Predicting the use of BA construction in Mandarin Chinese discourse: A modeling study with two verbs

Yao Yao
Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
Hong Kong
ctyaoyao@polyu.edu.hk

Abstract

This paper investigates the use of BA construction in Mandarin Chinese discourse with two frequently-occurring Chinese verbs 放 fàng (“v. to put”) and 拿 ná (“v. to take”). Previous literature suggests that the use of BA construction is influenced by a number of factors, including semantic meaning of the verb phrase and prominence and weight of the object NP. However, what is unclear is how these factors work together in conditioning word order variation (BA vs. SVO) in real context, especially regarding the effects of object NP prominence and weight and their interaction. In this study, we explore this issue by building corpus-based statistical models for predicting the use of BA construction in context. Our results show that for both verbs 放 fàng and 拿 ná, the use of BA construction is sensitive to the prominence (especially givenness) and weight of the object NP as well as structural parallelism, while no interaction effects were found. Furthermore, the weight effects are in opposite directions in the two models, raising new questions regarding the nature of heavy NP shift and revealing a great degree of cross-verb differences in word order variation.

1 Introduction

One of the most well-documented syntactic structure in Mandarin Chinese, the BA construction features an SOV word order with a preposed object noun phrase. An example of BA construction is shown in (a), with a corresponding sentence in canonical SVO word order shown in (b).

(a)	我	把	饭	吃完	了。
	I	BA	rice	eat-finish	ASP
	I ate the rice.				

(b)	我	吃完	了	饭。
	I	eat-finish	ASP	rice
	I ate the rice.			

A voluminous body of literature has been devoted to the study of the structural properties and historical development, as well as the semantic conditions for appropriately using BA construction. However, few studies have looked at the use of BA construction in context (see Liu 2007 for an exception). As Sun and Givón (1985) pointed out, the canonical SVO word order was used overwhelmingly in natural context, even in cases where the conditions for using BA construction were met. How to account for the variation between BA construction and canonical SVO in context then? Is there any general rule that can predict the use of BA construction in context or is the variation completely random and unpredictable? These are the questions that motivate the current research.

In this study, we model the use of BA construction and the alternative SVO construction in naturally produced discourse, using data from a large-scale Chinese corpus. The overarching goal of this research is to gain a comprehensive picture on the actual use of BA construction and

Copyright 2014 by Yao Yao

to unveil the mechanism of interactions between form, meaning, and context.

There are two lines of previous work that are relevant for this research: The first line concerns the linguistic properties of BA construction and the second line statistical modeling of syntactic variation. We will briefly review previous works along these two lines in the following section.

2 Background literature

2.1 Previous work on the use of BA construction

Numerous studies have examined when it is acceptable to use BA construction. The most well-acknowledged conditions include high prominence of object NP (i.e. definite, specific, or generic) and disposal meaning of the sentence verb (Li and Thompson 1974, 1975, 1981, Xu 1995). It is for this reason that the use of BA construction is often associated with notions of topic (Givón 1978, Tsao 1987, etc) and verb transitivity (Hopper and Thompson 1980, Liu 1999, Sun 1995, Thompson 1973). However, although these conditions seem to promote the use of BA construction, it is not clear how effective they are. As Sun and Givón (1985) has shown, BA construction could hardly compete with canonical SVO word order in real usage, even when these conditions were met. A more recent study by Liu (2007) examined the use of BA construction (and other types of object preposing) in context. From a corpus of about 400,000 characters, Liu collected 456 "structurally interchangeable" sentences which, regardless of their surface word order, could be expressed in the alternative word order without changing the meaning. Based on this dataset, Liu found a significant interaction of information status and weight on object preposing. If the object NP carried old information, it was more likely to be preposed if it was short; contrarily, if the object NP carried new information, it was more likely to be preposed if it was long. In other words, the general tendency was for a preposed NP to be either discourse-old + short or discourse-new + long, and for a postverbal NP to be the opposite. However, Liu's study only considered information status and weight but not any other syntactic, semantic or discourse properties (e.g. structural parallelism). Furthermore, Liu's study did not distinguish sentences with different verbs, which may lead to two potential problems. First, since the use of verbs in natural context is highly unbalanced, it is possible that the general tendency seen in the overall dataset was in fact driven by only a few high-frequency verbs. Second, the distribution of verbs may very well vary across sentence types (BA vs. SVO), therefore, it is possible that the observed pattern reflects more of verb-specific idiosyncrasies regarding word order as opposed to other factors (such as information status and weight).

