicy, hot’
d. *xiān-nènnèn* 鮮嫩嫩 ‘fresh and tender’
e. *qì-fēnfēn* 氣憤憤 ‘furious’
f. *yóu-nìnì* 油膩膩 ‘greasy’

(4) a. *xiāng-pēnpēn* 香噴噴 ‘fragrant’
b. *lěng-bīngbīng* 冷冰冰 ‘ice-cold’

(5) a. *bìng-wāiwāi* 病歪歪 ‘sickly-looking’
b. *mì-mámá* 密麻麻 ‘thickly clotted’

A derivational analysis like the one at hand here entails that AB or BA (type 2 and 3) are more basic than the ABB form. Perhaps diachronically one can trace such extensions, but it seems unlikely that the ABB form is less basic than AB. Under a Cognitive linguistics perspective (especially in Cognitive Grammar), these different combinations (AB, BA, BBA, etc.) certainly exist, but rather *alongside* ABB, instead of through rules that convert an underlying AB to ABB. Let us take *tián-mìmì* (3a) as an example (type 2a), for which *tián-mì* ‘sweet’ and *tiántián-mìmì* ‘sweet’ are also found. The conceptual content in these items is the same, namely A *tián* 甜 ‘sweet’ and B *mì* 蜜 ‘honey’. All three forms (AB, ABB and AABB) are well-known and not novel. In other words, they all have achieved unit status (Langacker 1987: 56ff.), but may differ with respect to the degree of entrenchment (Langacker 2017). It is the markedness of the ABB schema which highlights the vivid meaning of *tián-mìmì*. Note that schemas like AABB also confer special meaning, see Paul (2006) or Otting (2019), but these fall outside the scope of this study. To summarize, the occurrence of ABB forms may not necessarily be driven by derivation but rather application of constructional schemas to the same conceptual content.

The second main theme discussed in the literature concerns the internal morphology of ABB items. BB is often analyzed as an empty suffix. For instance, a recent reference grammar of Mandarin Chinese characterizes *hōnghōng* 烘烘 in *chòu-hōnghōng* 臭烘烘 ‘smelly, stinky’ as follows: “Such a suffix typically does not have very clear semantic content but does make a distinctive contribution to the overall meaning of the derived adjective” (Huang et al. 2016: 284). This distinctive contribution of BB is vivid in nature, according to the authors. But if there is a distinctive contribution in meaning, then that might lead us to the question whether BB is a content morpheme instead of an empty suffix. In the examples listed thus far, BB does not behave like other well-known suffixes in Chinese, such as the nominalizers *-zi* 子, *-er* 兒, or *-tou* 頭, an observation already made early on (Shào 1990). Such suffixes have undergone complete semantic bleaching when they occur as a suffix[5] (Packard 2000: 174). Thus, it stands to reason to investigate whether BB patterns truly fit in with these prototypical suffixes, or even affixes.

According to Taylor (2002), prototypical affixes are definitely (a) bound, (b) usually cannot be stressed, (c) often somewhat integrate in the phonological shape of a word of which they are a part, and (d) are highly selective to the items to which they attach. Prototypical words, Taylor argues, take the opposite of these four criteria. For the three nominalizing suffix

[5] The ‘full’ meaning of *zǐ* 子 is ‘son’, that of *ér* 兒 also ‘son’, and *tóu* 頭 ‘head’.

examples, criteria (a-b) and (d) intuitively seem true. Criterion (c) seems true for *-er* in the sense that er-integration '*érhuàyīn*' 兒化音 is a well-studied phenomenon in Mandarin. And the other two (*-zi, -tou*) are prone to lose their original lexical tone (at least in Northern variants).

Let us return to BB. It has often been noted that some BB forms can stand on their own, e.g., *mángmáng* 茫茫, while others are completely bound, e.g., *hūhū* 乎乎, *càncàn* 燦燦 (Shào 1990; Wang 2014). In other words, a number of BB elements are indeed argued to be (a) bound (Wang 2014). Continuing with the criteria for stemhood vs. affixhood, it (b) is possible to stress BB. BB's do not (c) integrate in the phonological shape (the "root" A); instead, they retain their full pronunciation. They are (d) highly selective to the items "to which they attach", or rather co-occur. These differences with the prototypical affix point in the direction that BB is not an affix, but rather a word.

Curiously, while Shào (1990) recognizes that BB does not behave exactly like the three nominalizing suffixes (*-zi, -er, -tou*), he still states that it makes sense to accept the traditional characterization of BB as a largely empty suffix currently undergoing grammaticalization. He finds evidence for this process in three manners. First, he states that BB only occurs after A, and that A and BB belong to the same part-of-speech category. The first proposition contradicts with Shào later stating that BBA forms also occur, as T'sou (1978) also pointed out. The second proposition seems doubtful, given that the literature also devotes considerable attention to which parts-of-speech can fill the A slot. Second, Shào points to the fact that some BB forms have the same pronunciation but different characters, but that this phonological similarity has led to orthographic fuzziness (6). Third, he mentions BB's that have different orthographic variants (7). These two issues make sense if BB is considered as ideophonic in nature, or occurring in an ideophonizing template (Dingemanse 2017; Akita and Imai 2022). This interpretation makes sense given that BB often is hard to define but is perceived as contributing a certain vividness to ABB, and the dynamic polysemy caused by similarity in the orthography of Chinese ideophones (Van Hoey and Lu 2019; Van Hoey 2020).

(6) Orthographic fuzziness
   a. *rè-hūhū* 熱呼呼 'warm', also written as 熱乎乎
   b. *xiě-hūhū* 血糊糊 'blood-stained', also written as 血乎乎

(7) Orthographic variants
   a. *xiào-mīmī* 笑咪咪 'smiling'
   b. *xiào-mīmī* 笑眯眯 'smiling'
   c. *xiào-mīmī* 笑迷迷 'smiling'

What emerges is that the distinction between stem and affix is not crystal clear in the case of A and BB, which is not entirely unexpected (Tuggy 1992), although it appears that the wordhood analysis of BB is stronger. This is corroborated by Yáo (2006), and also by Wang (2014), who notes that 91% of the 336 ABB items he observes are compounds of word-like elements rather than adjectival derivations with empty suffixes. And even in his attribution of derivation, e.g., *gān-bābā* 乾巴巴 'dry' or *shī-hūhū* 濕乎乎 'wet', in which BB elements are argued to have lost their lexical meaning and "do not contribute to the meanings of the whole words, so they should be treated as suffixes" (Wang 2014:354), one can wonder if derivation is truly at play (see above). After all, derivation quickly leads to the exclusionary fallacy and the rule/list fallacy (Langacker 1987: 28–29) and the process metaphor (Langacker 1987: 63). These principles hold that particular statements (lists) are to be excised from the grammar if general statements (rules) can be established that subsume them. For example, a noun like *stapler* presents a dilemma to a processual treatment of forms: if the form *stapler* is derived by rule, it is impossible to account for its meaning being more than 'something that staples'; if it is simply listed in the lexicon, the productive V + -er derivational pattern, which certainly can be identified, cannot subsume it (Langacker 1987: 28). Or in other words, as Wang states, if the BB form has no direct lexical meaning, it is a suffix and nothing else; if it has meaning, it must be following a processual rule that is the sum of its components, which then comes in a typology like the one presented before Table 1. Mindful of the rule/list fallacy, and the warnings raised against a pure derivation analysis, it appears that ABB forms truly do stand on their own. However, rather than using the dichotomy of lexical BB forms vs. empty BB forms, we are dealing with different degrees of transparency in their composition. We can illustrate with the examples shown in Figure 1.

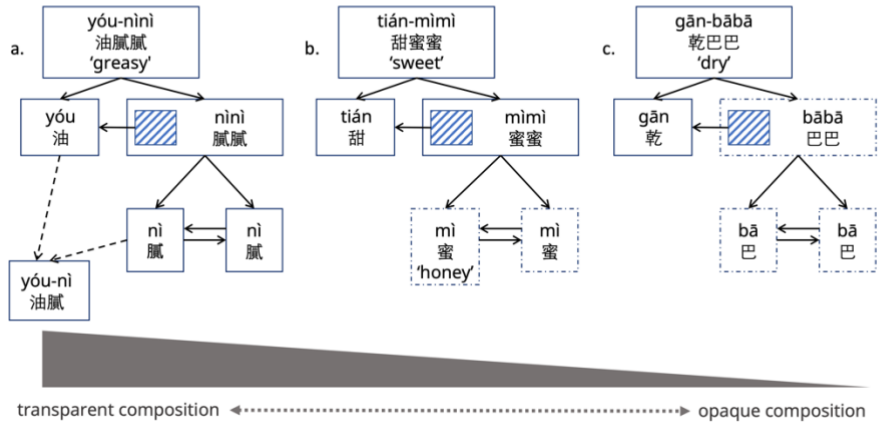

Figure 1. Three kinds of compositional transparency on a cline. Panel a shows full transparency. Panel b displays partial transparency that involves a semantic extension. Panel c is fully opaque.

The three diagrams shown in Figure 1 follow the typical conventions of Cognitive Grammar, with rectangles depicting so-called elaboration sites (Langacker 1987; Tuggy 1992). They show that on one level of description, the component is dependent on a more autonomous component. In the case of Figure 1, *nìnì* is dependent on *yóu*, just as *tián* is more autonomous than *mìmì*, and *gān* elaborates *bābā*. Panels A, B, and C in Figure 1 differ with respect to unit status (Langacker 1987) on the different levels of composition. In panel A, items have reached unit status on all levels. For panel B, the situation is slightly different. The smallest component *mì* by itself means 'honey'. The reading of *mìmì* as 'sweet' originates from 'honey' through a conventional metonymy of QUALITY FOR PRODUCT. In panel C, *bābā* is recognizable in a number of ABB patterns (Cáo 1995; Wang 2014), and has achieved unit status, but much less so than the other two cases. As a single character, *bā* is even less available with a clear meaning. The schema analysis of ABB that has been advanced here has led us to the realization of different levels of entrenchment, depending on the degree of unit status a component has reached. It is perhaps best to think of these three situations as appearing on a cline. Sometimes the composition is transparent, other times it remains opaque. For example, *hēi-qīqī* 黑漆漆 'black-as.lacquer', can be assumed to belong to the transparent after semantic extension kind, but an informal discussion with a young student showed that they were unaware that *qī* means 'lacquer' that its main feature is its black color, at which point the well-known type of *hēi-qīqī* should probably be situated further along the cline towards the third kind. In all these cases, the sum does seem more entrenched than its parts.

