

ORIGINAL ARTICLE

Open Access



# NP weight effects in word order variation in Mandarin Chinese

Yao Yao

Correspondence:

[ctyaoyao@polyu.edu.hk](mailto:ctyaoyao@polyu.edu.hk)

Department of Chinese and  
Bilingual Studies, The Hong Kong  
Polytechnic University, 11 Yuk Choi  
Rd, Hung Hom, Hong Kong

## Abstract

**Background:** In the literature of sentence production, both short-before-long and long-before-short word order tendencies have been observed across languages. Specifically, SVO languages such as English show the short-before-long noun phrase (NP) shift, placing heavy NPs near the end of the sentence; on the other hand, verb-final languages such as Japanese and Korean show the long-before-short NP shift, placing heavy NPs earlier in the sentence. In this paper, we examine the effects of NP weight on word order variation in Mandarin Chinese, which not only has a predominantly SVO word order but also allows a grammaticalized SOV construction (i.e., the *ba* construction).

**Methods:** We conducted a corpus analysis with two verb-specific datasets extracted from the 10 million-word Academia Sinica Balanced Corpus of Modern Chinese (Version 5.0; Chen et al., Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation, 1996). Each dataset contained more than 900 sentences, in either SVO or *ba* construction and possible to be converted to the other word order without changing sentence meaning. Generalized mixed-effects models were built to examine the effect of NP weight on the surface word order (SVO vs. *ba*), while controlling for other factors that are also known to influence the SVO-*ba* alternation (e.g., verb complement, animacy, and givenness of the object NP, sentence structure, and structural parallelism in the context). The accuracy of the modeling results was inspected by comparing the word order predictions made by the models with both actual word orders observed in the corpus and naturalness ratings of alternative word orders by native speakers in behavioral experiments.

**Results:** Our results show a U-shaped NP weight effect on SVO-*ba* alternation, in that both very short and very long NPs are more likely to be shifted to preverbal positions than NPs with medium weight. These results provide evidence that both conceptual and positional factors are operating in the preverbal domain in Mandarin.

**Conclusion:** Taken together with previous findings of positional factors operating in the postverbal domain in Mandarin, our results suggest that the relative sensitivity to conceptual and positional factors can vary within a language. We discuss findings in the framework of the sentence production model.

**Keywords:** Word order variation Mandarin Chinese, *ba* construction, Heavy NP shift

## 1 Introduction

### 1.1 NP weight and word order variation

Sentences are formed to convey messages. While the sequencing of words in a sentence can significantly change the content of the encoded message (e.g., John hit Mary vs. Mary hit John), language also provides a certain degree of flexibility that allows multiple

sequences to convey the same meaning—at least truth-conditionally (e.g., John hit Mary and Mary was hit by John, John sent Mary a gift and John sent a gift to Mary). Such flexibility has been widely observed, but the inventory of sequencing options differs across languages. Furthermore, for a given meaning and a given language, some sequences may be more favored than others (e.g., John sent it to Mary sounds better than John sent Mary it in English). While many factors may influence the choice of surface constituent order, we focus on one of the most widely studied factors, the weight of noun phrases (NP), measured by NP length in terms of the number of words throughout this paper.

Probably the most well-studied NP weight effect is the short-before-long tendency in English word order variation (Arnold et al. 2000; Bresnan et al. 2007; Stallings and MacDonald 2011; Stallings et al. 1998). The example in (1) shows two possible word orders of an English sentence with a verb phrase (VP) that contains two prepositional phrases (PP), one with a short NP (i.e., PP1) and the other a long NP that contains a relative clause (i.e., PP2). Probably no one would disagree that the sentence in (1a), which places the shorter PP1 before the longer PP2, sounds better than (1b), which places the longer PP2 before the shorter PP1.

- (1) a. John<sub>VP</sub>[gave<sub>PP1</sub>[in the garden] PP2[for the keys he recently borrowed from his friend Joe]].  
 b. John<sub>VP</sub>[looked<sub>PP2</sub>[for the keys he recently borrowed from his friend Joe] PP1[in the garden]].

Two prominent accounts have been proposed to explain the short-before-long tendency in English word order: a production-oriented account and a comprehension-oriented account. The production-oriented account attributes sequencing preferences to the properties of the production system. One of the most important properties is the accessibility (or availability) of individual words and phrases (Bock 1986; McDonald et al. 1993). Broadly speaking, accessibility describes the ease of accessing certain linguistic materials (words or phrases) by the speaker, or in other words, how “ready” the linguistic materials are to be used in production. The central thesis of the production-oriented account is that words and phrases that are more accessible to the speaker tend to occur earlier in the sentence. The rationale of this account is aligned with the general understanding that sentence production is incremental (Bock 1982; De Smedt 1994). Instead of waiting for the whole sentence to be planned out before starting the articulation, the speaker is engaged in sentence planning and articulation at the same time. As a result, the sentence begins to roll out—piece by piece—before the rest of the sentence is completely planned. Thus, it is natural that the pieces that are easier to retrieve and assemble are delivered first.

The notion of accessibility can be operationalized by different measures. For example, Bock (1986) showed that words and phrases that have recently been mentioned in the preceding context are easier to access than those that are new to the discourse. There is also a close relationship between accessibility and weight. As a heavy phrase contains more lexical items, a longer linear sequence, and (often) higher syntactic complexity than a lighter phrase; it seems natural that a heavy phrase requires greater effort in planning and preparation in the production process and hence should occur later than a lighter phrase. (However, as will be discussed below, weight and accessibility may also be associated in a different fashion.)

The comprehension-oriented account is promoted by considerations mainly from the comprehension side, regarding the efficiency of processing, or more specifically, parsing. Hawkins (2004, 2014) proposed a set of parsing-based principles of form variation, among which the most important is the principle to *minimize domain* (MiD), defined as follows:

“The human processor prefers to minimize the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed. The degree of this preference is proportional to the number of relations whose domains can be minimized in competing sequences or structures, and to the extent of the minimization differences in each domain.” (Hawkins 2004:31)

The essence of the MiD principle is to minimize the linear sequence of words that must be processed in order to construct relations of combination or dependency among constituents within a mother phrase. The relevant linear sequence of words—i.e., the *domain*—is also referred to as the *phrasal combination domain* (PCD). (The predecessors of MiD and PCD are the *early immediate constituent* (EIC) principle and the *constituent recognition domain* (CRD), respectively, both defined and discussed in Hawkins (1994).) Thus, the PCD of a VP should include at least its head (V), the heads of all the daughter constituents (e.g., NPs and PPs), and all the words in between. Using (1a) and (1b) as examples again (repeated below, with PCDs shown by dashed lines under the sentences), the PCD in (1a) has 5 words (looked ... for), while the PCD in (1b) has 12 words (looked ... in). In both cases, the PCD covers the head *looked*, the heads of both PP1 (*in*) and PP2 (*for*), and all the words in between. Thus, the MiD principle correctly predicts that (1a) is preferred over (1b), because the PCD of the VP is shorter in (1a). More generally, in a head-initial language like English, when the head of a phrase may be followed by multiple constituents, the MiD principle will always prefer the shorter constituent to be adjacent to the head or at least closer to the head than the longer constituent is, resulting in the short-before-long word order tendency.

(1). a. John<sub>VP</sub>[looked<sub>PP1</sub>[in the garden] PP2[for the keys he recently borrowed from his friend Joe]].

1            2 3 4            5.

-----.

b. John<sub>VP</sub>[looked<sub>PP2</sub>[for the keys he recently borrowed from his friend Joe] PP1[in the garden]].

1            2 3 4 5 6            7            8 9 10 11 12.

-----.

What is the psychological basis of the MiD principle? What the speaker can gain from MiD is dubious (see Hawkins 2004 for more discussion), but it does seem to provide some benefits for the listener. The MiD principle promotes the minimization of the linear sequence that the listener need to keep track of in order to recognize all the constituents within a phrase; as a result, it reduces the chance of long-distance syntactic relations that may cause structural confusion and increase working memory demands.<sup>1</sup>

As discussed above, both the accessibility-based account and MiD principle can account for the short-before-long preference in English. The fact that they produce confluent predictions is anything but surprising, as the production-oriented and comprehension-oriented accounts are often found to motivate the same phenomena in language production (e.g., Arnold 2008; Gahl and Johnson 2012). Nevertheless, if the scope of investigation is extended to word order variation in other languages, the predictions of the two accounts become more distinguishable. Crucially, the accessibility-based account—as has been described so far—would predict a universal short-before-long preference, as the relative ease of composing a short phrase is independent of the grammatical configuration of the language. By contrast, the MiD principle makes different predictions for languages with different headedness. In a head-initial language like English, the MiD principle favors a short-before-long sequence for phrases in the postverbal VP domain. Conversely, in a verb-final language like Japanese and Korean, the MiD predicts the opposite, long-before-short preference for phrases in the preverbal domain. An example of Japanese word order variation is shown in (2), with (2a) exemplifying a [PP NP V] internal structure of VP and (2b) a [NP PP V] structure. In this example, since PP is shorter than NP by one word, the PCD of VP in (2b) is slightly shorter (“o ... katta,” four words) than that of (2a) (“kara ... katta,” five words). The MiD principle predicts a slight preference for the word order in (2b), which puts the longer phrase (NP) before the shorter one (PP) in the linear sequence. If the length difference between the two phrases is larger, the preference will be stronger. Both corpus analyses and online sentence production experiments have confirmed the long-before-short preference in the preverbal domain in Japanese and Korean (Choi 2007; Hawkins 1994, 2004; Yamashita and Chang 2001).

- (2). a. Tanaka ga<sub>VP</sub>[<sub>PP</sub>[Hanako kara]<sub>NP</sub>[sono hon o] katta].  
 Tanaka \_NOM\_Hanako\_from\_that\_book\_ACC\_bought.  
*Tanako bought that book from Hanako.*  
 1        2        3        4        5.  
 -----.
- b. Tanaka ga<sub>VP</sub>[<sub>NP</sub>[sono hon o]<sub>PP</sub>[Hanako kara] katta].  
 Tanaka \_NOM\_that\_book\_ACC\_Hanako\_from\_bought.  
 Tanako bought that book from Hanako.  
 1        2                    3        4  
 -----
- (Example from Hawkins 2004:109)

To summarize, assuming that verb arguments occur in the postverbal domain in head-initial languages and in the preverbal domain in head-final languages, the MiD principle predicts an overall preference for a linear sequence that places the shorter constituents closer to the head and the longer constituents further away. This preference surfaces as a short-before-long pattern in head-initial languages and an opposite, long-before-short pattern in head-final languages.

While the MiD principle seems to be more successful in explaining cross-language differences in weight effects, questions remain as to how comprehension-oriented

factors can condition word order preferences in online sentence production. Both short-before-long and long-before-short patterns have been observed in online sentence production tasks, in English (Stallings et al. 1998) and Japanese (Yamashita and Chang 2001) respectively. The fact that these production tasks did not involve any role of listener—either real or imaginary—and offered no incentive for listener accommodation undermines a comprehension-oriented account of the results.

Meanwhile, can the production-oriented account accommodate cross-linguistic variations in the direction of heavy NP shift at all? Efforts have been made to revise the account in order to reconcile the English and Japanese data. Specifically, Chang and colleagues (Chang 2009; Yamashita and Chang 2001) distinguished two types of accessibility. In addition to the ease of constructing the form of a phrase, or “form accessibility,” which negatively correlates with phrase length as we have discussed before, Chang and colleagues also included “conceptual accessibility” in the consideration of word order preferences. Conceptual accessibility refers to the accessibility of relevant concepts associated with the phrase to be produced. Chang and colleagues argue that by virtue of having more lexical items modifying the head noun, a longer NP is semantically richer and more salient than a shorter one, which in turn “increases the overall accessibility of the phrase in the conceptual arena” (Yamashita and Chang 2001:B53). In other words, a longer phrase has both lower form accessibility and higher conceptual accessibility, compared to a shorter phrase. In this theory, the difference between Japanese and English is attributed to language-specific levels of relative sensitivity to the two types of accessibility: Japanese word order variation is more sensitive to meaning (hence conceptual accessibility), while English word order variation is more sensitive to form (hence form accessibility).

The underlying reason for the cross-linguistic difference proposed by Chang and colleagues may be related to the difference between the preverbal domain and the postverbal domain. A critical difference between the two languages is that word order variation takes place in the preverbal domain in Japanese but in the postverbal domain in English. Since the preverbal area is closer to the beginning of the sentence, where phrasal order is in general more sensitive to semantic and pragmatic factors such as topic status, animacy, and concreteness (McDonald et al. 1993), it is possible that word order variation in the preverbal domain—as it occurs in Japanese—is more sensitive to meaning than form.

Taken together, the comprehension-oriented account correctly predicts both the short-before-long pattern in English and the long-before-short pattern in Japanese, but the underlying mechanism that links the preference with language production is unclear. On the other hand, the production-oriented account can also explain the different weight effects in English and Japanese by considering two types of accessibility and the distinction between preverbal and postverbal domains.

Regarding the preverbal-postverbal distinction, we notice that previous investigation of constituent ordering has mostly focused on the variation within the preverbal or postverbal domain. What is less studied is the shift of a phrase from one domain to the other, i.e., across the verb. Why is a cross-domain phrasal shift important to study? For one thing, a within-domain phrasal shift concerns the relative order of (typically) two constituent phrases on the same side of the verb; thus, the weight difference of the two phrases is taken as the critical measure of the amount of weight effect in both

comprehension-oriented and production-oriented accounts. A cross-domain phrasal shift, however, concerns the relative order of one constituent phrase and the verb (i.e., whether the constituent phrase is before or after the verb). Assuming that the length of the verb is fixed, the weight effect—if any—has to be related to the absolute weight of the participating constituent phrase. Will we still observe weight effects, based on the weight of a single phrase, in the same directions as predicted by the theoretical accounts? Furthermore, with regard to the revised production-oriented account specifically, if constituent ordering is sensitive to different forces in the preverbal domain and the postverbal domain, how would the different forces interact if the language allows a phrase to be shifted from one domain to the other?

A potential example of cross-domain phrase shift has been discussed for Hungarian in the previous literature (Kiss 1981, 1987, 1998; see also Hawkins 1994). A canonical verb-initial language, Hungarian also allows the subject NP, the object NP, or both to be shifted to the preverbal domain to topic and focus positions. The sentence structure of a Hungarian sentence is shown in (3) (3a) is adapted from (4.3) in Hawkins 1994:130; (3b–d) from Table 1 in Kiss 1981).

- (3). a. [(TopicP) [(FocusP) [V NP1 NP2]<sub>V'</sub>]<sub>VP</sub>]<sub>S</sub>.  
 b. 'Szereti János Marit.  
 love<sub>V</sub> John<sub>NP1</sub> Mary<sub>NP2</sub>.  
*John loves Mary.*  
 c. <sub>FocusP</sub>['János] szereti Marit.  
 John<sub>NP1</sub> love<sub>V</sub> Mary<sub>NP2</sub>.  
*It is John who loves Mary.*  
 d. <sub>TopicP</sub>[János] <sub>FocusP</sub>['Marit] szereti .  
 John<sub>NP1</sub> Mary<sub>NP2</sub> love<sub>V</sub>.  
*As for John, it is Mary who he loves.*

(3b–d) are examples of Hungarian sentences when both NPs are in the postverbal domain ((3b)), when one NP is in the preverbal domain and the other in the postverbal domain ((3c)), and when both NPs are in the preverbal domain ((3d)). For sentences like (3b), Hawkins found that a short-before-long preference in the postverbal domain, with the shorter NP more likely to occur in front of the longer NP. But when one or both NPs are moved to the preverbal domain, “there were now many instances of longer NPs preceding shorter ones” (Hawkins 1994:131). We are not sure whether this pattern is specific to sentences with TopicP or FocusP or both. We take this to mean that in general, the short-before-long pattern observed in the postverbal domain in sentences like (3b) is at least reduced in sentences like (3c) or (3d) or both, although it is not clear to us whether a reverse, long-before-short pattern, is observed in sentences in the latter category. It is also not clear whether there is an effect of absolute NP weight

**Table 1** Number of sentences in each dataset

Verb	Number of sentences	Number of <i>ba</i> sentences	Number of SVO sentences
放 <i>fàng</i>	947	682	265
拿 <i>ná</i>	988	261	727

in the Hungarian data, in the direction that a longer NP will be more likely to be shifted from the postverbal domain to the preverbal domain than a shorter NP.

The above being said, Hawkins' explanation for the (potential) long-before-short tendency in Hungarian data is that NP shift involving the topic/focus positions in the preverbal domain in Hungarian is conditioned not by parsing efficiency but by the semantic-pragmatic status of the relevant NP(s). It should be noted that this account is in line with Chang et al.'s proposal that preverbal word order variation in Japanese is sensitive to meaning.

To the best of our knowledge, there has not been any detailed analysis of cross-domain NP shift apart from the brief analysis of Hungarian data. If a language allows word order variation both within a domain (among constituent phrases on one side of the verb) and across domains (phrasal shift across the verb), do the two types of variation follow the same or different principles? Can the relative sensitivity to meaning/form accessibility vary within a language as a property of the domain or is it a strictly language-specific property as a result of headedness? If the preverbal domain is indeed more sensitive to conceptual accessibility, does that mean that form sensitivity does not operate in the preverbal domain at all?

In this paper, we examine word order variation in Mandarin Chinese (hereafter "Mandarin" for short). Specifically, we focus on the alternation between SVO and SOV word orders in the language, and whether (and how) it is conditioned by NP weight. The reason we choose to investigate Mandarin is twofold. First and foremost, same as Hungarian, Mandarin allows word order variation both in the postverbal domain and across the preverbal-postverbal domains. While typically a SVO language (Sun and Givón 1985), Mandarin allows a grammaticalized *ba* construction that puts object NP before the verb, resulting in an optional SOV word order. Second, Mandarin presents an unusual combination of (typically) head-initial VPs and head-final NPs, which is rarely observed across languages (Matthews and Yeung 2000). Importantly, this property distinguishes Mandarin from both Japanese (with head-final VP and head-final NP) and English (with head-initial VP and head-initial NP) in a critical aspect (i.e., headedness) for both theoretical accounts. Given the mixed headedness of Mandarin, what will production-oriented and comprehension-oriented accounts predict for NP weight effects in word order variation?

In the current study, we modeled the alternation between SVO and the *ba* construction in the corpus-based data sets of two verbs, 放 *fàng* "to put" and 拿 *ná* "to take in one's hands." To preview the results, our analysis shows a significant non-linear effect of NP weight on object preposing. Object NPs on both ends of the weight scale (i.e., both very short NPs and very long NPs) are more likely to be preposed to the preverbal position (resulting in the *ba* construction) than NPs with medium weight. The results provide evidence that the SVO-*ba* alternation in Mandarin Chinese is conditioned by both conceptual salience and form accessibility. We discuss the modeling results in lieu of previous findings of postverbal word order variation in Mandarin and in the framework of the sentence production model.

In the remaining of the paper, we will first briefly describe the *ba* construction and the research on word order variation involving the *ba* construction; we will then give a short review of the existing literature on corpus-based statistical models of word order

variation; modeling methods and results of the 放 *fàng* and 拿 *ná* models will be discussed in detail, followed by a discussion of the implications for theories of sentence production.

### 1.2 *Ba* construction and Mandarin word order variation

Mandarin is a predominantly SVO language (Sun and Givón 1985), but the language is also equipped with a grammaticalized *ba* construction, in which an object NP is preposed to a preverbal position. The *ba* construction acquires its name from the grammatical marker 把 *bǎ*, which is used immediately before the preposed object NP in this construction (see (4a) for an example). Apart from 把 *bǎ*, it is also possible to use the lexical item 将 *jiāng* to mark a preposed object NP, especially in the formal registry (see (4b) for an example).

- (4). a. 他把那本書放下了。  
 tā\_bǎ\_nà\_bě\_nshū\_fàng\_xià\_le.  
 he\_BA\_that\_CL\_book\_put\_down\_ASP.  
*He put down that book.*
- b. 他將那本書放下了。  
 tā\_jiāng\_nà\_bě\_nshū\_fàng\_xià\_le.  
 he\_BA\_that\_CL\_book\_put\_down\_ASP.  
*He put down that book.*
- c. 他放下了那本書。  
 tā\_fàng\_xià\_le\_nà\_bě\_nshū.  
 he\_put\_down\_ASP\_that\_CL\_book.  
*He put down that book.*

Sentences (4a) and (4b), both of which are *ba* constructions with a preposed object NP (那本書 *nàběnshū*), have identical meanings, with the only difference being that (4b) may sound more formal than (4a). Sentence (4c), on the other hand, has the canonical SVO order, and is generally considered to be equivalent to both (4a) and (4b) in meaning.

Although the *ba* construction does not occur as often as the canonical SVO word order, it is one of the most well-studied topics in the literature of Chinese linguistics, and always occupies a special position in Chinese language teaching and learning. A voluminous body of research has been devoted to the structural and semantic properties of the *ba* construction. It is now widely acknowledged that at least two conditions must be met for a *ba* construction to be grammatical (Li and Thompson 1974, 1981, Li and Thompson 1975; Xu 1995; etc.): (i) the object NP is definite (or specific, generic), and (ii) the verb phrase conveys a disposal meaning and usually contains a verb complement or aspect marker. For these reasons, the *ba* construction has also been associated with notions of topic and topicality (e.g., Chu 屈承熹 1979; Mei 梅廣 1978; Sun and Givón 1985; Tsao 1987) and verb transitivity (Hopper and Thompson 1980; Liu 1999; Sun 1995; Thompson 1973). Specifically, Tsao (1987) proposed that in a *ba* construction, the initial NP is the primary topic, and the object NP following *ba* (hereafter, the “*ba* NP”) is a secondary topic. In addition to the fact that the *ba*



NP is always definite/generic/specific, Tsao listed a few other arguments for the status of the *ba* NP, including the possibility of inserting a (filled) pause after the *ba* NP (see (5a) below) and the common use of the *ba* NP as the head of a topic chain in the subsequent context (see (5b) below).

(5). a. 我把那本書(啊)賣給小明了。

wǒ\_bǎ\_nà\_běn\_shū\_(a)\_mài\_gěi\_xiǎo-míng\_le

I\_BA\_that\_CL\_book\_(filled pause)\_sell\_to\_Xiaoming\_ASP

*I sold that book to Xiaoming.*

(from example (18) in Tsao 1987)

b. 他把房子整修了一下,漆了漆,然後再賣出去。

tā\_bǎ\_fángzi\_zhěngxiū\_le\_yíxià,qī\_le\_qī,ránhòu\_zài\_mài\_chūqù

he\_BA\_house\_repair\_ASP\_a little, paint\_ASP\_paint, afterward\_then\_sell\_out

*He had the house repaired, painted, and then sold.*

(from example (22) in Tsao 1987)

The analysis of the *ba* NP as a secondary topic is not completely agreed upon (e.g., see LaPolla 1990), but it is consistent with the widely assumed tendency in Mandarin sentences, with given, topic information occurring near the beginning of the sentence in the preverbal domain and new, focus information occurring near the end of the sentence in the postverbal domain (LaPolla 1990, 1995, 2009; Li and Thompson 1981; Xu 2004).

While NP definiteness and the disposal meaning of the verb may be necessary for the use of the *ba* construction, they are by no means sufficient conditions. As shown in (4), when both conditions are met (*nà* “that” ensures the definiteness of the NP, and the verb complement *xia4* and aspect marker *le* ensure the disposal meaning of the verb), both the *ba* construction and the SVO word order are grammatical. Sun and Givón (1985) also found that even when both NP definiteness and verb disposal meaning are met, the likelihood of observing a *ba* construction (as opposed to SVO) is still slim in a corpus.

Thus, the question is, when both *ba* and SVO are possible, what conditions the surface word order? If one believes that the SVO-*ba* alternation is ultimately a reflection of the change in the information status of the NP, then this is equivalent to asking what conditions the topic/focus status of the object NP. Liu (2007) set out to tackle this question by examining a sample of 456 SVO and SOV sentences which could in theory be converted to the other word order (i.e., “structurally interchangeable”). It should be noted that Liu’s dataset includes not only explicit *ba* sentences but also SOV sentences without explicit *ba* marker. The unmarked SOV sentences comprise less than 16% of the SOV set. Liu’s analysis showed that overall, discourse old (“given”) object NPs are more likely to be preposed than discourse new NPs. Nevertheless, the trend of preposing given NPs only holds for light and medium-weight NPs. If the object NP is heavy, the effect of information status is reversed, with heavy + new object NPs being more likely to be preposed and heavy + old object NPs being more likely to be postverbal. Based on these results, Liu proposed that there is an interaction of givenness and weight that conditions the SVO-SOV alternation.