Finally, the third main theme in the literature is related to the number of type combinations, or n-to-n relations (Shào 1990; Cáo 1995; Zhāng 2005; Wang 2014). Some BB

forms only occur with a single A (8a-b), that is, there are no other A-*áiái* combinations. Others are characterized by relations involving different A's followed by the same BB (9a-c). In (9a-c), *méngméng* means 'indistinct, can't see clearly'. There are also some A's that only occur with one BB, which also only occurs with that A, in a unique combination (10). Group (8) differs from (10) in that the A *zuì* only occurs in an ABB form with *xūnxūn*, while (8b) and (9a) show that this is not the case for *bái*. To make this clearer, (11a-c) shows that there are indeed some A's that occur with many different BB's. A general reason for why some engage in many-to-many relations is lacking; it appears highly item-dependent.

(8) a. *hēi-qīqī* 黑漆漆 'pitch dark'
b. *bái-áiái* 白皑皑 'pure white'

(9) a. bái-méngméng 白濛濛 'hazy'
b. *huī-méngméng* 灰濛濛 'dusky'
c. *hēi-méngméng* 黑蒙蒙 'dark and indistinct'

(10) *zuì-xūnxūn* 醉醺醺 'drunk'

(11) a. *hēi-qīqī* 黑漆漆 'pitch dark'
b. *hēi-méngméng* 黑蒙蒙 'dark and indistinct'
c. *hēi-xūxū* 黑魆魆 'pitch black

By now it is clear that the ABB construction has attracted considerable attention because the schema is marked and open-class, in the sense that there are a considerable number of types that are agreed upon to belong to it. It has also been shown that traditional approaches tend to see relations between ABB and its recognizable components derivational in nature, but that this is not universally agreed upon. Furthermore, rather than BB being an empty suffix, ABB items vary in terms of compositional transparency. Combinations between A and BB are also item-dependent. If there are differences between members, that would strongly suggest that we are dealing with salience effects for the members. In other words, a prototypical structuring. In Section 3 this idea is further explored at a higher level, i.e., for the ABB schema itself.

3. Collocate and ideophone

The only morphological division that has been agreed upon by all studies on ABB is that there is a break between the A part and the BB part (Figure 1), i.e., A-BB, but that is not saying very much. It is now time to turn first to the supposed emptiness of BB, which we saw was the perspective taken by traditional approaches (T'sou 1978; Shào 1990; Huang et al. 2016).

Most studies indicate that ABB typically has a more vivid meaning or connotative feeling than if A (or AB or BA etc.) were to occur by itself. BB may indeed not have very clear semantic content, which is taken to mean descriptive content, but there definitely appears to be a depictive quality to it. Furthermore, many if not all of these ABB items relate to sensory imagery, and the synesthetic metaphorical mappings has been demonstrated to be A and BB rampant (Zhao 2021). This reference to sensory imagery fundamentally changes how we should see ABB: rather than viewing them as derivations of the A element, whether that be a noun, verb or adjective, they conform to the major criteria of Dingemanse's (2012; 2019) cross-linguistic definition for ideophones. After all, I have established that that the ABB schema is (a) marked, and consists of (b) open-class (c) words that (d) depict (e) sensory imagery. It should be noted again that this is not the first time ABB is regarded as a typical ideophone. In fact, the introduction already mentioned the pioneering work of T'sou in the 1970s. And recently, it was also explored in Japanese (She 2015).

Note that this account does not deny the substantial degree of conventional arbitrariness present in the form-meaning mappings of ABB items (not all these items are iconic), nor does it preclude analyses based on other cognitive mechanisms like metaphor or metonymy, e.g., the aforementioned fully transparent *yóu-nìnì* or semi-transparent *tián-mìmì* in Figure 1. Such mechanisms are likely also active alongside the ideophonizing pattern of ABB (Dingemanse 2017; Akita and Imai 2022), which highlights the vividness of these form-meaning mappings. In other words, the core of ABB items likely contains "real and iconic" BB ideophones, while many items undergo an ideophonizing coercion by virtue of the template. In what follows, I will refer to all such BB items as ideophones, with the caveat that this may be an oversimplification of the items that participate in the construction..

Continuing along the ideophone track, we find Mok (2001), who studied sound symbolism in Mandarin, Cantonese and Hakka from a phonological perspective. She correctly differentiates the nature of a prosaic A and more sound symbolic BB. She recodes them as X and A, resulting in many observed patterns that go beyond the full reduplication we have encountered thus far. This attention to the full reduplication (BB in ABB = AA in XAA) is presumably the result of the one-syllable-one-meaning principle referred to above: full reduplication tends to stand out more than partial reduplication, especially if you are dealing with written data in the form of Chinese characters. Mok (2001), on the other hand, already includes partial reduplication, e.g., Cantonese *caau*^{4}*-mi*^{1}*mang*^{1} 巢咪嗼 'wrinkled' or *faa*^{1}*-li*^{1}*luk*^{1} 花哩碌 'colorful, flowery', which both contain alliterative ideophones.

Now we will take a few steps that follow from Mok's (2001) characterization of a prosaic X and a sound symbolic AA, and the notion of different degrees of combinability. First, it appears that the combinations are largely idiomatic, in the sense that it is not wholly predictable

which prosaic elements can combine with which sound symbolic elements, but there are most definitely accepted combinations[6]. Still, it may then be of analytic interest to see X and AA as being instantiations of a higher-level constructional schema involving a collocate and ideophone of the full reduplication type, to see what can be generalized and what is particular to the exemplars.

In a subsequent step, we can investigate if there are other lower-level schemas that are sanctioned by the higher-level collocate-ideophone schema. Thus far we have been considering trisyllabic schemas of the form A-BB/X-AA. But Mok's (2001) study shows that full reduplication is not the only trisyllabic pattern to be taken into consideration: we need at least rhyming and alliterative ideophones, i.e., X-AB (where AB stands for two different syllables = two different characters).

While ideophones certainly come in more syllabic numbers, the Chinese Ideophone Database (Van Hoey and Thompson 2020) indicates that disyllabic ideophones by far outweigh the other types. They find that 3479 types out of 4948 collected consist of two syllables (1 syllable 5%; 2 syllables 70%; 3 syllables 6%; 4 syllables 12%; other 7%). Furthermore, the disyllabic ideophones in the Chinese Ideophone Database have a proportion of 60/40 (2271 items have different characters, 1458 items are composed of a reduplication). Thus, it makes sense to increase the scope of the ideophonic part to disyllabic forms. A full picture of collocate-ideophone constructions may also need to look at ideophones of a different number of syllables, but in this first study a presumed coverage of 70% is a good starting point. One can also venture that straying too far from the marked disyllabic ideophone pattern in the original ABB pattern reduces the family resemblance of the items under investigation. On the other side of the construction, there are no a priori theoretical limitations for the number of syllables for the prosaic collocate. For example, Yáo (2006) also states that it may be fruitful to look at collocates that are longer than a single syllable, e.g., (12-13). After all, we might miss out on some of the meaning elaborations of ideophones if we restrict ourselves only to monosyllabic collocates. For instance, it could be important that *yīnyīn* only describes the fresh and verdant green of grass (13b), but perhaps not of the leaves on trees — trees and grasses being two different ontological plant categories within the Chinese lexicon (Chao 1953). Additionally, not increasing the scope of collocates means we will miss out on items like (14), which do not occur in a trisyllabic construction, at least not in the corpus used here.

(12) a. *bái-ái'ái* 白皑皑 'white-gleaming.white'

[6] A personal anecdote: one of the authors once overheard a group of young people play an ABB game in a restaurant in Taipei to pass the time. Each participant had to say an ABB form one after another and if they produced a meaningless combination or something the group didn't approve of, that participant had lost and the game continued.

b. *bái-xuě-ái'ái* 白雪皑皑 'white-snow-gleaming.white'

(13) a. *lǜ-yīnyīn* 綠茵茵 'green-verdant'

b. *lǜ-cǎo-yīnyīn* 綠草茵茵 'green-grass-verdant'

(14) *fēng-chén-púpú* 風塵僕僕 'wind-dust-dusty'

One more issue must be tended to before we can investigate what the implications are of treating the original ABB as an instantiation of collocate-ideophone constructional schemas, namely the fact that some ABB forms can also occur as BBA, e.g., *xiāng-pēnpēn* 香噴噴 vs *pēnpēn- xiāng* 噴噴香 (type 3 in Table 1). It would thus seem that we are dealing not only with collocate-ideophone types but also with the reverse, ideophone-collocate schemas. An illustration of the schemas that are conceptually included in the collocate-ideophone constructions is shown in Figure 2. Ideophone-collocate schemas can be diagrammed by mirroring the collocate and ideophone in the constructional slots, e.g., AA-X vs. X-AA (with different exemplars).

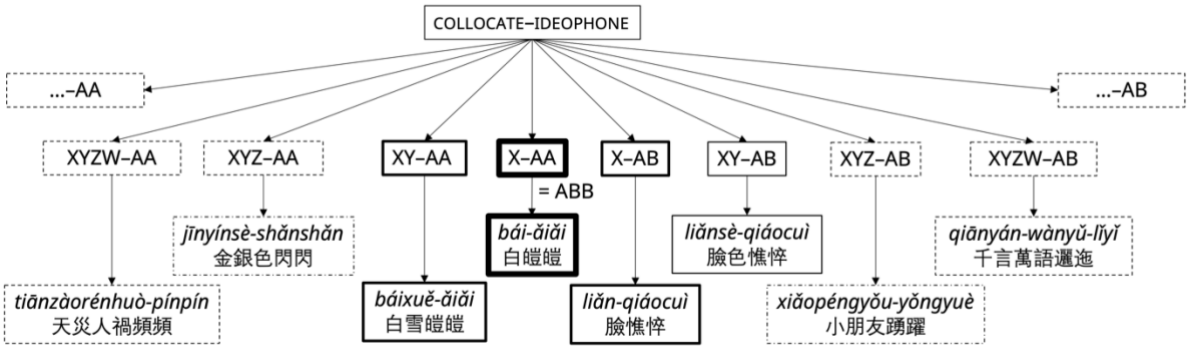

Figure 2. Collocate-ideophone schemas with exemplars (taken from the corpus presented below). The letters A and B refer to ideophonic syllables; the letters X, Y, Z, and W to syllables in the prosaic collocate.