Several caveats should be noted about Liu’s study. Apart from givenness and weight, Liu’s study did not control for other factors that are known to affect surface word order,

such as the semantic features of the NP and the use of parallel structures in the discourse. Most importantly, if the use of BA construction is related to verb transitivity, it is possible that individual verbs have inherent biases toward or against a certain surface word order. Liu's dataset mixed sentences of different verbs, with insufficient information about the distribution of each verb type, which makes it hard to evaluate whether the observed pattern holds across verbs or is only driven by certain verbs. We also notice that Liu's analysis divided object NPs into discrete weight categories (light, medium, heavy), and the critical evidence for the givenness  $\times$  weight interaction comes from the heavy category, of which there are only a couple dozen tokens in the dataset (~ 5%). Last but not the least, Liu did not provide a unified account for the observed pattern. If there is indeed an interaction of givenness and weight, why should they operate in tandem, and why should they both affect the alternation between SVO and SOV?

In a more recent study, Yao and Liu (2010) looked at a related but slightly different phenomenon: the word order variation in Mandarin ditransitive sentences. Similar to English, Mandarin allows variable orderings of direct object (DO) and indirect object (IO) in the postverbal domain (see (4a–b)). Moreover, Mandarin also allows the DO to be preposed before the verb, as in a monotransitive *ba* construction (see (4c)). Yao and Liu found NP weight effects in both postverbal word order variation (i.e., (4a) vs. (4b)) and postverbal-to-preverbal preposing of DO (i.e., (6a–b) vs. (6c)), but the direction of the effect differs between the two domains. Specifically, a longer DO (relative to the IO) has an increased likelihood to be preposed before the verb; meanwhile, if DO and IO are both postverbal, a longer DO (relative to the IO) is more likely to be shifted to the end. In other words, a two-way heavy NP shift was observed. Since the authors only investigated ditransitive sentences, it is unclear whether the observed NP weight effect on SVO-*ba* alternation can be extended to monotransitive sentences.

- (6). a. 小王送<sub>DO</sub>[那本書]給<sub>IO</sub>[妹妹]。  
 xiǎowáng sòng nà běn shū gěi mèimèi  
 Xiaowang give\_that\_CL\_book\_to\_sister  
 Xiaowang gave that book to (his) sister.
- b. 小王送給<sub>IO</sub>[妹妹]<sub>DO</sub>[那本書]。  
 xiǎowáng sòng gěi mèimèi nà běn shū  
 Xiaowang give\_to\_sister\_that\_CL\_book  
 Xiaowang gave (his) sister that book.
- c. 小王把<sub>DO</sub>[那本書]送給<sub>IO</sub>[妹妹]  
 xiǎowáng bǎ nà běn shū sòng gěi mèimèi  
 Xiaowang BA\_that\_CL\_book\_give\_to\_sister  
 Xiaowang BA that book gave to (his) sister.

Taken together, Liu (2007) and Yao and Liu (2010) suggest complex patterns of how weight might condition the ordering of NP constituents in a Mandarin sentence. More specifically, evidence has been found for preposing both short and long NPs in the postverbal-to-preverbal shift.

One remaining question is the difference between *ba* constructions and SOV sentences with unmarked preverbal object NPs (hereafter “unmarked SOV”). As

mentioned above, both *ba* constructions and unmarked SOV sentences were included in Liu's dataset of SOV sentences. However, in this study, we focus on the variation between *ba* and SVO sentences, excluding unmarked SOV sentences. The decision to exclude unmarked SOV sentences is based on two concerns. First, unmarked SOV sentences may not have sufficient representation in our corpus. Previous corpus studies have found that unmarked SOV sentences are a small minority of SOV sentences in natural language production (~ 16% in Liu's (2007) dataset and 21% in Sun and Givón's (1985) dataset), occurring much less often than *ba* constructions. Second, despite the apparent similarity in word order, unmarked SOV sentences and *ba* constructions are subject to significantly different semantic and pragmatic conditions in their use (Iemolo and Arcodia 2014; Liu 2007; Sun and Givón 1985). We elaborate this point with a comparison of (7a), an example of an unmarked SOV, and (7b), repeated from (4a) above.

- (7). a. ?他那本書放下了。  
 tā\_nà\_běn\_shū\_fàng\_xià\_le  
 he\_that\_CL\_book\_put\_down\_ASP  
 He put down that book.  
 b. 他把那本書放下了。  
 tā\_bǎ\_nà\_běn\_shū\_fàng\_xià\_le.  
 he\_BA\_that\_CL\_book\_put\_down\_ASP.  
*He put down that book.*

To many native speakers, (7a) may sound acceptable but will need more context to sound better. This is because an unmarked SOV sentence requires the preverbal NP to be in contrastive emphasis (Ernst and Wang 1995, among others). When the context does not supply a contrast to the object NP, as is the case in (7) where there is no contrast to 那本書 *nàběnshū*, the unmarked SOV sentence in (7a) sounds incomplete by itself. By contrast, such a requirement is not imposed for *ba* constructions, as shown in (7b). The importance of contrastive emphasis for unmarked SOV sentences is more clearly shown in the comparison of (8a) and (8b) below, which differ by the presence/absence of the contrast between 花生 *huāshēng* "peanut" and 海鮮 *hǎixiān* "seafood." Furthermore, while the *ba* construction requires the verb to have a disposal meaning, this requirement is not applied to unmarked SOV sentences. Therefore, while the unmarked SOV sentence in (8a) is grammatical with a modal verb 能 *néng* "can," the *ba* constructions in both (8c) and (8d) are ungrammatical.

- (8). a. 我花生能吃,海鮮不能吃。  
 wǒ\_huāshēng\_néng\_chī, hǎixiān\_bù\_néng\_chī  
 I\_peanut\_can\_eat, seafood\_NEG\_can\_eat  
*I can eat peanuts, but cannot eat seafood.*  
 b. ??我花生能吃。  
 wǒ\_huāshēng\_néng\_chī  
 I\_peanut\_can\_eat  
*I can eat peanuts.*  
 c. \*我把花生能吃,把海鮮不能吃。

wǒ\_bǎ\_huāshēng\_néng\_chī, bǎ\_hǎixiān\_bù\_néng\_chī

I\_BA\_peanut\_can\_eat, BA\_seafood\_NEG\_can\_eat

*I can eat peanuts, but cannot eat seafood.*

d. \*我把花生能吃。

wǒ\_bǎ\_huāshēng\_néng\_chī

I\_BA\_peanut\_can\_eat

*I can eat peanuts.*

### 1.3 Theoretical accounts of word order variation in Mandarin

What would the theories of word order variation predict for SVO-*ba* alternation in Mandarin? The production-oriented account will predict a longer-NP-preverbal pattern if meaning processing is prioritized and a shorter-NP-preverbal pattern if form processing is prioritized. Based on what we know about Japanese, it is possible that the longer-NP-preverbal preference is stronger. The prediction by the comprehension-oriented MiD principle depends on the phrase structure analysis of the *ba* construction. The current consensus on the syntax of *ba* construction is that the preverbal object NP and the V head constitute a VP, which is in turn contained in a *ba*P headed by the *ba* marker (Huang et al. 2009 and references therein). As shown in (9), under this analysis, the phrasal combination domain (PCD) of the VP (i.e., the linear distance between the V and the N nodes) is longer in the SVO word order (i.e., (9a)) than in the *ba* construction (i.e., (9b)), because Mandarin noun phrases are head-final. Nevertheless, in the *ba* construction, we also need to consider the PCD of the *ba*P, which is the linear sequence from the head *ba* marker to the V node. The PCD of the *ba*P, just like the PCD of the VP in the SVO word order, contains the full NP. Thus, when the object NP is heavy, there is no substantial benefit in terms of parsing efficiency associated with either SVO word order of the *ba* construction. Under this analysis, the comprehension-oriented account predicts no NP weight effect in SVO-*ba* word order variation.<sup>2</sup>

(9). a. PCD of the VP in the SVO word order

VP [V XP NP [..... N]]

-----

b. PCD of the VP and *ba*P in the *ba* construction

*ba*P [BA VP [NP [..... N] V]]

-----

One may also argue that Mandarin may pattern exactly like Hungarian, in that cross-domain NP shifts involving preverbal (secondary) topic positions are conditioned not by processing efficiency but by semantic-pragmatic factors. Under this analysis, NP weight effects in the SVO-*ba* alternation—if any—will be beyond the scope of prediction of the comprehension-oriented account.

To summarize, the comprehension-oriented account would either predict no NP weight effect or make no prediction, depending on what one believes to be the jurisdiction of the MiD principle.

#### 1.4 Modeling word order variation

Before delving to the methods of the current study, we will briefly review the existing literature that used similar methodology in the research of word order variation. With the help of corpus data and statistical models, Bresnan and colleagues (Bresnan et al. 2007; Bresnan and Ford 2010; Kendall et al. 2011; Kuperman and Bresnan 2012; Tily et al. 2009; Wolk et al. 2013; etc.) successfully modeled the phenomena of English dative variation (e.g., John gave Mary a gift vs. John gave a gift to Mary) and genitive variation (e.g., John's book vs. the book of John), revealing that word order variation in English is conditioned by a potpourri of semantic, structural, and contextual factors including NP accessibility, pronominality, definiteness, weight, verb semantics, structural parallelism, etc. Furthermore, the models' predictions of surface word order probabilities were shown to be strongly correlated with native speakers' natural judgment (Bresnan 2006) and processing times in reading tasks (Tily et al. 2009), confirming the validity of the model estimates. Following Bresnan et al.'s work, the corpus-based modeling methodology has been applied to the investigation of word order variation in other languages including Mandarin (e.g., Yao and Liu 2010 discussed above), Cantonese (Starr 2015), Estonian (Klavan et al. 2015), Persian (Faghiri et al. 2014), and Swedish (De Cuyper et al. 2014).

## 2 Methods

All the data for the corpus study come from the 10-million-word Academia Sinica Balanced Corpus of Modern Chinese (Version 5.0; Chen et al. 1996), which contains part-of-speech-tagged texts from both spoken and written sources. A preliminary search for *ba* constructions in the corpus found more than 11,000 sentence tokens, featuring over 1600 different verbs. Over 90% of the verbs have less than 25 tokens. Only 5 verbs have more than 200 tokens of *ba* constructions in the corpus: 放 *fàng* "to put" ( $N = 495$ ), 當 *dāng* "to consider as" ( $N = 341$ ), 帶 *dài* "to take/bring (with someone)" ( $N = 271$ ), 送 *sòng* "to give" ( $N = 230$ ), and 拿 *ná* "to take in one's hands" ( $N = 207$ ). Among these five, only 放 *fàng* and 拿 *ná* are used in the current analysis. The other three verbs are excluded for different reasons. 當 *dāng* is excluded because further corpus search shows that it is overwhelmingly used in *ba* construction rarely appears in SVO word order; therefore, there is not enough word order variation regarding 當 *dāng* that can be modeled. 送 *sòng* and 帶 *dài* are excluded because they are both used predominantly in dative sentences, with both a direct object NP (DO) and an indirect object NP (IO). As shown in Yao and Liu (2010), Mandarin dative sentences exemplify a three-way word order variation (Subj V DO IO vs. Subj V IO DO vs. Subj BA DO V IO), which is more complicated than the binary variation between SVO and *ba* construction. Since the main goal of the current study is to investigate the SVO-*ba* variation, we decide to exclude both 送 *sòng* and 帶 *dài* from the analysis.

To complement the *ba* sentence set, we compiled a SVO sentence set for each target verb (放 *fàng*, 拿 *ná*) by searching for all the sentences that contain the target verbs but not the *ba* marker in the corpus. The SVO-*ba* datasets were manually examined to exclude sentences that fall into the following categories: (i) sentences that cannot be converted to the alternative word order (mostly SVO sentences with bare verbs); (ii) SOV sentences with preverbal object NPs but not explicitly marked by *ba*; (iii) sentences with omitted object NPs; (iv) when the target verb phrase is part of an idiomatic expression; (v)

when the target verb is not used with the target sense (e.g., 放*fàng* can also mean “to release (e.g., someone from prison)”); (vi) when there is no reliable measure of context (e.g., when the sentence is in the beginning or the end of the text excerpt). Table 1 below shows the number of sentence tokens in the final datasets (see Yao 2014 for more details of dataset compilation). Four example sentences, one of each verb and each word order, are shown in (10).

- (10).a. 他卻死也不肯把槍放下。(Verb = 放*fàng*; word order = BA)  
 tā\_què\_sǐ\_yě\_bù\_kěnbǎ\_qiāng\_fàng\_xià  
 he\_but\_die\_also\_NEG\_willing to\_BA\_gun\_put\_down  
*But he would rather die than put down the gun.*
- b. 大人小孩都放下手上的工作及課業。(Verb = 放*fàng*; word order = SVO)  
 dàrén\_xiǎohái\_dōufàng\_xià\_shǒu\_shàng\_de\_gōngzuò\_jíkèyè  
 adult\_child\_all\_put\_down\_hand\_above\_DE\_work\_homework  
*All the adults and the children put down the work in their hands.*
- c. 政府是不是能把未來十年的教育採購經費一次拿出來?(Verb = 拿*ná*; word order = BA)  
 zhènfǔ\_shì\_bù\_shìnéng\_bǎ\_wèilǎi\_shínián\_de\_  
 jiàoyù\_cǎigòujīngfèiyī\_cì\_ná\_chūlái  
 government\_be\_NEG\_be\_can\_BA\_future\_ten\_year  
 \_DE\_education\_procure\_funds\_one\_CL\_take\_out  
*Can the government take out the education procurement funds for the next ten years in one go?*
- d. 阿媽一開口我就拿出我的小小紙條來(Verb = 拿*ná*; word order = SVO)  
 āmā\_yī\_kāikǒuwǒ\_jiù\_ná\_chū\_wǒ\_de\_xiǎozhǐtiáolái  
 mom\_once\_speak\_I\_then\_take\_out\_I\_DE\_little\_note ASP  
*As soon as mom started to speak, I took out my little note.*

All the sentence tokens were annotated for 14 properties pertaining to the surface word order (WordOrder), verb complement (VerbComp), the style of the text (Text-Mode), the object NP (ObjAnimacy, ObjIsPronoun, ObjHasPronDem, ObjLen), sentence structure (AdvP\_before, VP\_before, VP\_after), and the context (BA\_before, BA\_after, ObjMention\_before, ObjMention\_after). For all context-related measures, the scope of context is defined as the preceding and following 10 units (separated by comma, period, exclamation mark, or question mark) around the target structure. Below is the complete list of the 14 properties and their coding schemes:

### **ObjLen**

Critical variable. Number of characters in the object NP.

### **AdvP**

Whether the target VP is modified by a preceding adverbial phrase. If the target VP is modified by an adverbial phrase preverbally (e.g., 慢慢地把書放下 *mànmàndebǎshūfàngxià* “slowly put down the book”), then TRUE; otherwise, FALSE.

### **BA\_after**

Whether another *ba* sentence is used in the following context. If the following context has a *ba* sentence, then TRUE; otherwise, FALSE.

**BA\_before**

Whether another *ba* sentence is used in the preceding context. If the preceding context has a *ba* sentence, then TRUE; otherwise, FALSE.

**ObjAnimacy**

Whether the object NP is animate. If the object NP refers to an entity with life, then TRUE; otherwise, FALSE.

**ObjHasPronDem**

Whether the object NP contains a pronoun. If the object NP contains a pronoun (e.g., 他的書 *tādeshū* “his book”) or a demonstrative (e.g., 那本書 *nàběnshū* “that book”), then TRUE; otherwise, FALSE. By definition, ObjHasPronDem is true when ObjIsPron is TRUE.

**ObjIsPron**

Whether the object NP is a pronoun. If the object NP is a pronoun (e.g., 他 *tā* “he”) then TRUE; otherwise, FALSE

**ObjMention\_after**

Whether the object NP is mentioned later. If the object NP is mentioned in the following context, then TRUE; otherwise, FALSE.

**ObjMention\_before**

Whether the object NP is mentioned before. If the object NP is mentioned in the preceding context, then TRUE; otherwise, FALSE.

**TextMode**

Genre of the source text, as coded in the corpus. Possible categories: spoken, written, spoken-to-be-written, written-to-be-read, and written-to-be-spoken.

**VerbComp**

Complement or aspect marker of the verb, for example the complement 下 *xià* “down” in 放下 *fàngxià* “put down.”

**VP\_after**

Whether the target VP is followed by another VP. If the target VP is followed by another VP (e.g., 他放下書走了 *tāfàngxiàshūzǒule* “he put down the book bag (and then) left”), then TRUE; otherwise, FALSE.

**VP\_before**

Whether the target VP is preceded by another VP. If the target VP is preceded by another VP (e.g., 他幫我們把書放下 *tābāngwǒmenbǎshūfàngxià* “he helped us put down the books”), then TRUE; otherwise FALSE.

**WordOrder**

Dependent variable. If the target sentence is a *ba* construction, then 1; otherwise (i.e., if the target sentence is an SVO structure) 0.

Table 2 shows the distribution of the variables in each dataset. In the current study, we are most interested in the effects of weight (ObjLen) on surface word order (WordOrder). All the other properties will be entered into the model as control factors (random or fixed effects), as they have been shown to affect word order in previous studies on English and Chinese (Bresnan et al. 2007; Yao and Liu 2010).

As shown in Table 1, the two verbs, 放 *fàng* and 拿 *ná*, show different tendencies toward the *ba* construction. While the majority (> 70%) of the 放 *fàng* sentences use the *ba* construction, the majority (> 70%) of the 拿 *ná* sentences use the SVO word order,

**Table 2** Distribution of model predictor variables in the datasets

Variable	Distribution in the 放 <i>fàng</i> dataset	Distribution in the 拿 <i>ná</i> dataset
ObjLen	Raw: [1, 39], mean = 4.21, SD = 3.89 Log-transformed: [0, 3.66], mean = 1.15, SD = 0.72	Raw: [1, 33], mean = 5.22, SD = 4.04 Log-transformed: [0, 3.50], mean = 1.40, SD = 0.72
AdvP	False = 733; true = 214	False = 712; true = 276
BA_after	False = 707; true = 240	False = 778; true = 210
BA_before	False = 708; true = 239	False = 816; true = 172
ObjAnimacy	False = 837; true = 110	False = 984; true = 4
ObjHasPronDem	False = 805; true = 142	False = 835; true = 153
ObjIsPron	False = 891; true = 56	False = 963; true = 25
ObjMention_after	False = 672; true = 275	False = 805; true = 183
ObjMention_before	False = 656; true = 291	False = 811; true = 177
TextMode	Spoken = 32; spoken-to-be-written = 10; written = 886; written-to-be-read = 7; written-to-be-spoken = 12	Spoken = 70; spoken-to-be-written = 2; written = 880; written-to-be-read = 16; written-to-be-spoken = 20
VerbComp	14 different complement/aspect after 放 <i>fàng</i> (上 <i>shàng</i> "up"; 下 <i>xià</i> "down"; 入 <i>rù</i> "into"; 到 <i>dào</i> "into"; 回 <i>huí</i> "back"; 在 <i>zài</i> "at"; 好 <i>hǎo</i> "properly"; 妥 <i>tuǒ</i> "properly"; 平 <i>píng</i> "flat"; 進 <i>jìn</i> "into"; 錯 <i>cuò</i> "wrong"; 了 <i>le</i> ASP-PERF; 低 <i>dī</i> "low"; 給 <i>gěi</i> "to")	16 different complement/aspect after 拿 <i>ná</i> (上 <i>shàng</i> "up"; 下 <i>xià</i> "down"; 出 <i>chū</i> "out"; 去 <i>qù</i> "go"; 光 <i>guāng</i> "empty"; 回 <i>huí</i> "back"; 住 <i>zhù</i> "stay"; 走 <i>zǒu</i> "away"; 來 <i>lái</i> "come"; 掉 <i>diào</i> "remove"; 給 <i>gěi</i> "to"; 進 <i>jìn</i> "into"; 開 <i>kāi</i> "away"; 過 <i>guò</i> "over"; 到 <i>dào</i> "reach" (e.g., 拿到兩本書 <i>ná dào liǎng běn shū</i> "take (and reach) two books"; 到 <i>dào</i> "to" (e.g., 拿兩本書到那兒 <i>ná liǎng běn shū dào nà er</i> "take two books (to) there")
VP_after	False = 821; true = 126	False = 619; true = 369
VP_before	False = 877; true = 70	False = 874; true = 114

suggesting that the two verbs may further differ in terms of how their inclination to one or the other word order may be influenced by the predicting variables. In order to address such cross-verb variation while maintaining relatively simple model structures, we decided to model the two verbs separately and compare the model results in a discussion of verb-specific features in word order variation (see Section 4.2). Importantly, doing so allows us to avoid the use of higher-order interactions (e.g., the interaction between weight, givenness, and verb), which can be hard to interpret.

Two generalized mixed-effects models were built for the datasets of 放 *fàng* and 拿 *ná*, respectively, using the `lmer()` function in the `lme4` package (Bates et al. 2015; version 1.1–18–1) of R (R Core Team 2017; version 3.4.0). Each model started with `WordOrder` as the outcome variable, the full set of predictor variables as fixed effects, and `VerbComp` as the random effect. The critical predictor, `ObjLen`, was log-transformed and centered (Baayen 2008) before being entered into the models. Furthermore, since previous studies suggested a non-linear effect of NP weight (i.e., both very short and very long NPs are more likely to be preposed), we added a quadratic term of `ObjLen` in the models. Whether or not the quadratic effect of `ObjLen` (if any) is due to the interaction between weight and givenness is tested in subsequent model analysis.

The initial models were submitted to backward elimination, where non-significant predictors—predictors whose elimination did not change model fit significantly—were removed from the models. After eliminating non-significant predictors, we also tried to



add random slopes of the critical predictors, ObjLen and ObjLen<sup>2</sup>, to the models but the resulting models did not converge. Thus, we only report results from the final models with random intercepts.

### 3 Results

#### 3.1 Results of the verb 放 *fàng*

The final model of 放 *fàng* contained seven significant fixed-effect predictors. Table 3 below shows a summary of the fixed effects in the model. Critical effects with ObjLen are shown in bold.

As shown in Table 3, everything else being equal, ObjLen has a significant linear effect and a significant quadratic effect on the likelihood of using the *ba* construction. The positive coefficient of the quadratic term ( $\beta_{\text{ObjLen}^2} = 0.82$ ) indicates that the effect of ObjLen on *ba*-likelihood follows a U-shaped parabolic curve that opens upward (see Fig. 1). In other words, as ObjLen goes up, the likelihood of *ba* decreases first and then increases.

As one can roughly estimate from Fig. 1, the turning point from the downward trend to the upward trend (i.e., the bottom of the parabolic curve) happens near the value of 0.5 on the  $x$  axis. The exact value of the turning point can also be calculated. A parabolic curve like the one in Fig. 1 is symmetrical about a vertical line that goes through the turning point. As shown in (11), a parabolic curve defined by  $ax^2 + bx$  is symmetrical about  $x = -b/2a$ . Thus, with the coefficients of the quadratic term ( $\beta_{\text{ObjLen}^2} = 0.82$ ) and linear term ( $\beta_{\text{ObjLen}} = -0.80$ ) of ObjLen, the curve in Fig. 1 is symmetrical about  $x = -\beta_{\text{ObjLen}}/(2\beta_{\text{ObjLen}^2}) = 0.80/(2 \times 0.82) = 0.49$ , which also gives the exact value of the turning point. If we convert the transformed ObjLen back to the raw values, the turning point is around five characters (i.e.,  $e^{(0.49 + \text{mean}(\log(\text{ObjLen})))} = e^{(0.49 + 1.15)} = 5.15$ ). That is to say, controlling for other factors, when the object NP is less than five characters, the shorter it is, the more likely to use a *ba* construction; when the object NP is more than five characters, the longer it is, the more likely to use a *ba* construction.