Given all the attention ABB (X-AA in Figure 2) has received in the literature, it is expected that it is prototypical among these collocate-ideophone constructions. Figure 2 shows this difference in salience through varying line thickness. XY-AA and X-AB are less prototypical than X-AA, but still quite familiar. However, increasing syllables in the prosaic collocate (XYZ, XYZW, …) may instantiate exemplars that are less apprehensible as instances of the same construction. To verify if X-AA (ABB) is truly the prototype of these constructions, as the literature strongly indicates, a usage-based study may be in order. This can either be done in an experimental manner (for example, Rosch 1975; Kelly, Bock and Keil 1986), or we can use a corpus-based

method, which has become increasingly favored in Cognitive Linguistics (Tummers et al. 2005; Divjak and Arppe 2013; Gries 2019).

In what follows, I will present relevant corpus data. The identified corpus patterns will be related to the schemas we identified in Figure 2. Finally, I will go beyond the schemas and demonstrate the advantage that increasing the scope of ABB studies to larger collocate-ideophone constructions brings.

4. Materials

There are two main sources of data I rely on for the analysis. The first is the Academia Sinica Balanced Corpus of Modern Chinese[7] (CKIP group and Academia Sinica 2013) and the second is the Chinese Ideophone Database (Van Hoey and Thompson 2020). This combination of database and corpus stands in contrast with most ABB studies, which rely on types collected in wordlists found in dictionaries. The data and R code for the analysis can be found in the accompanying OSF repository[8].

Collocate-ideophone constructions were identified in three operationalized groupings. The first operationalized group of data comes from disyllabic ideophones listed in the Chinese Ideophone Database, together with a word preceding them in the corpus, e.g., *lù-kǎnkě* 路坎坷 'road-bumpy'. The second group contains all structurally valid ABB expressions found in the corpus. This was done to make sure BB items not listed in the Chinese Ideophone Database would not be left out. However, this overgenerated hits, e.g., words of the type *Huáng jiějiě* 黃姊姊 'big sister Huang'. As a result, this group was cleaned by omitting items with tags like noun, foreign word, numeral etc. A third group of data was selected according to the criteria of the first group, but now looked beyond the word border, with word defined in terms of the ASBC corpus tagging. That is to say, while in the first group the ASBC had segmented *huáng-dēngdēng* 黃澄澄 'glistening yellow' as one word ("黃澄澄"), now in the third group items like "白色 蒼茫" (*báisè-cāngmáng* 'indistinct white') are caught, which have a space between them. After all, these also conform to the larger collocate-ideophone constructional family.

Ideophone-collocate constructions were obtained in the same, but mirrored, manner. These groups were all merged into one dataset, excluding duplicate values. These were manually inspected. The final dataset consisted of 5634 tokens, divided over 16 constructional subschema types.

[7] The ASBC corpus can be accessed here http://asbc.iis.sinica.edu.tw/ . It comprises about 10 million words. The version consulted in this study was institutionally accessible to the authors.
[8] https://osf.io/2m58e/

5. Relative salience of subschemas and types

We can now investigate the relative salience of the subschemas. It has been argued before that the productivity of a schema is largely determined by its type frequency (Bybee 2001). In other words, schemas with a higher type frequency can be considered more salient. Salience here is understood as meaning 'highly entrenched in a given context' (Geeraerts 2017; see also Schmid and Günther 2016).

Table 2 shows that there are 2690 types with 5634 tokens. These are not uniformly distributed. Instead, we find the collocate-ideophone schemas of X-AA, XY-AB, XY-AA, as well as the ideophone-collocate schemas of AB-XY, and AA-XY are particularly bountiful. If we just limit our scope to trisyllabic expressions, X-AA (or traditionally "ABB") truly stands out as the only schema with such a high type frequency, but also token frequency. The other major productive patterns are all tetrasyllabic. This is not entirely surprising, as Chinese is known to possess many different idioms consisting of four characters, and also ideophone combinations in four characters that date back to at least the *Book of Odes (*詩經, ca. 11th-7th century BCE) (Smith 2015; Féng 2016). Note that in tetrasyllabic schemas, partial reduplication (AB) had a higher frequency than full reduplication (AA), a pattern observed in both collocate-ideophone and ideophone-collocate schemas.

Table 2. Type and token frequencies for all identified subschemas

Collocate-Ideophone			Ideophone-Collocate		
Subschema	Type frequency	Token frequency	Subschema	Type frequency	Token frequency
X-AA	465	1779	AA-X	74	176
X-AB	90	137	AB-X	43	58
XY-AA	532	1268	AA-XY	240	304
XY-AB	775	1363	AB-XY	325	393
XYZ-AA	40	41	AA-XYZ	16	17
XYZ-AB	62	67	AB-XYZ	13	14
XYZW-AA	5	5	AA-XYZW	2	2
XYZW-AB	8	9	AB-XYZW	1	1

It has thus become clear that there are different degrees of salience to be found amongst these subschemas. But are there similar distributions of type-token frequencies to be found within the subschemas? After all, one could reason that since the traditional literature takes

wordlists as the main source of data, the descriptions therein are supported by the idea that all ABB forms are equal in terms of structure, and hence equally good exemplars of ABB. Of course, the native speaker intuitions provided in the discussions provide some nuance: some exemplars are more representative of ABB than others (because they keep getting mentioned). One can wonder what differences in salience the corpus data harbor. In other words, are there types with a particular high token frequency, indicating that they are more entrenched and more salient, i.e., better exemplars of their respective subschema?

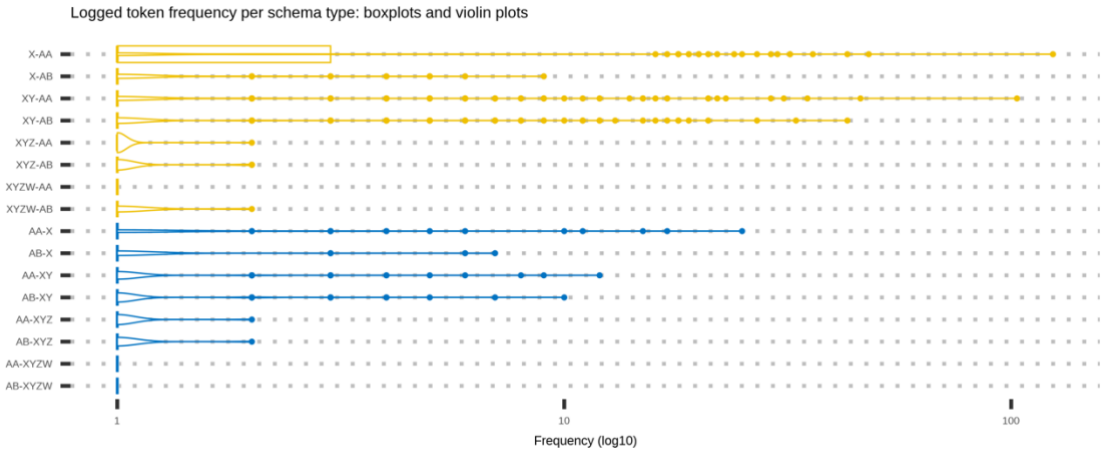

Figure 3. Logged token frequency per subschema type. The boxplots are overlayed with violin plots. Note that only for X-AA a visible boxplot could be drawn; the other subschemas only consisted of outliers.

As Figure 3 shows, every subschema has a median type-token frequency of 1. That means that most types only occur once in the whole corpus. Except for XYZW-AB, there are no outliers for hexasyllabic subschemas, which may be due to the overall low token frequency. The tetrasyllabic and trisyllabic subschemas, however, all have outliers. Note that X-AA ("ABB") is the only schema with a higher upper quartile, i.e., the token frequency of 3! In terms of outliers, we also find that X-AA has the highest outlier, i.e., *huó-shēngshēng* 活生生 'lively' (token frequency of 124, logged frequency of 2.09). This outlier is followed by *xiǎoxīn-yìyì* 小心翼翼 'very cautiously' of the XY-AA subschema. One important observation of Figure 3 is that there are more outliers for collocate-ideophone subschemas than for ideophone-collocate subschemas. Additionally, there are more outliers when syllables of collocates are short. A linear regression model with subschema frequency as dependent variable and weight of collocate and collocate/ideophone order as predictors showed that frequency, and relatedly these distributions with outliers, is indeed driven by collocate weight and order (significant

intercept for monosyllabic collocate and the order collocate-ideophone, p = 0.009), but at the same time no other variable settings significantly differ from the intercept.

The implications of Table 2 and Figure 3 are that when ABB is reconsidered as a subschema of schemas involving collocates and ideophones, there are other subschemas that also emerge as salient. However, the distribution is not uniform ($\chi^2(7) = 159,\ p < 0.001$)[9]. ABB, now rephrased as X-AA, does stand out as the most productive trisyllabic pattern. It is the only subschema for which type-token frequencies show a larger spread. Furthermore, the utility of corpus data on top of wordlist data has proved useful: theories built around ABB should pay attention to the non-equality of exemplars used. It makes more sense to figure out exemplars like *huó-shēngshēng* (token frequency of 124) rather than – *bīng-lěngleng* 冰冷冷 'ice-cold'[10], which has a token frequency of 1. It could also be, of course, that there are other good reasons to build a theory on exemplars like *bīng-lěngleng*, e.g., a unique pattern for a formal rule (Table 1), or a special attraction in a collostructional pattern, of the kind Stefanowitsch and Gries (2003) advanced.

6. Corpus-based tuples

Given that the corpus data has placed ABB words in a new light, it leads one to wonder what other perspectives the corpus can shed on these collocate and ideophone constructions. To respond to that question, I turn to the notion of tuples (Gries 2019). Gries argues that many sources of information presented in corpus linguistic studies conflate two or more dimensions of information, such as frequency and effect size of association measures, direction of association, or dispersion. His proposed remedy consists of calculating single sources of information, which can then be put together to create a more coherent picture, so-called tuples. Here I follow that line of thinking. Since we already have the corpus data for collocate-ideophone and ideophone-collocate constructions, the other single sources of information can be relatively easily obtained. I will include four sources of information: cue validity, token frequency of ideophones, dispersion of ideophones, and ideophones' constructional preference. Finally, by combining these dimensions of information, a picture of collocate-ideophone constructions emerges that is rampant with non-uniformity throughout — it just depends how you look at it.

[9] Because some of the expected values were smaller than 5, we also ran a Fisher Exact Test ($p < 0.001$).
[10] Note this is not *lěng-bīngbīng* 冷冰冰 from example 4b, which has a token frequency of 15.