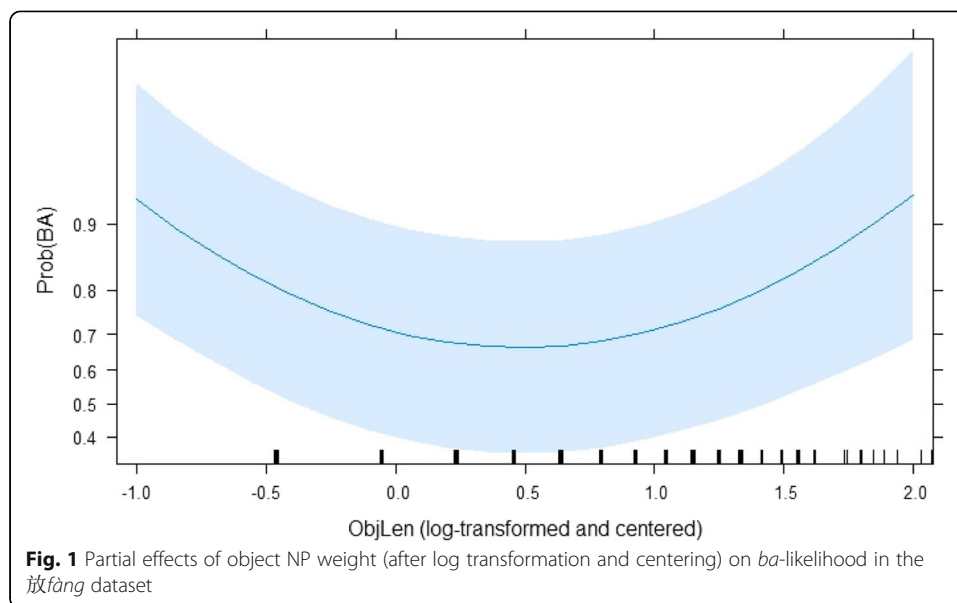
(11).

$$ax^2 + bx = a\left(x^2 + \frac{b}{a}x\right) = a\left(x - \left(\frac{-b}{2a}\right)\right)^2 - \left(\frac{-b}{2a}\right)^2$$

Other fixed effects in the model suggest that the likelihood of using *ba* increases when: (i) the target verb phrase is not preceded by an adverbial phrase; (ii) there is a *ba*

**Table 3** Summary of fixed effects in the model on 放 *fàng*

Predictor	Coefficient ( $\beta$ )	$z$	$p(> z )$
(Intercept)	0.47	0.71	0.48
AdvP_before = true	-1.34	-5.54	< .001
BA_after = true	0.90	3.44	< .001
ObjMention_before = true	1.00	3.58	< .001
ObjHasPronDem = true	1.13	2.72	.006
ObjLen	-0.80	-4.08	< .001
ObjLen <sup>2</sup>	0.82	4.06	< .001



construction in the following context; (iii) the object NP is mentioned in the previous context; (iv) the object NP contains a pronoun. Most of these effects are predicted, given what we know from previous literature about *ba* construction and word order variation. In particular, the effect of BA\_after aligns with the documented effect of structural parallelism (Bresnan et al. 2007; Yao and Liu 2010). The effects of ObjMention\_before are compatible with the effects of givenness (Bresnan et al. 2007; Yao and Liu 2010). The effect of ObjHasPronDem is consistent with previously observed tendencies of definite NPs and pronouns occurring earlier in the sentence (Bresnan et al. 2007; Li and Thompson 1981; among others). In addition, the model also shows that when the target verb phrase is preceded by an adverbial phrase, it is less likely to use the *ba* construction (see (12) for an example). One possible explanation is that the preverbal adverbial phrase prefers to be closer to the verb that it modifies and therefore prevents the preposing of the object NP, as in a *ba* construction.

- (12). 我輕輕放下刀叉。  
 wǒ\_qīngqīng\_fàng\_xià\_dāo\_chā  
 I\_lightly\_put\_down\_knife\_fork  
 I put down the knife and fork lightly.

Apart from the fixed effects, the model also estimates an adjustment to the intercept for each unique verb complement (VerbComp). Such adjustments are modeled as a random effect, because they represent the idiosyncratic propensities for the *ba* construction associated with each unique 放*fàng* + VerbComp combination. Given the assumptions of mixed-effects models, coefficients of a random effect are drawn from a normal distribution around zero. Table 4 lists the number of sentences of each VerbComp in the dataset as well as the by-VerbComp adjustments the model generates. As can be seen in Table 4, the distribution of VerbComp is anything but even, with the majority of the sentences containing verb complements such as 在 *zài* and 下 *xià*. The by-VerbComp adjustments also show significant cross-VerbComp differences. For

**Table 4** Random effects in the 放fàng model

VerbComp	上 shàng	下 xià	入 rù	到 rù	回 huí	在 zài	好 hǎo	妥 tuǒ	平 píng	進 jìn	錯 cuò	了 le	低 dī	給 gěi
Number of sentences	2	181	87	35	15	506	5	1	1	82	2	2	7	2
Adjustment to the intercept	-1.2	-	-	2.3	2.6	1.6	0.28	0.36	0.86	0.57	0.78	-	-	-
		2.9	0.63									1.9	2.0	1.7

example, compared to other 放fàng + VerbComp combinations, 放fàng + 在 zài has a stronger preference for *ba* constructions ( $\beta = 1.6$ ). On the other hand, the combination of 放fàng + 下 xià strongly favors the SVO word order ( $\beta = -2.9$ ). Is there any principled explanation for such variation? Can the by-VerbComp preference for the *ba* construction be predicted by telicity or other semantic or morphological features of the resultative compound? These are questions that need further investigation in future research.

The predictions that the model generates are likelihoods of using the *ba* construction. If we use 0.5 as the cut-off likelihood for a binary SVO-*ba* decision, the model makes the correct prediction over 90% of the time ( $(189 + 631)/947 = 86.6\%$ ; see Table 5). The baseline accuracy—when the model always guesses *ba*, which occurs more often than SVO—is  $(51 + 631)/947 = 72.0\%$ . In other words, by adding all the predictors, there is an increase of  $(86.6\% - 72.0\%)/72.0\% = 20.2\%$  in model accuracy.

To further evaluate the validity of the *ba*-likelihoods predicted by the model, we conducted a separate behavioral experiment with naturalness judgment. The experimental stimuli were a sample of 100 放fàng sentences taken from the corpus dataset, regardless of whether or not the model correctly predicted the surface word order. In order to maximize the diversity of the sentence stimuli, the selected sentences were distributed roughly evenly between the two surface word orders (SVO, *ba*) and across 10 equal *ba*-likelihood bands from 0 to 1 (as predicted by the model). On average, there were five sentences per surface word order per *ba*-likelihood band (SD = 1.91; range = [1, 9]). It should be noted that this design entails that about half of the sentence stimuli are not correctly predicted by the model (i.e., SVO sentences with high *ba*-likelihood and *ba* sentences with low *ba*-likelihood), even though in the complete dataset, such sentences only comprise 13.4% (i.e.,  $100\% - 86.6\%$ ) of the total.

A matching set of 100 sentences were constructed (hereafter the “constructed sentences”) by converting the sentence stimuli selected from the corpus (hereafter the “corpus sentences”) to the alternative word order (SVO- > *ba*; *ba*- > SVO). Twenty five native Chinese speakers (19 female, 6 male; mean age = 21.3 years old; SD = 3.32), all born and raised in Mainland China, participated in the experiment. The participant’s task was to rate the naturalness of different versions of a critical sentence in given contexts. In each trial, the participant was presented with the preceding and following context of a critical sentence (10 sentences before and 10 sentences after, the same as in the model dataset) and a place holder in the middle. The participant was then asked to

**Table 5** 放fàng model accuracy. Each cell shows the number of sentences given the surface word order and predicted word order

	Surface SVO	Surface <i>ba</i>
Predicted SVO	189	51
Predicted <i>ba</i>	76	631

consider a list of highly similar sentences in different word orders (*ba* or SVO)—with no knowledge of which one is from the corpus and which one is constructed—and rate how natural each sentence would sound if it is to replace the place holder in the given context, on a scale of 1 (=extremely unnatural) to 10 (=extremely natural), allowing any real number in between. An example trial is shown in [Appendix A](#). The complete list of corpus sentences is included in Additional file 1.

All the experimental sessions were administered on a computer in a self-paced manner. The experimental trials were randomly ordered, and the relative order of corpus and constructed sentences on the list was balanced for each surface word order. Three stimuli were excluded from the analysis due to critical errors in the target sentences. For each remaining stimulus, we calculated an overall *ba* rating (RatingBA; averaged across all participants) for the *ba*-version of the critical sentence and an overall SVO rating (RatingSVO; averaged across all participants) for the SVO-version. The difference between the two scores, RatingBA—RatingSVO, with a possible range of [− 10, 10] was derived as a measure of *ba*-propensity for the stimulus.

If corpus predictions and human intuition about sentence naturalness are highly aligned, we should observe a high correlation between the *ba*-likelihoods generated by the model and the *ba*-propensities from the experiment, meaning that a sentence predicted by the model to be very likely in the *ba* word order is also rated by human speakers as more natural in the *ba* word order than the alternative SVO word order, and vice versa. What we observe in the experimental dataset is an overall medium level of correlation ( $r = 0.43$ ) between the two data sources. However, recall that half of the experimental stimuli are not correctly predicted by the model. Is it possible that human speakers' intuition about these “difficult” cases actually align with the corpus and therefore differ from the model predictions? This hypothesis is borne out by a post hoc analysis. If we only look at the “difficult” stimuli, the correlation between corpus predictions and experimental ratings is basically non-existent ( $r = 0.06$ ). On the contrary, if we only look at the “easy” stimuli, which are correctly predicted by the model (i.e., *ba* in corpus and *ba*-likelihood > 0.5 OR SVO in corpus and *ba*-likelihood < 0.5), the correlation is much stronger ( $r = 0.71$ ). In other words, human speakers did not experience the same level of difficulty—as the model had—when predicting the word order for “difficult” cases.

The above being said, we should note that human speakers also had weaker intuition about “difficult” stimuli than “easy” stimuli. As shown in [Table 6](#), for “easy” stimuli, the ratings of *ba*-propensity clearly distinguish the stimuli that appear in the *ba* form in the corpus (mean = 2.08) from those that appear in SVO (mean = − 1.58). The perception of relative *ba*-propensity for “difficult” stimuli is much weaker, although the ratings are on average positive for stimuli that appear in the *ba* form in the corpus (mean = 0.37) and negative for those that appear in SVO (mean = − 0.18). That is to say, the “difficult” stimuli may very well be difficult to judge (or predict) for both human speakers and the model.

Taken together, given that “difficult” stimuli are the major source of model-experiment discrepancy and that such stimuli have a much higher presence in the experiment (50%) than in a natural corpus dataset (< 15%), we think it is fair to conclude that the *ba*-likelihoods predicted by the model are largely corroborated by human speakers' judgment of *ba*-propensity based on naturalness ratings.

We also constructed an alternative model to examine whether the observed quadratic effects of object NP weight could be fully explained by the interaction between weight

**Table 6** Comparison of experimental ratings of *ba*-propensities and model predictions of *ba*-likelihoods for “easy” and “difficult” stimuli regarding the verb 放 *fàng*

	Number of experimental stimuli	Mean <i>ba</i> -propensities (experiment)	Mean <i>ba</i> -likelihoods (model)
“Easy” stimuli (i.e., correctly predicted by the model)	N = 49 (i.e., 51%)	<i>ba</i> in corpus, 2.08	<i>ba</i> in corpus, 0.79
		SVO in corpus, -1.58	SVO in corpus, 0.28
“Difficult” stimuli (i.e., wrongly predicted by the model)	N = 48 (i.e., 49%)	<i>ba</i> in corpus, 0.37	<i>ba</i> in corpus, 0.28
		SVO in corpus, -0.18	SVO in corpus, 0.74

and givenness, as Liu (2007) proposed. The alternative model had a highly similar structure to the 放 *fàng* model, except that the quadratic term of ObjLen was replaced by ObjLen × ObjMention\_before. Contra to Liu’s proposal, only ObjMention\_before produced a positive effect on *ba*-likelihood (i.e., previously given object NP is more likely to be preposed); there was no significant effect of ObjLen or ObjLen × ObjMention\_before on *ba*-likelihood (both  $|z| < 1$ ,  $p(>|z|) > .3$ ). In other words, without the quadratic term, ObjLen in its linear form has no significant effect on the use of *ba*, probably because both short and long object NPs tend to be preposed, and the quadratic effect of ObjLen could not be fully predicted by the interaction of ObjLen and ObjMention\_before.