6.1. Cue validity

The first dimension of information that may be of interest is the attraction between collocate and ideophone. As the literature has copiously discussed, some BB forms only occurred with one A, others with many different A's, see (8-11). In other words, we can think of collocate and ideophone as a construction with two slots: how likely is a given collocate a cue for a given ideophone, and vice versa. This relation between cue and an outcome has been found to be a strong determinant of prototype formation (Stefanowitsch and Gries 2003; Ellis and Ferreira-Junior 2009), and is often called 'cue validity' (Stefanowitsch and Gries 2003).

This can be solved with directional association measures like ΔP (Ellis and Ferreira-Junior 2009; Gries 2019), or more general association measures like the Kullback-Leibler divergence D_{KL} (Baayen 2011; Gries and Durrant 2020), both of which we find correlated ($r > 0.8$) in our data. These association measures all work from the same premise: you need to calculate contingency tables which show the frequencies of co-occurrence combinations for the phenomena you are interested in. Let us take *huó-shēngshēng* as an example[11]. Table 3 shows four logical combinations: (a) is the co-occurrence of the collocate and ideophone in question; (b) is the collocate with other ideophones; (c) the ideophone in combination with other collocates; (d) the rest of the combinations. This type of analysis has also been termed co-varying collexeme analysis (Gries and Stefanowitsch 2004). Note that no distinction is made between collocate-ideophone and ideophone-collocate constructions, as that is a different source of information.

Table 3. Example of the contingency table of *huó-shēngshēng* 活生生 'lively'

Collocate	**Shēngshēng 生生**	**¬ Shēngshēng 生生**	**Row totals**
Huó 活	(a) 124	(b) 11	(a+b) 135
¬ Huó 活	(c) 43	(d) 5456	(c + d) 5499
Column totals	(a + c) 167	(b + d) 5467	5634

The calculation of D_{KL} is quite complex. $D_{KL:ideophone \rightarrow collocate}$, the attraction an ideophone has on a collocate,(15) shows that given an ideophone, how likely another collocate is to be found as well. $D_{KL:collocate \rightarrow ideophone}$ (16) shows the relative attraction of a collocate unto an ideophone. These are normalized (17), so that higher values indicate a stronger cue validity. For *huó-shēngshēng*, it seems that *huó* attracts *shēngshēng* slightly more than the other way around, since $D_{KL:ideophone \rightarrow collocate\ norm}$ equals 0.96, and $D_{KL:collocate \rightarrow ideophone\ norm} = 0.99$. Note that in these normalized cue validity measures,

[11] This example was chosen as a didactic example because all cells in Table 3 have values.

a higher score means stronger relative attraction, with values close to 1 indicating near-perfect attraction.

(15) $$D_{KL_{ideophone \rightarrow collocate}} = \frac{a}{a+c}\ log_2 \frac{a\,(a+b+c+d)}{(a+b)(a+c)} + \frac{c}{a+c} log_2 \frac{c\,(a+b+c+d)}{(a+c)(c+d)}$$

(16) $$D_{KL_{collocate \rightarrow ideophone}} = \frac{a}{a+b}\ log_2 \frac{a\,(a+b+c+d)}{(a+b)(a+c)} + \frac{b}{a+b} log_2 \frac{b\,(a+b+c+d)}{(a+b)(b+d)}$$

(17) to normalize: $1 - e^{-D_{KL}}$

The major benefit of using these directional association measures lies in verifying the observed variability between collocates and ideophones. Let us illustrate with two examples. It has been noted that *zuì-xūnxūn* 醉醺醺 'drunk' is a unique combination. *Zuì* only prefers *xūnxūn* and vice versa. If that is true, then their D_{KL} values are predicted to be (1,1). This appears as predicted in Figure 4. Both when *zuì* (panel A) or *xūnxūn* (panel B) are queried, only one value is returned, namely *zuì-xūnxūn*, at the position of (1,1). This means that in this case the wordlist data and intuitions were corroborated by the corpus data.

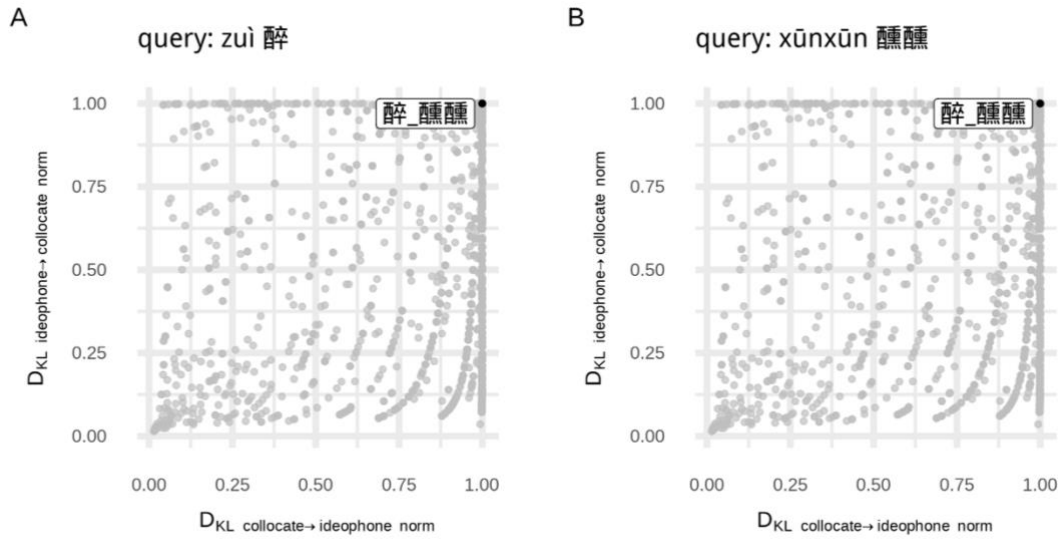

Figure 4. Cue validity (Kullback-Leibler divergence) for *zuì* (panel A) and *xūnxūn* (panel B). Grey points indicate the calculated D_{KL} values for all ideophone and collocate combinations. The bold black points show ideophone and collocate items that belong to a given query; in this case both queries result in the same plot.

Another fully reduplicated ideophone for which the literature has pointed out that there is only one collocate, is *qīqī*, yielding the predicted unique combination *hēi-qīqī* 黑漆漆 'black as lacquer' (example 8a). On the other hand, it has been pointed out in many places that *hēi* can have many different BB's. Based on the literature, panel B is expected to have a single value, namely *hēi-qīqī*. Furthermore, *hēi-qīqī* is predicted to have a high y-axis value but not necessarily a high x-axis value. Panel A should then show many different items that combine with *hēi*.

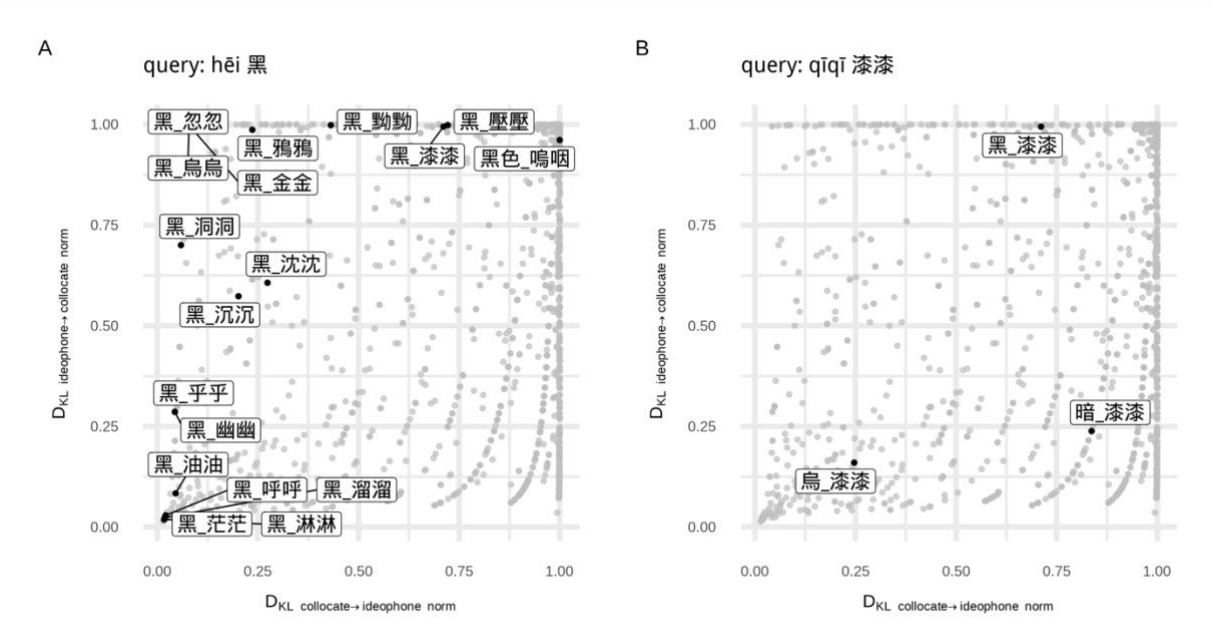

Figure 5. Cue validity (Kullback-Leibler divergence) for *hēi* (panel A) and *qīqī* (panel B). Note that many of observed combinations in this corpus may not be entrenched for the full population but may be particular to the corpus used for generating these statistics, e.g., *hēi-línlín* 黑淋淋.

As Figure 5 shows, the predictions were only half correct. Panel A indeed shows many different combinations for the collocate *hēi*, and *hēi-qīqī* does have a high value on the y-axis. However, *qīqī* also occurs with two other collocates. Does this then mean that the traditional account of *hēi-qīqī* is fundamentally wrong? The answer is negative; *wū* 烏 also means 'black, dark < raven' and *àn* 暗 also means 'dark'. The appearance of *wū* may be a Taiwanese influence, see example (1). At the same time, this may also be a case of sociolinguistic variation or language change, namely, the emergence of a lesser dominant pattern.

This single dimension of information (cue validity) can act as a catalyst for further lexicographic exploration, with benefits for second language acquisition as well: idiomatic expressions are harder to learn when they are not transparent (Nippold and Taylor 2002; Tabossi,

Fanari and Wolf 2008), which is often the case for ideophone-collocate constructions. The potential pedagogical benefit of calculating association measures lies in curriculum development: first focus on the ideophone-collocate constructions that have high cue validity, either from the ideophone as cue or from the collocate as cue.

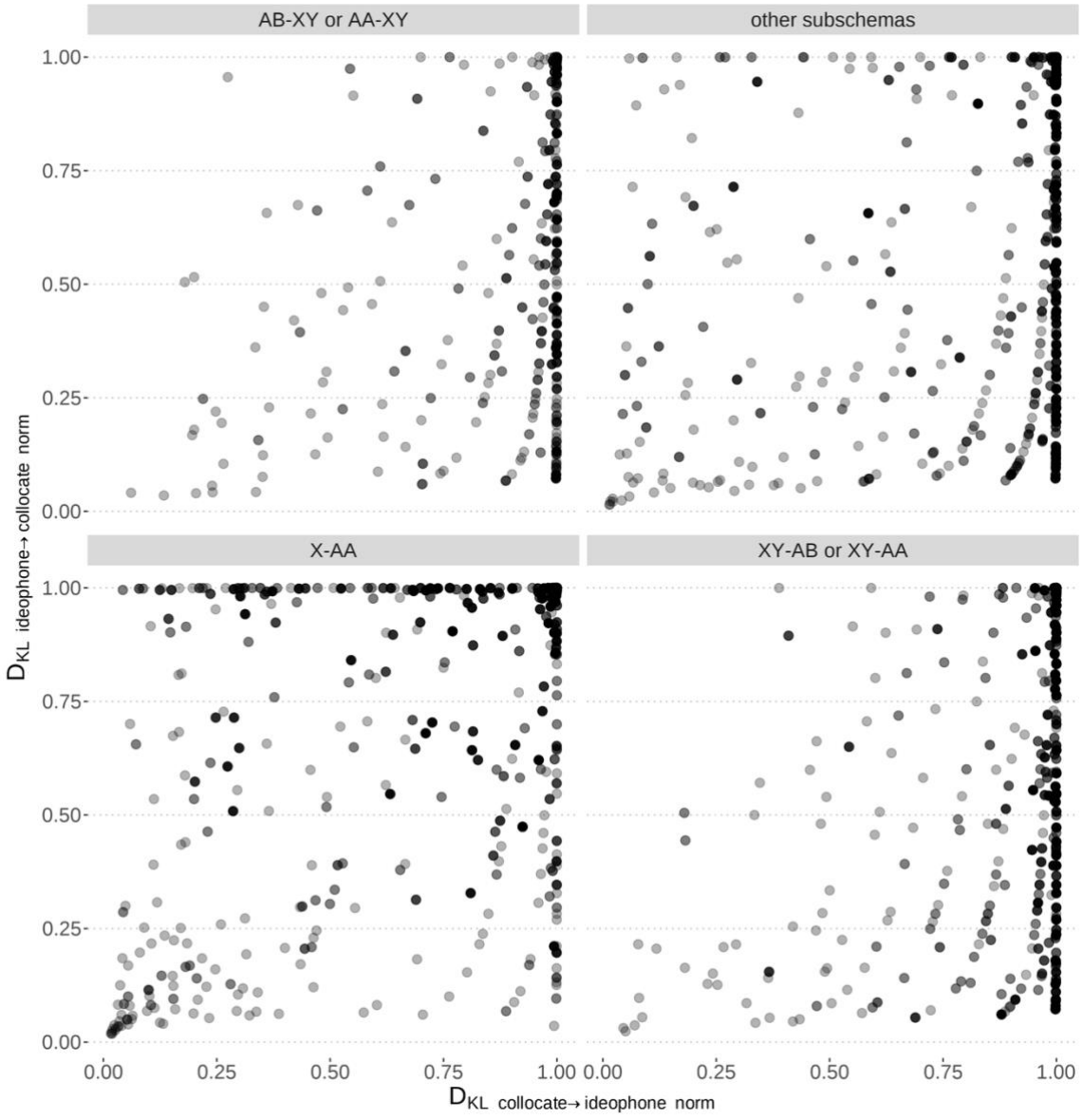

Figure 6. Differences in cue validity (Kullback-Leibler divergence) for four groups of subschemas

Before we move on to the next single dimensions of information, let us briefly inspect the configuration of the two cue validity measures. Figure 6 consists of four subsets of the data. Based on the highly productive type and token frequencies presented before (Table 2), it is

useful to inspect separately ideophone-collocate schemas like AB-XY and AA-XY (upper left), collocate-ideophone schemas like XY-AB and XY-AA (lower right), the prototype X-AA (lower left) and the rest of the less productive schemas (upper right). It immediately stands out that collocation patterns of most schemas are driven by the collocation, rather than the ideophone (higher values on the x-axis). The prototypical collocate-ideophone construction X-AA, however, is different: more high values on the y-axis are observed, suggesting that it is the ideophone cue that attracts the collocate rather than other way around. This may present another piece of evidence that the so-called suffix BB is not empty after all, but very much meaningful.

6.2. Frequency and dispersion

A second dimension of information constitutes token frequency, which has already been addressed for all instances of the identified subschemas in Figure 1 and Table 2. We can also calculate the token frequency of the ideophones separately, as it has already been pointed out that the majority of full expressions have a token frequency of 1, see Section 5.

Related to token frequency is a third dimension of information: dispersion. Given a corpus, it has been argued (Gries 2008; 2020; 2021) that it is actually dispersion and not so much token frequency that is responsible for well-entrenched units. Types that are found throughout many different corpus parts have a higher dispersion. Given that they occur in many different contexts, they are presumably well-known to a variety of language users. However, jargon-like words may have a much higher overall token frequency, yet be restricted to a certain genre or even text or corpus part.

Dispersion can also be calculated with many different measures. Gries has proposed DP in the past (Gries 2008; Lijffijt and Gries 2012), but has recently also advocated for Kullback-Leibler divergence application for dispersion (Gries 2021). Here we also follow this measure of $D_{KL:dispersion}$. The calculation follows similar principles as the Kullback-Leibler divergence measures for cue validity presented in Section 6.1[12]. A helpful tutorial is offered in Gries (2021), and the code for our calculations is provided in the supplementary materials on OSF. Here (Figure 7) we inspect the main tendencies of the dispersion of the ideophones in our dataset. The corpus parts I take into consideration are the genre divisions of the Academia Sinica Balanced Corpus of Modern Chinese, i.e., literature, life, society, science, philosophy, and art.

Figure 7 shows the logged token frequency of ideophones in our dataset (657 types across 5634 tokens, the same as for all expressions) in relation to the normalized dispersion values.

[12] In principle, the formula is $D_{KL:dispersion} = (\frac{posterior}{data} || \frac{prior}{theory})$: how much does the probability distribution of the two corpus parts, given the word we are currently looking at (the posterior), diverge from the percentage distribution of the corpus sizes (the prior).

These normalized values indicate that higher values represent wider dispersion. In other words, a dispersion of 0 would mean occurrence only in one genre of the Academia Sinica corpus, while 1 indicates a balanced occurrence across the different genres of the Academia Sinica corpus. There is a clear negative linear correlation between the frequency of ideophones and skewed dispersion. That being said, the density plot (cloud) and the boxplot (field) shows that the majority of ideophones are reasonably well dispersed.

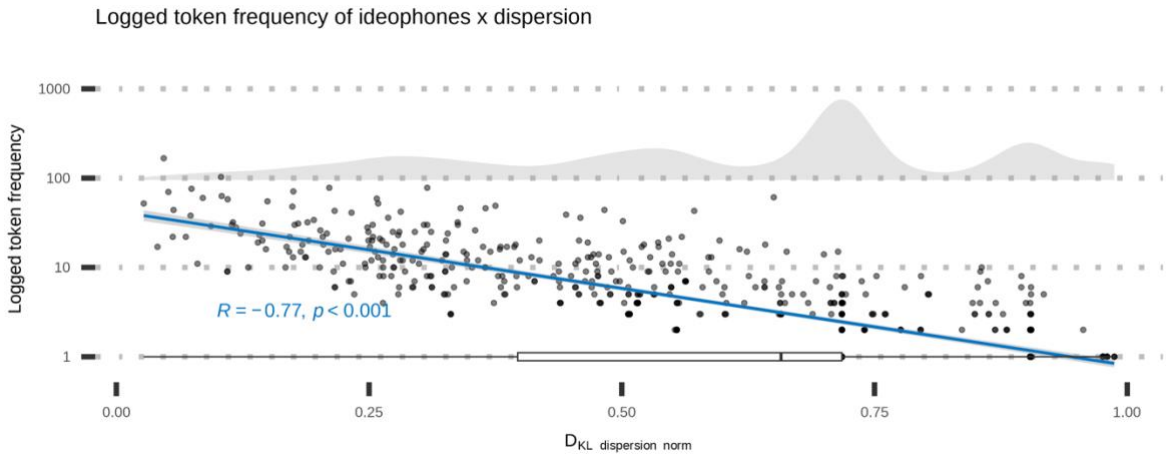

Figure 7. The dispersion rate of ideophones in the corpus is negatively correlated with (logged) frequency, i.e., highly frequent ideophones tend to occur only in one section of the corpus. At the same time, the boxplot and distribution show that most ideophone types have a relatively high dispersion rate but low frequency (Pearson correlation of $R = -.77, p < .001$).

6.3. Collocate-ideophone and ideophone-collocate

The fourth dimension of information that is of interest to this study is the preference combinations of collocate and ideophone have. Early on, it was already noted that some ABB expressions can also occur as BBA (T'sou 1978), but not all. We have seen examples like *xiāng-pēnpēn* 香噴噴 'fragrant' vs *pēnpēn-xiāng* 噴噴香 'fragrant' being mentioned in the literature. Given these two possible orders (collocate followed by ideophone and ideophone followed by collocate), it stands to reason to assume that each item has a preference. In the case of types that only occur in, say, collocate-ideophone, that preference will be very clear. But in the case of *xiāng* and *pēnpēn* that will be less straightforward. Once more we can turn to contingency tables and collostructional methods (Gries 2019) to address this problem. Because the corpus data only has *xiāng-pēnpēn*, we illustrate the method with *cuōtuō* 蹉跎 'dawdle, waste time' in

Table 4. After calculating all values for all ideophones, the association measure used here is the Fisher-Yates Exact test (Stefanowitsch and Gries 2003), a bidirectional test that ranks in this case how much an ideophone is relatively attracted to the collocate-ideophone construction. The distributions are shown in Figure 8.