### 3.2 Results of the verb 拿 *ná*

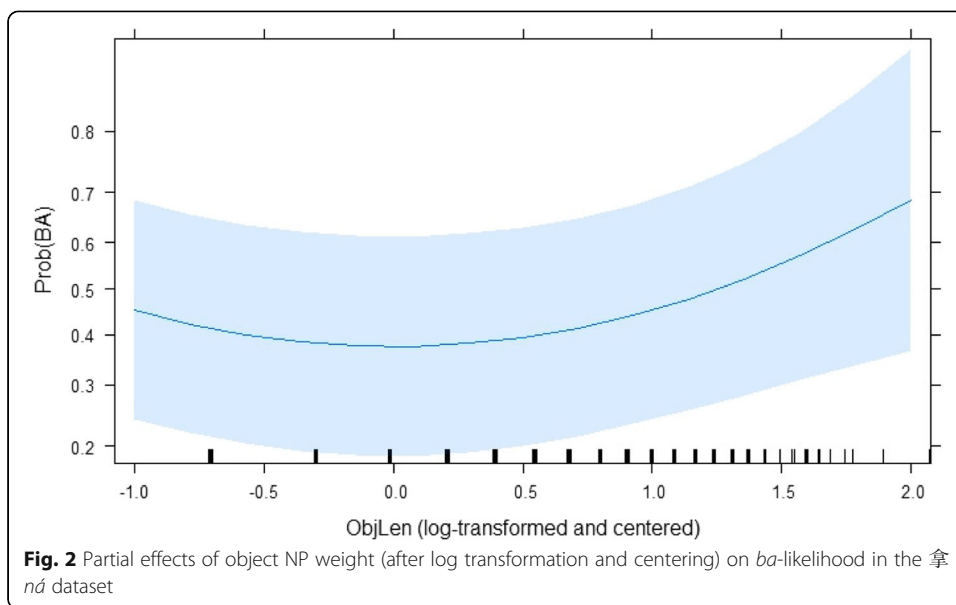
The model on the verb 拿 *ná* contains six significant fixed effects (see Table 7).

Similar to the model on 放 *fàng*, the model on 拿 *ná* also reveals a significant quadratic effect of ObjLen ( $\beta_{\text{ObjLen}^2} = 0.32$ ), indicating a U-shaped effect of ObjLen (see Fig. 2). However, the linear term of ObjLen is not present in the model, suggesting that the parabolic curve is symmetrical about  $x = 0$ , and that the turning point of the effect is when ObjLen = 0. A transformed ObjLen at 0 can be converted back to a raw value of ObjLen at  $e^{(0 + \text{mean}(\text{ObjLen}))} = e^{(1.40)} = 4.05$ , i.e., around four characters long. In other words, when the object NP has less than four characters, the shorter it is, the more likely to use the *ba* construction; when the object NP is longer than four characters, the longer it is, the more likely to use the *ba* construction.

Similar to the 放 *fàng* model, the 拿 *ná* model also shows significant effects of the use of *ba* construction in the following context, mentions of the object NP in the surrounding context, and the use of pronouns in the object NP, all in the same directions as one would predict based on previous literature. Furthermore, the 拿 *ná* model also shows

**Table 7** Summary of fixed-effects predictors in the model on 拿 *ná*

Predictor	Coefficient ( $\beta$ )	$z$	$p(> z )$
(Intercept)	-0.96	-1.90	0.06
BA_after = true	0.48	2.22	.03
ObjMention_before = true	1.23	5.20	< .001
ObjMention_after = true	0.53	2.28	0.02
ObjHasPronDem = true	0.83	3.49	< .001
VP_before = true	-0.84	-2.66	.0008
ObjLen <sup>2</sup>	0.32	2.39	0.02



an effect of the presence of a preceding VP ( $\beta_{ObjLen^2} = -0.84$ ), suggesting that if the target verb phrase is preceded by another VP, it is less likely to use the *ba* construction. (13) shows an example of a 拿 *ná* phrase preceded by another VP headed by the verb 要求 “request.” We are not sure what causes this effect. It could be that the preceding verb phrase, which is often not a *ba* construction, discourages the use of a *ba* construction right after.

- (13). 印尼政府應該要求志願救援工作小組拿出證據。  
 yìní zhèngfǔ yīnggāi yāoqiú zhìyuàn jiùyuán xiǎozǔ ná chū zhèngjù  
 Indonesian\_government\_should\_  
 request\_volunteer\_rescue\_team\_take\_out\_evidence  
*The Indonesian government should request the voluntary rescuing team to supply evidence.*

The model also shows by-VerbComp variation in terms of the tendency toward the *ba* construction. As shown in Table 8, other things being equal, some complements, such as 下 *xià* “down” ( $\beta = 1.4$ ) and 到<sub>2</sub> *dào* “to” ( $\beta = 2.3$ ) are more likely to promote the use of *ba* construction than other complements, whereas 到<sub>1</sub> *dào* “reach” ( $\beta = -4.7$ ) has a much stronger tendency toward the SVO word order than any other complement. Such variation may be related to properties of the event described by the 拿 *ná*

**Table 8** Random effects in the 拿 *ná* model

VerbComp	上 <i>shàng</i>	下 <i>xià</i>	出 <i>chū</i>	去 <i>qù</i>	回 <i>huí</i>	光 <i>guāng</i>	住 <i>zhù</i>	走 <i>zǒu</i>
Number of sentences	3	10	392	73	29	1	3	24
Adjustment to the intercept	0.35	1.4	-1.3	-0.33	-0.38	1.2	-0.006	0.18
VerbComp	來 <i>lái</i>	掉 <i>diào</i>	進 <i>jìn</i>	給 <i>gěi</i>	開 <i>kāi</i>	過 <i>guò</i>	到 <sub>1</sub> <i>dào</i>	到 <sub>2</sub> <i>dào</i>
Number of sentences	91	38	4	38	5	6	216	29
Adjustment to the intercept	0.90	0.44	1.2	-0.96	0.38	-0.58	-4.7	2.3

+ VerbComp combination. The exact nature of such variation awaits further investigation in future research.

Table 9 shows the accuracy of word order predictions of the 拿 *ná* model. Overall accuracy rate is  $(674 + 147)/988 = 83.1\%$ , compared with the baseline accuracy of  $(677 + 50)/988 = 73.5\%$  by always guessing the more frequent SVO word order.

The validity of the *ba*-likelihoods predicted by the 拿 *ná* model is confirmed by a naturalness rating experiment with a separate group of 25 native Chinese speakers (19 female, 6 male; mean age = 22.2 years old; SD = 3.63). A total number of 91 sentence stimuli were sampled from the 拿 *ná* dataset (complete sentence list in Additional file 2), roughly evenly distributed between the two surface word orders and across predicted *ba*-likelihood bins (0–0.1, 0.1–0.2, ... 0.9–1). Although we aimed to have at least four stimuli per surface word order per *ba*-likelihood bin, a few likelihood bins close to 0 or 1 have fewer stimuli, due to the small number of representations in the dataset. On average, there are 4.6 stimuli per word order per likelihood bin (SD = 2.08; range = [1,8]). More than half of the stimuli are correctly predicted by the model (i.e., “easy” stimuli) while the rest are predicted wrong (i.e., “difficult” stimuli). Experimental procedure and data analysis followed that of the experiment for 放 *fàng*. Not surprisingly, human speakers’ judgment of *ba*-propensity had a sizable correlation with the model’s prediction of *ba*-likelihoods ( $r = 0.56$ ) for the easy stimuli ( $N = 55$ ), but the correlation was negative for the difficult stimuli ( $r = -0.55$ ;  $N = 36$ ), resulting in an overall low correlation between experimental ratings and model predictions ( $r = 0.23$ ). Furthermore, human speakers also had stronger intuitions about the SVO-*ba* distinction in the “easy” stimuli than in the “difficult” stimuli (Table 10).

We also tested if the quadratic effect of ObjLen could be explained as an interaction between weight and givenness. In the alternative model that replaced the quadratic term  $\text{ObjLen}^2$  with the linear term ObjLen and its interaction with ObjMention\_before, most of the control predictors remained significant, with effects in the same directions and similar magnitude as in the original 拿 *ná* model (see Table 11, compared with Table 7). However, while the main effect of ObjMention\_before remains significant and positive, both ObjLen and  $\text{ObjLen} \times \text{ObjMention\_before}$  turn out to be significant in the new model. Specifically, ObjLen has a significant positive effect ( $\beta_{\text{ObjLen}} = 0.35$ ) when ObjMention\_before is false (i.e., the baseline value of ObjMention\_before), and the effect becomes positive when ObjMention\_before is true ( $\beta_{\text{ObjLen}} + \beta_{\text{ObjLen} \times \text{ObjMention\_before}} = 0.35 - 0.76 = -0.41$ ). That is to say, when the object NP is given, the effect of weight is negative (i.e., longer NPs are less likely to be preposed), but when the object NP is new, the weight effect is in the positive direction (i.e., longer NPs are more likely to be preposed). This crisscross pattern is clearly shown in Fig. 3, which plots the partial effects of the interaction of ObjLen and ObjMention\_before in the 拿 *ná* model with the interaction term.

**Table 9** 拿 *ná* model accuracy. Each cell shows the number of sentences given the surface word order and predicted word order

	Surface SVO	Surface <i>ba</i>
Predicted SVO	674	114
Predicted <i>ba</i>	53	147

**Table 10** Comparison of experimental ratings of *ba*-propensities and model predictions of *ba*-likelihoods for “easy” and “difficult” stimuli regarding the verb 拿 *ná*

	Number of experimental stimuli	Mean <i>ba</i> -propensities (experiment)	Mean <i>ba</i> -likelihoods (model)
“Easy” stimuli (i.e., correctly predicted by the model)	<i>N</i> = 55 (i.e., 60%)	<i>ba</i> in corpus, 1.43 SVO in corpus, -2.20	<i>ba</i> in corpus, 0.78 SVO in corpus, 0.22
“Difficult” stimuli (i.e., wrongly predicted by the model)	<i>N</i> = 36 (i.e., 40%)	<i>ba</i> in corpus, 1.68 SVO in corpus, -0.97	<i>ba</i> in corpus, 0.28 SVO in corpus, 0.70

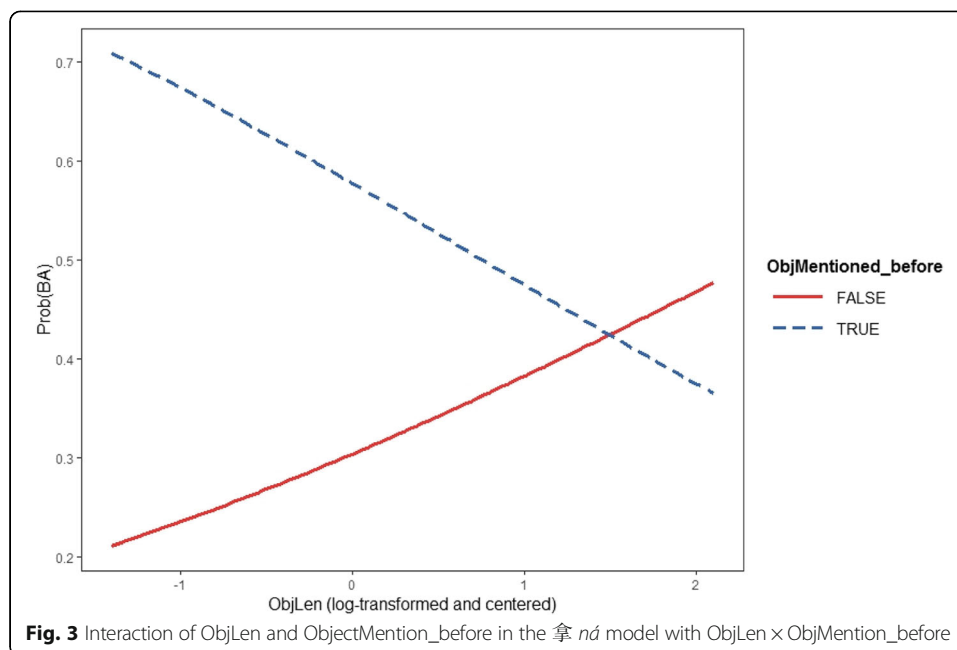
How do we explain the results presented above, with regard to the effect of *ObjLen* × *ObjMention\_before* when the quadratic term *ObjLen*<sup>2</sup> is not in the model? Since a quadratic effect (e.g., *ObjLen*<sup>2</sup>) is essentially the interaction of a variable with itself (e.g., *ObjLen* × *ObjLen*), the apparent interchangeability between *ObjLen*<sup>2</sup> and *ObjLen* × *ObjMention\_before* leads us to speculate that *ObjLen* is somehow closely related with *ObjMention\_before* in this dataset. The speculation is not completely baseless, as object NPs given the previous context (*ObjMention\_before* = true) can often take concise forms and therefore shorter in length. A post hoc analysis found that previously mentioned object NPs (mean = 3.61, SD = 2.98) are indeed significantly shorter than new object NPs (mean = 5.58, SD = 4.15; *t*(343) = -7.36, *p* < .001). In other words, the quadratic effect of *ObjLen*<sup>2</sup> can be explained as the interaction of *ObjLen* × *ObjMention\_before*, with the use of *ba* construction showing a downward trend with *ObjLen* when the object NP is given (i.e., when *ObjLen* is short) and an upward trend with *ObjLen* when the object NP is new (i.e., when *ObjLen* is long). This is compatible with the interaction of givenness and weight found in Liu (2007).