Table 4. Example of contingency table for *cuōtuō* 蹉跎 'dawdle'

Ideophone	**Collocate-ideophone**	**Ideophone-collocate**	**Row totals**
Cuōtuō 蹉跎	(a) 3	(b) 2	(a+b) 5
¬ Cuōtuō 蹉跎	(c) 4666	(d) 963	(c + d) 5629
Column totals	(a + c) 4669	(b + d) 965	5634

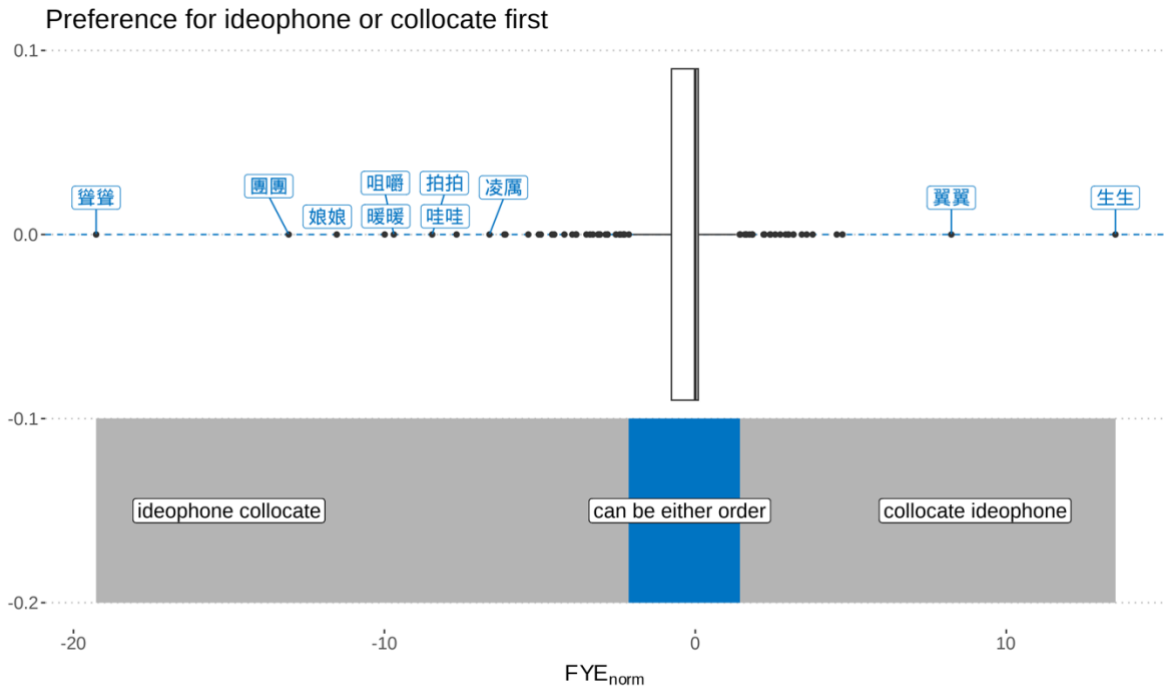

Figure 8. Preference for the order ideophone-collocate (left) or collocate-ideophone (right). Zero is the strict boundary between the two poles. However, the blue zone indicates a 'border zone'.

Figure 8 displays information regarding this constructional preference. First, a boxplot shows that most items prefer the collocate-ideophone order, as the median lies right of 0. The third quartile is also very close to it, but the first quartile is stretched relatively far away from 0. Major outliers for the collocate-ideophone order are *yìyì* from *xiǎoxīn-yìyì* 小心翼翼 'be careful'and *shēngshēng*, as in *huó-shēngshēng* 活生生 'lively'. On the other hand, there are a number of items that strongly prefer an ideophone-collocate order, the strongest outlier of which

is *sǒngsǒng-jiān* 聳聳肩 'shrugging shoulder'[13]. The blue zone in the ribbon, which ranges from the first quartile to the third, can be said to contain ideophones without strong preference, i.e., those that can occur in both orders. For instance, *cuōtuō* 'dawdle' has a value of -0.68.

6.4. A holistic picture

We have now calculated cue validity, frequency, dispersion, and preference. With these four single sources of information in hand, it is possible to present a holistic picture of constructions involving ideophones and collocates. The data is plotted with the plotly package (Plotly Technologies Inc. 2015). The html document in the supplementary materials (see Data Availability Statement) allows for the reader to interact with the plot. Here I present two stills from the same visualization, which contains the cue validity measures $D_{KL:ideophone \rightarrow collocate\ norm}$, $D_{KL:collocate \rightarrow ideophone}$, and logged token frequency of ideophones on the main axes. The color gradient shows the normalized dispersion $D_{KL:dispersion}$. Symbols indicate the preferred construction order: a square for collocate-ideophone, a diamond for ideophone-collocate, and a circle for moderate attraction to either. The two perspectives are presented in Figure 9 and Figure 10.

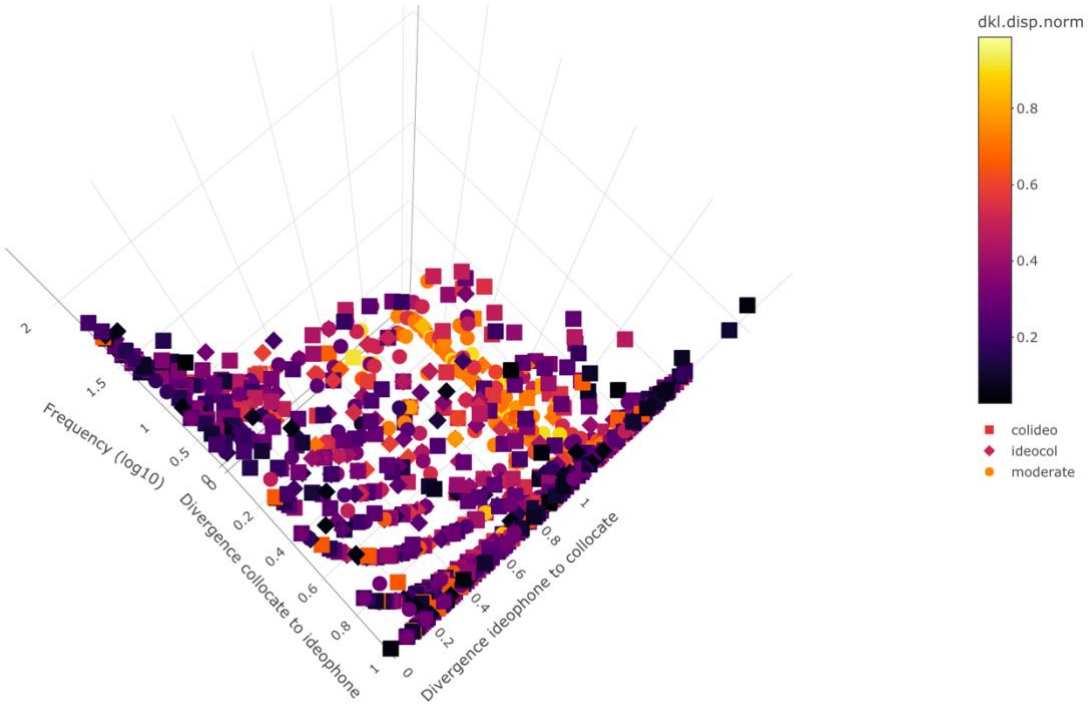

Figure 9. Still 1 of the 3D tuple combining of single dimensions of information

[13] It must be noted that sǒngsǒng-jiān can also be seen as a verbal diminutive (Sui 2018): briefly shrugging one's shoulders. However, because it still fits the cross-linguistic concept of an ideophone, it was kept in the data.

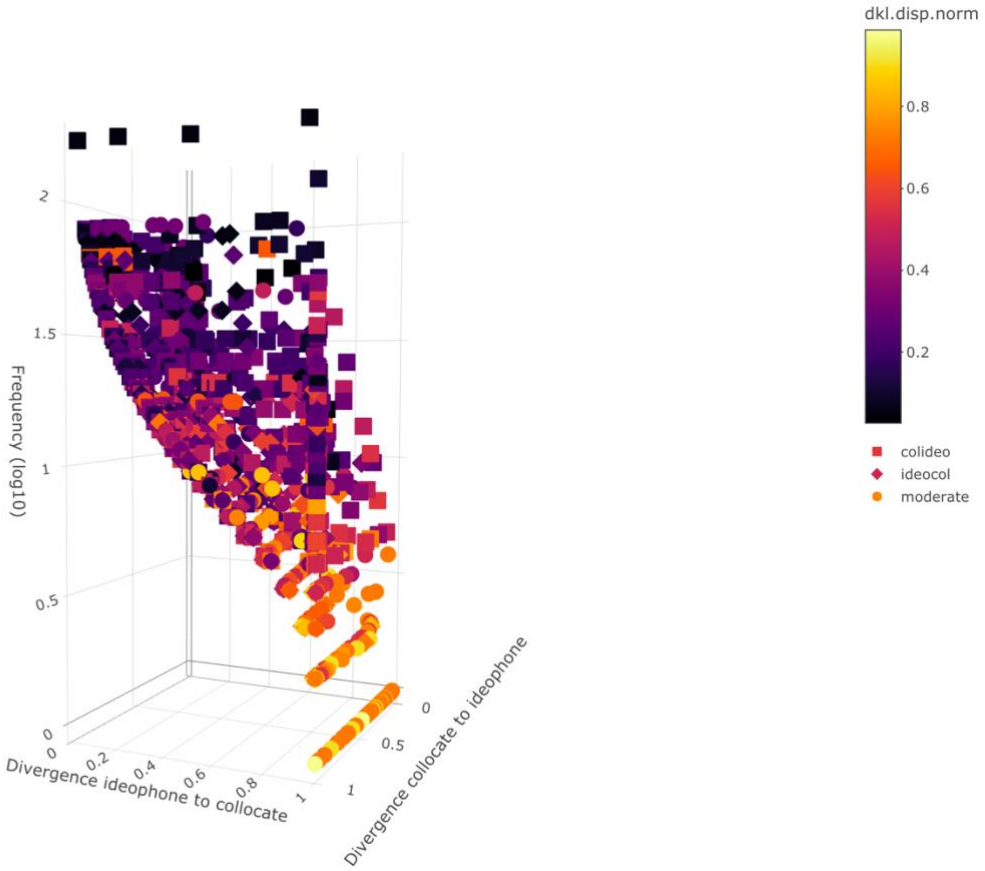

Figure 10. Still 2 of the 3D tuple combining of single dimensions of information

Figure 9 is reminiscent of the figures presented in Section 6. This tuple shows that the curved lines in those figures are somehow a frequency effect, which could be a vestige from a binary categorization conforming to "hapax" or "not hapax". In the code and data supplement, I have also generated plots marking hapaxes and omitting hapaxes from this figure. In any case, more follow-up work is needed to correctly interpret if the usual cognitive mechanisms are correlated with these frequency effects (see Bybee and Hopper 2001). We can also see, somewhat surprisingly, that items with the highest dispersion values appear to correlate with high values for $D_{KL:ideophone \rightarrow collocate\ norm}$. This can be related once again to the prototype status that ABB (X-AA) subschemas harbor in this dataset, since they mostly have high values for divergence with the ideophone as cue. From the completely different angle in Figure 10, it can be seen that ideophone token frequency follows an elegant curve, at least in the lower values. This surprising structure can be taken as a reminder that we should not solely base our understanding of ABB and related constructions on wordlist data alone, but also include observable corpus data.

7. Discussion and conclusion

I have started out with a survey of ABB-related literature (Section 2). First, the traditional derivational analysis that treats BB as an empty suffix was questioned. The potential involvement of other cognitive mechanisms notwithstanding (Section 3), we have seen that viewing ABB as a compound consisting of a prosaic collocate A and an ideophonic part BB can be quite fruitful. Based on a parallel line of research, we followed Mok (2001) in the notation of X-AA for fully reduplicated ideophones, but also incorporated partially reduplicated ones: X-AB. This led to the inclusion of longer prosaic parts (X, XY, XYZ, XYZW, …) that were combinable with disyllabic ideophones (AA, AB). Given that the order of collocate and ideophone has also been found as ideophone-collocate, the final schematic network that emerged consisted of 16 subschemas of ideophone-collocate constructions (Figure 2). In that network, it was assumed that the X-AA (ABB) lower-level schema would have a salient place, given all the attention received in the literature. To observe this, we turned to the Academia Sinica Balanced Corpus of Modern Chinese (Section 4) and obtained type and token frequencies of the schemas under investigation. It was found that X-AA is the most prominent trisyllabic schema. However, some other, tetrasyllabic, schemas were also found to be productive: XY-AA, XY-AB, AB-XY, AA-XY (Section 5).

Apart from an analysis on the level of the schema, we also looked at how exemplars fitted in with these subschemas (Section 6). By calculating four single sources of information (cue validity, token frequency, dispersion, and constructional preference), a tuple presentation (Gries 2019) containing all four dimensions was presented, that suggests that some of these variables hang together.

Two questions then remain: did it make sense to treat ABB as a prototypical schema sanctioned by a main schema involving collocates and ideophones? And what is the impact of doing so? The first question can be answered on two levels, by invoking a four-way characterization of prototype effects, which rests on effects of non-equality non-discreteness at the extensional or intensional levels (Geeraerts 2010: 189): (a) differences of typicality and membership salience, (b) clustering into family resemblances, (c) fuzziness at the edges and membership uncertainty, (d) absence of necessary and sufficient conditions. Given the domain of collocate-ideophone constructions, we do find prototype effect (a), with X-AA for instance being more salient than XYZW-AA; effect (b) with X-AA and X-AB being closer to each other in terms of number of syllables, but from another perspective X-AA and XY-AA in terms of reduplication type in the ideophonic part; and effect (c) to some degree with a question raised in Section 3 concerning the limit of syllables the ideophonic part can have. Effect (d) is not entirely present: the necessary and sufficient condition for ideophone-collocate constructions is

that there is a part that can be termed ideophonic and a part that can be seen as prosaic. The prototypicality effects observed on the level of schemas is thus similar to that of the category 'bird' (Geeraerts 2010: 191).

On the level of exemplars in relation to the subschemas, we can similarly observe effect (a) through the differences in dispersion value and token frequency; effect (b) with families that share the same collocate or the same ideophone, see examples (8-11); effect (c) with some instances like the AA-X (BBA) exemplar of *sǒngsǒng-jiān* 'shrug shoulders' (footnote 7) having somewhat of an ideophonic quality to it but not entirely. For effect (d), on the level of schemas conditions were clear, but on the level of exemplars it is actually not so easy what the necessary and sufficient conditions for ideophonehood and iconicity are (Dingemanse 2019; Dingemanse et al. 2020), especially in Chinese (Van Hoey 2020). All in all, we can then conclude that on the level of exemplars, all four prototypicality effects are present, reminiscent of the effects observed for the concept 'fruit' (Geeraerts 2010: 190).

Let us turn to the second question: what is the impact of treating ABB as a subschema of collocate-ideophone constructions? The main benefit is that this allows traditional studies on ABB to reconverge with the other line of research that has been developed parallel to it, as well as provide a place for this especially tight coupling between collocate and ideophone within the typological literature. That is, while collocates are often identifiable for ideophones in different languages (Nuckolls 2014), ideophones are argued to be mostly used in holophrastic utterances (Dingemanse and Akita 2016; Dingemanse 2017). The tight integration of well-known collocate-ideophone constructions shows that this is not necessarily the case in Mandarin and other Sinitic languages.

From the perspective of cognitive linguistics, this case study provides another piece of evidence that it is still useful to think of constructions in terms of prototypically structured assemblies, and that this can be observed with corpus data. This study has also attempted to take Gries's (2019) advice into account, by calculating first single sources of information, that can then be put together and lead to further exploration in future studies. It is important to note that this study has only been possible by taking seriously the notion that ideophones are words (Dingemanse 2019) first and foremost, with concomitant lexical effects that can be observed in the corpus in the first place. However, we should also be cognizant of whether these corpus-observable effects are truly present inside the heads of speakers. After all, linguistic categorization often distinguishes more fine-grained categories than many language users are aware of (Tuggy 1999; Divjak and Arppe 2013). To that end, it may be useful to supplement the corpus-based findings from this study with subjective ratings. One can expect correlations between dispersion, frequency on the one hand and familiarity ratings on the other. There may also exist correlations between topical genre of the corpus and valence, or arousal. Such

triangulation can provide converging evidence from context-rich data and context-free data, leading to a more holistic picture of how ideophones in ideophone-collocate constructions behave, in Mandarin, but also in other languages. In other words, I hope that this paper can provide a new beginning for ABB constructions as the prototype of collocate-ideophone constructions.

8. Acknowledgments

This paper is an expanded and updated version of a section in my dissertation (Van Hoey 2020). As such, I would like to thank Chiarung Lu and our lab members working on Cognitive and Functional linguistics at National Taiwan University for helping me to work out the general concept. I am also very grateful for the feedback garnered during my postdoctoral fellowship at the University of Hong Kong in the Language Development Lab, in particular to Youngah Do and Xiaoyu Yu for inspiring discussions on ABB and ideophonehood. Finally, a huge thanks to three anonymous reviewers who challenged a number of the assumptions made here, and who also helped me to improve the quality of the manuscript in a significant manner. The usual disclaimers apply.

9. Data availability statement

Supplementary data for this paper is available in the OSF repository at https://osf.io/2m58e/ . This repository holds the data, an R markdown file detailing how the different figures and numbers were obtained, as well as a folder with the figures presented in this paper.

10. References

Akita, Kimi & Mutsumi Imai. 2022. The iconicity ring model for sound symbolism. In Sara Lenninger, Olga Fischer, Christina Ljungberg & Elżbieta Tabakowska (eds.), *Iconicity in cognition and across semiotic systems* (Iconicity in Language and Literature 18), 27–46. Amsterdam: John Benjamins. https://doi.org/10.1075/ill.18.

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

Bodomo, Adams. 2006. The structure of ideophones in African and Asian languages: The case of Dagaare and Cantonese. In John Mugane, John P. Hutchison & Dee A. Worman (eds.), *Selected Proceedings of the 35th Annual Conference on African Linguistics*, 203–213. Somerville, MA: Cascadilla Proceedings Project.

Bybee, Joan L. 2001. *Phonology and language use* (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.

Bybee, Joan L. & Paul J. Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45). Amsterdam: Benjamins.

Cáo, Ruìfāng 曹瑞芳. 1995. Pǔtōnghuà ABB shì xíngróngcí de dìngliàng fēnxī 普通话 ABB 式形容词的定量分析 [quantitative analysis of Mandarin Chinese ABB adjectives]. *Yǔwén yánjiū* 语文研究 3. 22–25.

Chang, Yufen. 2009. The phonology of ABB reduplication in Taiwanese. In Yun Xiao (ed.), *Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21): Vol. 1*, 28–41. Smithfield, Rhode Island: Bryant University.

Chao, Yuen Ren. 1953. Popular Chinese plant words: A descriptive lexico-grammatical study. *Language* 29(3). 378–414.

CKIP group & Academia Sinica. 2013. Academia Sinica Balanced Corpus of Modern Chinese (ASBC 4.0) (Zhōngyāng yánjiūyuàn Xiàndài Hànyǔ biāojì yǔliàokù 中央研究院現代漢語標記語料庫). https://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm.

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

Dingemanse, Mark. 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, Mark. 2019. "Ideophone" as a comparative concept. In Kimi Akita & Prashant Pardeshi (eds.), *Ideophones, mimetics and expressives* (Iconicity in Language and Literature, ILL 16), 13–33. Amsterdam: John Benjamins. https://doi.org/10.1075/ill.16.02din.

Dingemanse, Mark & Kimi Akita. 2016. An inverse relation between expressiveness and grammatical integration: On the morphosyntactic typology of ideophones, with special reference to Japanese. *Journal of Linguistics* 1–32. https://doi.org/10.1017/S002222671600030X.

Dingemanse, Mark, Marcus Perlman & Pamela Perniss. 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.

Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). https://doi.org/10.1515/cog-2013-0008.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 188–221. https://doi.org/10.1075/arcl.7.08ell.