To summarize, the two models of 放 *fàng* and 拿 *ná*, respectively, reveal highly similar results regarding model accuracy, the general patterns of SVO-*ba* alternation, the effects of fixed-effects predictors, and individual differences across verb complements. What is most important for the current study is the presence of a quadratic effect of object NP length in both models. We discuss the interpretation of the quadratic length effect in more detail in the next section.

**Table 11** Summary of fixed-effects predictors in the alternative model on 拿 *ná*, with an interactive effect of *ObjLen* × *ObjMention\_before*

Predictor	Coefficient ( $\beta$ )	<i>z</i>	<i>p</i> (>  <i>z</i>  )
(Intercept)	-0.83	-1.67	0.09
<i>BA_after</i> = true	0.45	2.07	.04
<i>ObjMention_after</i> = true	0.53	2.28	0.02
<i>ObjHasPronDem</i> = true	0.83	3.48	< .001
<i>VP_before</i> = true	-0.74	-2.36	.02
<i>ObjMention_before</i> = true	1.14	4.38	< .001
<i>ObjLen</i>	0.35	2.27	0.02
<i>ObjLen:ObjMention_before</i> = true	-0.76	-2.34	0.02





## 4 Discussion

### 4.1 Summary of the findings

So far, we have shown that the effect of object NP weight on *ba*-likelihood is more complicated than a simple linear effect. In the model of the verb 放 *fàng*, the weight effects are best characterized by a U-shaped quadratic curve, with the lowest point around (untransformed) ObjLen = 5 characters. The effect is not confluent with an interaction of NP weight and givenness. In fact, the interaction of object weight and givenness is not significant, even when the quadratic effect of NP weight is not included in the model. On the other hand, in the model of 拿 *ná*, although the same quadratic effect of NP weight is observed on *ba*-likelihood, this effect seems to be confluent with an interactive term between weight and givenness, compatible with the pattern suggested in Liu (2007), which was based on observations from a dataset of mixed verbs. Thus, our results indicate that the quadratic effect of NP weight on *ba*-likelihood is present across verbs but the interaction between weight and givenness is not. If the dataset happens to have a very high percentage of heavy + new and light + given NPs, it is possible for the two to be confluent.

Model predictions of *ba*-likelihoods from both models are largely corroborated by human intuition gathered from the naturalness rating experiments, especially for the items that the models can make correct predictions for ( $r = 0.71$  for 放 *fàng*;  $r = 0.56$  for 拿 *ná*), which are the majority of the items (> 80%) in the corpus datasets but only about half of the stimuli in the experiments (< 60%). The discrepancy between model predictions and human ratings can be attributed mainly to the items that the models fail to predict correctly—which only comprise a small percentage of the corpus datasets but about half in the experimental stimuli. The differences in dataset design may further lead to different ranges and distributions of the critical variables (e.g., the corpus datasets have a much larger range of ObjLen than the experimental stimuli), which becomes another possible source of discrepancy between model predictions and experimental results (Arnold et al. 2000). Last but not the least, the

discrepancy may also stem from inherent differences between written language and spoken language. While the corpus data are predominantly in the written genre from published sources, it is not clear whether experimental participants are also operating in the written-language mode or the spoken-language mode when they rate the naturalness of the sentence stimuli.

#### 4.2 Cross-verb variation

Despite the similarity in model results between 放 *fàng* and 拿 *ná*, the current study also reveals some interesting cross-verb differences. In addition to the presence (absence) of a significant interaction between weight and givenness, as discussed above, the two verbs 放 *fàng* and 拿 *ná* differ greatly in terms of *ba* propensity. While 放 *fàng* is highly biased toward *ba* construction (baseline *ba* probability > 60%), 拿 *ná* is much more likely to appear in SVO word order (baseline *ba* probability < 30%). Such cross-verb differences are also observed in other verbs that occur in the corpus but not included in the current study. For example, the verb 當 *dāng* predominantly occurs in *ba* constructions and does not seem to allow word order alternation at all. Furthermore, such verb-specific *ba* tendencies are echoed by the complement-specific *ba* propensities that we have observed within each verb in the models reported above. These idiosyncratic verb-specific or complement-specific *ba* tendencies may be related to the semantic properties of the verb/complement and/or the properties of the event described by the verb/complement, but the exact patterns of these effects are beyond the scope of the current study.

The current study only examines two verbs, due to the limitations of the corpus data we use. A full scale of the investigation of cross-verb differences will be possible if a larger and more comprehensive dataset becomes available, which will allow multiple verbs to be either modeled separately or together in a mixed-effects model with verb as one of the random effects.

#### 4.3 Interpretation of the NP weight effects

How to interpret the current results regarding the effect of NP weight? As discussed above, although previous literature has mostly focused the relative order of two NPs in the preverbal or postverbal domain, the two theoretical accounts—production-oriented or comprehension-oriented—can both be extended to make predictions for the cross-domain shift of a single NP. In a nutshell, the production-oriented account predicts that both longer-NP-preverbal and shorter-NP-postverbal preferences are possible, depending on whether the word order is more sensitive to conceptual or positional factors. The comprehension-oriented MiD principle, on the other hand, either predicts a null effect of NP weight or makes no prediction regarding NP weight effects, depending on the assumed realm of prediction of the MiD principle. The current modeling results are more compatible with the production-oriented account. Below we discuss in detail how the U-shaped weight effects may arise from production-oriented account.

The underlying principle of the production-oriented account is that easy-to-access (i.e., high-accessibility) NPs will appear earlier in the sentence, giving more time for other phrases that demand more resources to be accessed and assembled and thus maximizing the smoothness of the incremental process of sentence production. In this

regard, NP weight is a complex measure, because it is a proxy for a number of properties that could affect the ease/difficulty of production of an NP, including pronominality, identifiability, givenness, semantic richness, and sequence complexity. We know that pronouns are usually lighter than regular full NPs, and that NPs with highly identifiable or previously given referents tend to be lighter as they require less lexical information to specify. Object NPs with these properties (pronominality, high identifiability, givenness) are easier to access and tend to be placed earlier in the sentence. In our models, these properties are explicitly coded by variables like *ObjIsPronoun*, *ObjHasPronDem*, and *ObjMention\_before*, and the predicted effects on promoting NP preposing (as in a *ba* construction) are consistently observed in the models of both verbs. Therefore, what is captured in the effects of NP weight in the models must be beyond the effects of pronominality, identifiability, and givenness. We consider the weight variable *ObjLen* in the models mainly as the proxy of semantic richness and sequence complexity.

Semantic richness and sequence complexity are both associated with heavy NPs; however, the two operate have opposite effects on NP accessibility. Heavy NPs contain more lexical information (i.e., high semantic richness), at the same time they also have more complicated syntactic structures (i.e., high sequence complexity) as they are longer sequences. Nevertheless, as reviewed above, semantic richness, which operates in the conceptual (i.e., meaning) arena, leads to conceptual salience and accessibility, and as a result, promotes the earlier placement of heavy NPs. On the other hand, sequence complexity, which operates in the positional (i.e., form) arena, directly informs the level of difficulty of assembling a linear sequence and is negatively correlated with form accessibility. Thus, the effect of sequence complexity is to promote the earlier placement of light NPs.

Previous studies explained the long-before-short tendency in Japanese and short-before-long tendency in English as cross-language differences in the relative weightings of conceptual accessibility and form accessibility. The fact that we observe weight effects in both directions in U-shaped parabolic curves suggests that both conceptual and positional factors are at work in the SVO-*ba* alternation. A likely scenario is as follows: conceptual accessibility promotes heavier (and thus more salient) NPs to an earlier, preverbal position, while ease of sequencing prefers lighter (and thus easier) NPs to occur earlier. Crucially, we have to assume that the default word order is SVO, and object preposing only occurs when either conceptual accessibility or ease of sequencing is highly activated. Consequently, both very light and very heavy NPs are more likely to be preposed compared to their medium-weight counterparts, which do not excel in either type of accessibility. A different scenario is one in which both conceptual and form accessibilities are at work simultaneously, without assuming SVO as the default word order. In this scenario, conceptual accessibility favors heavy NPs in the preverbal position and lighter NPs in postverbal positions, while form accessibility favors light NPs in the preverbal position and heavy NPs in postverbal positions. For there to be an overall U-shaped weight effect, the effect of conceptual accessibility must outweigh that of form accessibility for NPs near the heavy end, and the effect of form accessibility must outweigh that of conceptual accessibility for NPs near the light end. In addition, the sum of the two effects on medium-weight NPs must place them in the postverbal domain. Only when all three conditions are met will we observe the overall

pattern of both heavy and light NPs being more likely to be in the preverbal domain than their medium-weight counterparts.

Although the current findings do not provide direct support for the comprehension-oriented account, they are, however, more compatible with the analysis that considers topic/focus positions in the preverbal domain to be outside of the jurisdiction of the MiD principle. Instead, this type of postverbal-to-preverbal NP shift (into topic/focus positions) is hypothesized to be influenced by semantic-pragmatic factors. This account is compatible with our observation of conceptual accessibility being one of the conditioning factors for the Mandarin NP shift from a postverbal position to the preverbal *ba* NP position, which has often associated with topicality in the literature. Along this line, one may further hypothesize that cross-domain NP shift as well as NP shift in the preverbal domain of Mandarin sentences are only sensitive to accessibility-based considerations, whereas NP shift in the postverbal domain of Mandarin sentences may be sensitive to both accessibility-based and parsing-related considerations.

To summarize, we find evidence of conceptual and positional factors working at the same time—in opposite directions—in the SVO-*ba* alternation in Mandarin. The results are compatible with other findings of Chinese word order variation. As mentioned above, Yao and Liu (2010) found in Mandarin dative sentences a tendency for heavy direct object NPs (relative to the indirect object NPs) to be preposed to the *ba* position, but when both direct and indirect object NPs are postverbal, there is a clear short-before-long preference. Combined with the current findings, we argue that in Mandarin sentence production, conceptual factors have stronger influence in the preverbal domain and positional factors are more prominent in the postverbal domain. This pattern is in general consistent with what has been observed in languages that differ in headedness.

Finally, going back to the questions raised at the end of Section 1, results from the Mandarin corpus analysis provide unambiguous evidence that both conceptual and positional factors are operating in the preverbal domain, but positional factors are much more prominent than conceptual factors in the postverbal domain. These results also showed that the relative sensitivity to conceptual (or positional) factors can vary within the language and is unlikely to be have risen solely from headedness.

## 5 Conclusion

In this study, we showed that the effects of NP weight on word order variation are more complicated in Mandarin Chinese than what have been previously documented for other languages (English, Japanese, Korean, etc.). Specifically, both short-before-long and long-before-short tendencies were observed in word order variation across the preverbal and postverbal domains. We interpret the current results as evidence for both conceptual and positional factors operating in Mandarin Chinese, especially in the preverbal domain. These findings contribute to the general understanding of the underlying mechanisms for word order variation across languages.

## 6 Endnotes

<sup>1</sup>However, it should be noted that the PCD as defined above may not be adequate for the listener to recognize the correct constituent structure. For example, when the

listener encounters the head of the second PP (i.e., for in (1a) and in in (1b)), it is not clear whether the second PP is a daughter of the VP or a complement of the preceding NP. In reality, the listener may need to use the semantic/pragmatic information contained in the words after the head of the second PP for disambiguation.

<sup>2</sup>Matthews and Yeung (2000) examined the variation between VO and OV word orders in another Chinese language, Cantonese, with a reading experiment. They found a processing advantage for OV sentences over the VO counterparts when the object NP was heavy, i.e., containing a relative clause that modifies the N head. Matthews and Yeung attributed the processing advantage to the shortened PCD of the verb phrase in OV word order. It should be noted that the preverbal stimuli Matthews and Yeung's study were all OSV sentences and did not include the *ba* construction, although Cantonese shares a similar *jiang* construction. Thus, the analysis in (9) does not apply to Matthews and Yeung's study, neither does Matthews and Yeung's analysis apply to the current study.

## 7 Appendix A

An example trial on the behavioral experiment on 放 *fàng*. It should be noted that all the text was presented in simplified Chinese characters during the actual experimental sessions.

### Chinese version:

...要說很多的感謝,感謝這群人,其實說這些是人,我覺得他們已經愈來愈不像人,愈來愈像神。這幾年我都一直擔心再接到學校的電話,通訊簿我都不大敢寫,因為有次昱農溺水,朱朱打電話給我的當下,我真的不知道我說了些什麼, \_\_\_\_\_[S]\_\_\_\_\_,我就一直很擔心,這個學校會不會因為這樣而以後不辦了。.....我真的很感謝這些人,辦了一個學校,讓我可以很放心地把兩個小孩子都放在這。"謝謝林林和羅羅,在我三年級不斷偷東西的時候,和我一直談話,..."

S1: 可是當我把電話放下之後  
(1 = "很奇怪,極不自然"; 10 = "極為自然"。可以用小數點。)

S2: 可是當我放下電話之後  
(1 = "很奇怪,極不自然"; 10 = "極為自然"。可以用小數點。)

### English glosses:

.....(I want to) say many thanks, thanks to these people. In fact, I think they are less and less like people but more and more like gods. In the past few years, I have been afraid of receiving a call from the school, afraid of writing my phone number on the contact list, because that time when Linong had the drowning accident, when Zhuzhu called me, I really did not know what to say. \_\_\_\_\_[S]\_\_\_\_\_, I was really worried. Will the school have to close because of accidents like this? ... I really want to thank these people, who created a school like this, so that I could leave my kids here without worrying. "Thanks to Linlin and Luoluo, who kept talking to me when I had shoplifting problems in the third grade..."

S1: But after I put down the phone  
(1 = extremely unnatural; 10 = extremely natural. Any number in between allowed.)

S2: But after I *ba* phone put down  
(1 = extremely unnatural; 10 = extremely natural. Any number in between allowed.)

## 8 Additional files

**Additional file 1:** Experimental stimuli for the verb 放 *fàng*. (PDF 1064 kb)

**Additional file 2:** Experimental stimuli for the verb 拿 *ná*. (PDF 980 kb)

**Acknowledgements**

The author is grateful to Chang Liu for research assistance and to the audience at the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC 28) for helpful comments.

**Funding**

This work was supported by the Hong Kong Research Grants Council's Early Career Scheme [grant number 558913].

**Availability of data and materials**

Please contact author for data requests.

**Authors' contributions**

YY conceived of the study, carried out the corpus analysis and behavioral experiments, and drafted the manuscript. The author read and approved the final manuscript.

**Ethics approval and consent to participate**

The study was approved by the Human Subjects Ethics Application Review System at the Hong Kong Polytechnic University (no. HSEARS20130406001-02), and written consent was obtained from all the participants.

**Consent for publication**

Not applicable.

**Competing interests**

The author declares that there are no competing interests.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 13 April 2018 Accepted: 13 November 2018

Published online: 20 December 2018

**References**

- Arnold, Jennifer E. 2008. Reference production : Production-internal and addressee-oriented processes. *Language and Cognitive Processes* 23 (4): 495–527.
- Arnold, Jennifer E., Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76 (1): 28.
- Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. New York: Cambridge University Press.
- Bates, Douglas M., Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67 (1): 1–48.
- Bock, J.Kathryn. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 89 (1): 1–47.
- Bock, J.Kathryn. 1986. Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory* 12 (4): 575–586.
- Bresnan, Joan. 2006. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In *Roots: Linguistics in search of its evidential base*, ed. Sam Featherston and Wolfgang Sternefeld, 75–96. Berlin: Mouton de Gruyter.
- Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, ed. Gerlof Boume, Irene Kraemer, and Joost Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.
- Bresnan, Joan, and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86 (1): 168–213.
- Chang, Franklin. 2009. Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language* 61 (3): 374–397.
- Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design methodology for balanced corpora. In *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*, ed. B.-S. Park and J. B. Kim, 167–176. Seoul: Kyung Hee University.
- Choi, Hye-Won. 2007. Length and order: A corpus study of Korean dative-accusative construction. *Discourse and Cognition* 14: 207–227.
- Chu, Chauncey C 屈承熹. 1979. Word order and the issues of word order change in Chinese 漢語的詞序與詞序變遷中的問題. In *Lectures on linguistics: Theories, application and Chinese grammar 語言學論集:理論,應用及漢語法*, ed. Chauncey C. Chu, 87–121. Taipei: Crane Publishing Co.
- De Cuyper, Ludovic, Kristof Baten, and Gudrun Rawoens. 2014. A corpus-based analysis of the Swedish passive alternation. *Nordic Journal of Linguistics* 37 (2): 199–223.
- De Smedt, Koenraad. 1994. Parallelism in incremental sentence generation. In *Parallel Natural Language Processing*, ed. Geert Adriaens and Udo Hahn, 421–447. Norwood, NJ: Ablex.
- Ernst, Thomas, and Chengchi Wang. 1995. Object preposing in Mandarin Chinese. *Journal of East Asian Linguistics* 4 (3): 235–260.
- Faghiri, Pegah, Pollet Samvelian, and Barbara Hemforth. 2014. Accessibility and word order : The case of ditransitive constructions in Persian. In *Proceedings of the 21st international conference on head-driven phrase structure grammar*, 217–237. New York: University at Buffalo.

- Gahl, Susanne, Yao Yao, and Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66 (4): 789–806.
- Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford and Hong Kong: Oxford University Press.
- Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.
- Hawkins, John J.A. 1994. *A performance theory of order and constituency*. Cambridge (England) and New York: Cambridge University Press.
- Hopper, Paul J., and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56: 251–299.
- Huang, Cheng-Teh James, Yen-hui Audrey Li, and Yafei Li. 2009. *The syntax of Chinese*. Cambridge: Cambridge University Press.
- Iemmolo, Giorgio, and Giorgio Francesco Arcodia. 2014. Differential object marking and identifiability of the referent: A study of Mandarin Chinese. *Linguistics* 52 (2): 315–334.
- Kendall, Tyler, Joan Bresnan, and Gerard van Herk. 2011. The dative alternation in African American English. *Corpus Linguistics and Linguistic Theory* 7 (2): 229–244.
- Kiss, Katalin E. 1981. Structural relations in Hungarian, a “free” word order language. *Linguistic Inquiry* 12 (2): 185–213.
- Kiss, Katalin E. 1987. *Configurationality in Hungarian*. Dordrecht: Reidel.
- Kiss, Katalin E. 1998. Identificational focus versus information focus. *Language* 74 (2): 245.
- Klavan, Jane, Maarja-Liisa Pilvik, and Kristel Uibo. 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. *SKY Journal of Linguistics* 28: 187–224.
- Kuperman, Victor, and Joan Bresnan. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66 (4): 588–611.
- LaPolla, Randy. 1990. *Grammatical relations in Chinese: Synchronic and diachronic considerations*. Berkeley: University of California.
- LaPolla, Randy. 1995. Pragmatic relations and word order in Chinese. In *Word Order in Discourse*, ed. Pamela Downing and Michael Noonan, 297–330. Amsterdam and Philadelphia: John Benjamins Publishing Company.
- LaPolla, Randy. 2009. Chinese as a topic-comment (not topic-prominent and not SVO) language. In *Studies of Chinese Linguistics: Functional Approaches*, ed. Janet Zhiqun Xing, 9–22. Hong Kong: Hong Kong University Press.
- Li, Charles N., and Sandra A. Thompson. 1974. Historical change of word order: A case study in Chinese and its implications. In *Historical linguistics*, ed. John Mathieson Anderson and Charles Jones, 199–217. Amsterdam: North-Holland.
- Li, Charles N., and Sandra A. Thompson. 1975. The semantic function of word order: A case study in Mandarin. In *Word Order and Word Order Change*, ed. Charles N. Li, 163–195. Austin: University of Texas Press.
- Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley and Los Angeles: University of California Press.
- Liu, Feng-hsi. 1999. Structure-preservation and transitivity. In *Functionalism and Formalism in Linguistics II: Case Studies*, ed. Michael Darnell, Edith A. Moravcsik, Michael Noonan, Frederick J. Newmeyer, and Kathleen Wheatley, 175–202. John Benjamins Publishers.
- Liu, Feng-hsi. 2007. Word order variation and ba sentences in Chinese. *Studies in Language* 31 (1): 649–682.
- Matthews, Stephen, and Y.Y. Yeung. 2000. Processing motivations for topicalization in Cantonese. In *Cognitive-functional linguistics in an East Asian context*, ed. Kaoru Horie and Shigeru Sato, 81–102. Tokyo: Kurocio Publishers.
- McDonald, Janet L., Kathryn Bock, and Michael H. Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology* 25 (2): 188–230.
- Mei, Kuang 梅廣. 1978. The ba-sentences in modern Chinese 把字句. *NTU Humanities Taiwanica 臺大文史哲學報* 27: 145–180.
- Stallings, Lynne M., and Maryellen C. MacDonald. 2011. It’s not just the “heavy NP”: Relative phrase length modulates the production of heavy-NP shift. *Journal of Psycholinguistic Research* 40 (3): 177–187.
- Stallings, Lynne M., Maryellen C. MacDonald, and Padraig G. O’Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language* 39 (3): 392–417.
- Starr, Rebecca L. 2015. Predicting NP forms in vernacular written Cantonese. *Journal of Chinese Linguistics* 43 (1A): 54–89.
- Sun, Chaofen. 1995. Transitivity, the ba construction and its history. *Journal of Chinese Linguistics* 23 (1): 159–195.
- Sun, Chaofen, and Talmy Givón. 1985. On the so-called Sov word order in Mandarin Chinese: A quantified text study and its implications. *Language* 61 (2): 329–351.
- Team, R. Core. 2017. *R: A language and environment for statistical computing*. Vienna: R Foundation for statistical computing URL <https://www.R-project.org/>.
- Thompson, Sandra A. 1973. Transitivity and some problems with the Ba construction in Mandarin Chinese. *Journal of Chinese Linguistics* 1 (2): 208–221.
- Tily, Harry, Susanne Gahl, Inbal Arnon, Neal Snider, Anubha Kothari, and Joan Bresnan. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1 (2): 147–165.
- Tsao, Feng-fu. 1987. A topic-comment approach to the Ba construction. *Journal of Chinese Linguistics* 15 (1): 1–54.
- Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in late modern English: Exploring cross-constructional variation and change. *Diachronica* 30 (3): 382–419.
- Xu, Liejiong. 1995. Definiteness effects on Chinese word order. *Cahiers de Linguistique - Asie Orientale* 24 (1): 29–48.
- Xu, Liejiong. 2004. Manifestation of informational focus. *Lingua* 114 (3): 277–299.
- Yamashita, Hiroko, and Franklin Chang. 2001. “Long before short” preference in the production of a head-final language. *Cognition* 81 (2): B45–B55.
- Yao, Yao. 2014. Predicting the use of BA construction in Mandarin Chinese discourse: A modeling study with two verbs. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC28)*. Phuket.
- Yao, Yao, and Feng-hsi Liu. 2010. A working report on statistically modeling dative variation in Mandarin Chinese. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 1236–1244. Beijing.