Féng, Jìng 冯静. 2016. Shījīng *zhōng de nǐhuìcí yánjiū* 《诗经》中的拟绘词研究 [Study of ideophones in the Book of Odes]. Harbin: Heilongjiang University Master thesis.
Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford Univ. Press.
Geeraerts, Dirk. 2017. Entrenchment as onomasiological salience. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (Language and the Human Lifespan), 153–174. Berlin: De Gruyter Mouton.
Gries, Stefan Th. 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, Stefan Th. 2019. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3). 385–412. https://doi.org/10.1075/ijcl.00011.gri.
Gries, Stefan Th. 2020. Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. Cham: Springer.
Gries, Stefan Th. 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, Stefan Th. & Philip Durrant. 2020. Analyzing co-occurrence data. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 141–159. Cham: Springer.
Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Covarying collexemes in the *into-causative*. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture and mind*, 225–236. Stanford, Calif: CSLI Publications.
Hsieh, Shu-Kai. 2006. *Hanzi, concept and computation: A preliminary survey of Chinese characters as a knowledge resource in NLP*. Tübingen: Universität Tübingen PhD dissertation.
Huang, Shi-Zhe, Jing Jin & Dingxu Shi. 2016. Adjectives and adjective phrases. In Chu-Ren Huang & Dingxu Shi (eds.), *A reference grammar of Chinese*, 276–296. Cambridge: Cambridge University Press.
Institute of Linguistics, Chinese Academy of Social Sciences (ed.). 2016. *Xiàndài Hànyǔ cídiǎn* 现代汉语词典 [Dictionary of Modern Chinese]. 7th edn. Beijing: Commercial Press.
Kelly, Michael H., J. Kathryn Bock & Frank C. Keil. 1986. Prototypicality in a linguistic context: Effects on sentence structure. *Journal of Memory and Language* 25. 59–74.
Kwok, Bit-Chee, Andy C. Chin & Benjamin K. Tsou. 2016. Grammatical diversity across the Yue dialects. *Journal of Chinese Linguistics* 44(1). 109–152. https://doi.org/10.1353/jcl.2016.0002.
Kwok, Bit-Chee 郭必之. 2012. Cóng Nánníng Yuèyǔ de zhuàngmàocí kàn Hànyǔ fāngyán yǔ mínzú yǔyán de jiēchù 从南宁粵语的状貌词看汉语方言与民族语言的接触 [Studying language contact between Chinese dialects and folk languages through by investigating ideophones in Nanning Yue]. *Mínzú yǔwén* 民族语文 3. 16–24.
Lai, Yik-Po 黎奕葆. 2015. Xiānggǎng yuèyǔ shuāng yīnjié zhuàngmào hòuzhuì de yīnyùn tèsè 香港粵語雙音節狀貌後綴的音韻特色 (Phonology of disyllabic ideophonic suffix in Hong Kong Cantonese). *Language and Linguistics* 16(5). 691–729.
Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar 1: Theoretical prerequisites*. Stanford, California: Stanford University Press.
Langacker, Ronald W. 2017. Entrenchment in Cognitive Grammar. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (Language and the Human Lifespan), 39–56. Washington, DC : Berlin: American Psychological Association ; De Gruyter Mouton.

Lǐ, Jìnróng 李劲荣. 2008. ABB xíngshì róng cí de gòuchéng fāngshì ABB 形式容词的构成方式 (The structure type of adjectives in the form ABB). *Journal of Gannan Normal University* 赣南师范学院学报 1. 87–91.

Lǐ, Liàng 李靓. 2019. *Xiàndài Hànyǔ ABB shì xíngróngcí yǔyì yánjiū* 现代汉语 ABB 式形容词语义研究 *(The semantic study of ABB adjectives in Contemporary Chinese)*. Jinan: Shandong University Master thesis.

Lǐ, Xíngjiàn 李行健. 2013. *Xiàndài Hànyǔ guīfàn cídiǎn* 现代汉语规范词典 [Standard dictionary of Modern Chinese]. 3rd edn. Foreign Language Teaching and Research Press.

Lijffijt, Jefrey & Stefan Th. Gries. 2012. Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics* 17(1). 147–149. https://doi.org/10.1075/ijcl.17.1.08lij.

Lǔ, Shūxiāng 吕叔湘 (ed.). 1980. *Xiàndài Hànyǔ bābǎi cí* 現代漢語八百詞 [800 words of Modern Chinese]. 5th edn. Beijing: Commercial Press.

Mok, Waiching Enid. 2001. *Chinese sound symbolism: A phonological perspective*. Hawai'i: University of Hawai'i PhD dissertation.

Nippold, Marilyn A. & Catherine L. Taylor. 2002. Judgments of idiom familiarity and transparency: A comparison of children and adolescents. *Journal of Speech, Language, and Hearing Research* 45(2). 384–391. https://doi.org/10.1044/1092-4388(2002/030).

Nuckolls, Janis B. 2014. Ideophones' challenges for typological linguistics: The case of Pastaza Quichua. *Pragmatics and Society* 5(3). 355–383. https://doi.org/10.1075/ps.5.3.03nuc.

Otting, Nadine Faye. 2019. *An analysis of AABB reduplicated plurals in Mandarin Chinese*. Hong Kong: The University of Hong Kong PhD dissertation.

Packard, Jerome Lee. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge: Cambridge University Press.

Paul, Waltraud. 2006. Zhu Dexi's two classes of adjectives revisited. In Christoph Anderl & Halvor Eifring (eds.), *Studies in Chinese language and culture: Festschrift in honour of Christoph Harbsmeier on the occasion of his 60th birthday*, 303–315. Oslo: Hermes Academic Publishing.

Plotly Technologies Inc. 2015. Collaborative data science. Montréal: Plotly Technologies Inc. https://plot.ly.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104(3). 192–233.

Schmid, Hans-Jörg & Franziska Günther. 2016. Toward a unified socio-cognitive framework for salience in language. *Frontiers in Psychology* 7. https://doi.org/10.3389/fpsyg.2016.01110.

Sew, Jyh Wee. 2008. From South Asian echo formation to Cantonese phonetic repetition. *The International Journal of Language, Society and Culture* 24. 72–83.

Shào, Jìngmǐng 邵敬敏. 1990. ABB shì xíngróngcí dòngtài yánjiū ABB 式形容词动态研究 [Dynamicity in ABB adjectives]. *Shìjiè Hànyǔ jiāoxué* 世界汉语教学 1(11). 19–26.

She, Shanshan 佘姗姗. 2015. *Chūgokugo ni okeru ABB-gata keiyōshi no onomatope-sei ni tsuite* 中国語における ABB 型形容詞のオノマトペ性について *[On Chinese ABB adjectives]*. Osaka: Osaka University Master thesis.

Smith, Jonathan. 2015. Sound Symbolism in the reduplicative vocabulary of the *Shijing*. *Journal of Chinese Literature and Culture* 2(2). 258–285. https://doi.org/10.1215/23290048-3324236.

Stefanowitsch, Anatol & Stefan Th. Gries. 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.

Sui, Yanyan. 2018. Affixation or compounding? Reduplication in Standard Chinese. In *Exact repetition in grammar and discourse* (Trends in Linguistics: Studies and Monographs 323), 127–157. Berlin: Mouton de Gruyter.

Sun, Jingtao. 1999. *Reduplication in Old Chinese*. Vancouver: University of British Columbia PhD dissertation.

Sūn, Wànmì 孙万蜜. 2012. Guānyú ABB shì xíngróngcí zhōng BB de dúyīn wèntí 关于 ABB 式形容词中 BB 的读音问题 (The pronunciation problems of the BB in the ABB construction adjectives). *Yǔyán yánjiū* 语言研究 5. 117–120.

Tabossi, Patrizia, Rachele Fanari & Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(2). 313–327. https://doi.org/10.1037/0278-7393.34.2.313.

Taylor, John R. 2002. *Cognitive grammar* (Oxford Textbooks in Linguistics). Oxford: Oxford University Press.

T'sou, B.K. 1974. Sound symbolism in Chinese: A preliminary study. Paper presented at VII International Conference on Sino-Tibetan Language and Linguistics, Georgia State University, Atlanta.

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.

Tuggy, David. 1992. The affix-stem distinction: A Cognitive Grammar analysis of data from Orizaba Nahuatl. *Cognitive Linguistics* 3(3). 237–300.

Tuggy, David. 1999. Linguistic evidence for polysemy in the mind: A response to William Croft and Dominiek Sandra. *Cognitive Linguistics* 10(4). 343–368.

Tummers, Jose, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2). 225–261. https://doi.org/10.1515/cllt.2005.1.2.225.

Van Hoey, Thomas. 2020. *Prototypicality and salience in Chinese ideophones: A cognitive and corpus linguistics approach* 中文擬聲(態)詞的原型與顯著特徵:以認知與語料庫語言學方法探討. Taipei: National Taiwan University PhD dissertation.

Van Hoey, Thomas & Chiarung Lu. 2019. Lexical variation of ideophones in Chinese classics: Their implications in embodiment and migration. In Janice Fon (ed.), *Dimensions of diffusion and diversity* (Cognitive Linguistics Research 63), 195–226. Berlin; Boston: De Gruyter Mouton. 10.1515/9783110610895-008.

Van Hoey, Thomas & Arthur Lewis Thompson. 2020. The Chinese Ideophone Database (CHIDEOD). *Cahiers de linguistique Asie orientale* 49(2). 136–167. https://doi.org/10.1163/19606028-bja10006.

Wáng, Tīngjiàn 王烃鉴. 2020. *Xiàndài Hànyǔ ABB shì zhuàngtài xíngróngcí rèn zhī yánjiū* 现代汉语 ABB 式状态形容词认知研究 (A cognition study of ABB-type state adjectives in Mandarin Chinese). Nanchong: China West Normal University Master thesis.

Wang, Zhijun. 2010. The head of the Chinese adjectives and ABB reduplication. In Lauren Ebyd Clemens & Chi-Ming Louis Liu (eds.), *Proceedings of the 22nd North American Conference on Chinese Linguistics (NACCL-22) & the 18th International Conference on Chinese Linguistics (IACL-18): Vol. 1*, 232–245. Cambridge (Mass.): Harvard University Press.

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

Yáo, Zhànlóng 姚占龙. 2006. Tónggēn ABB shì zhuàngtài xíngróngcí jí qí liàng jí kǎochá 同根 ABB 式状态形容词及其量级考察 (An investigation of statal adjectives of the ABB pattern and their gradability). *Shìjiè Hànyǔ jiàoxué* 世界汉语教学 3. 72–79.

Zhāng, Dān 张丹. 2005. Hànyǔ zhōng ABB xíng zhuàngtài xíngróngcí de gòuchéng fènxī 汉语中 ABB 型状态形容词的构成分析 [An analysis of the formation of ABB type state adjectives in Mandarin Chinese]. *Journal of Shenyang Agricultural University (Social Sciences)* 沈阳农业大学学报(社会科学版) 7(1). 124–126.

Zhào, Qīngqīng 赵青青. 2021. Tōnggǎn yǐnyù shìjiǎo de Xiàndài Hànyǔ ABB shì zhuàngtài xíngróngcí 通感隐喻视角的现代汉语 ABB 式状态形容词 (ABB-pattern state adjectives in Mandarin: A study from the perspective of synaesthetic metaphor). *Shìjiè Hànyǔ jiāoxué* 世界汉语教学 35(2). 206–219. https://doi.org/10.13724/j.cnki.ctiw.2021.02.005.

Zhèng, Huáidé 郑怀德 & Qìnghǎi 孟庆海 Mèng. 2003. *Hànyǔ xíngróngcí yòngfǎ cídiǎn* 汉语形容词用法词典 *[Dictionary of adjective usage in Chinese]*. Shanghai: Commercial Press.