What we see as lacking in the current literature is a study of BA vs. SVO variation in natural context that (1) considers all related factors and (2) provides sufficient control for verb type. The current study used data from a well-annotated 10-million-word Chinese corpus. As a result, our dataset was big enough to allow a wide range of predictor factors to be considered and sentences of different verbs to be modeled separately. Before we introduce the modeling detail of the current study, we will first review some major relevant works on statistical modeling of word order variation in both Chinese and English and briefly introduce what previous works have found to be significant predictors for surface word order.

2.2 Previous work on statistical modeling of syntactic variation

Using corpus data and statistical modeling methods, Bresnan and colleagues (Bresnan 2007, Bresnan et al. 2007, Bresnan and Ford 2010, Tily et al. 2009, Wolk et al. 2011, etc) have successfully modeled English dative variation (e.g. *I gave John a book* vs. *I gave John a book*) and genitive variation (e.g. *John's book* vs. *the book of John*). The models showed that both variation phenomena were sensitive to a wide range of properties pertaining to different sentence components (e.g. semantic type of the verb, NP accessibility, pronominality, definiteness, syntactic complexity, etc) and context (e.g. presence of parallel structures). After being trained with large corpus datasets, model accuracy reached over 90% when predicting the surface dative/genitive form in unseen sentences. Bresnan et al.'s research has been extended to other varieties of English (e.g. Australian English; Bresnan and Ford, 2010) as well as historical English (Wolk et al. 2011).

Similar modeling techniques have been applied in the investigation of syntactic variation in Chinese, too (e.g. Yao and Liu 2010, Starr under review). Yao and Liu (2010) examined Chinese dative variation in written texts. Two statistical models were constructed to investigate the three-way contrast among Chinese dative constructions (PREVERBAL: 我把书送给小王; POSTVERBAL DATIVE: 我送书给小王; POSTVERBAL DOUBLE OBJECT: 我送小王书). The models were overall quite successful in predicting surface form (accuracy > 87%). However, in the model for preverbal-postverbal variation, in contrast with the proposal in Liu (2007), Yao and Liu (2010) did not find a significant interaction between information status and weight on object preposing. Instead their results suggested a weak weight effect, as heavy direct object NPs were slightly more likely to be preposed than light direct object NPs. The discrepancy between Yao and Liu (2010) and Liu (2007) provides another motivation for reexamining the word order variation regarding BA and SVO constructions.

3 Methods

3.1 Data

This study uses data from the Academia Sinica Balanced Corpus of Modern Chinese (Version 5.0; *Sinica 5.0* for short; (Chen et al., 1996)), which contains about 10 million words of text (both spoken and written) and has been tagged with part of speech. To compile a dataset, we first created an initial list of BA sentences by locating instances of the BA markers (either 把 bǎ or 将 jiāng tagged as preposition) in the corpus. The initial list contained more than 11 thousand BA sentences of over 1600 different verbs, among which the five most frequent verbs were 放 *fàng* "v. to put", 当 *dāng* "v. to consider as", 带 *dài* "v. to bring", 送 *sòng* "v. to give", 拿 *ná* "v. to take". In this study, we chose to focus on 放 *fàng* and 拿 *ná* because (1) 带 (dài) and 送 (sòng) are both typical transfer verbs and are therefore involved in the complex three-way variation of dative constructions; (2) preliminary corpus analysis suggests that the verb 当 (dāng) is predominantly used in BA construction with little variation in word order.

The next step was to construct a comparable set of SVO sentences of the two target verbs (放 *fàng* and 拿 *ná*). Since the key was to find structurally interchangeable sentences, we narrowed the scope of search to sentences that not only contained a target verb but also an aspect marker (e.g. 了, 过, 着) or verb complement (e.g. 上, 下, 来, 去, 回, 完) that has been used in at least one BA sentence with the same target verb and an explicit object noun phrase (NP). As pointed out in previous studies, an important condition for using BA construction is the disposal meaning of the sentence, which is often expressed by aspect markers and verb complements in Mandarin Chinese (for examples, the verb 吃 "v. to eat" in (a) and (b) is followed by a complement 完 "finish"). In fact, sentences with bare verbs were nearly extinct in the BA sentence set of the current study.

Both the BA sentence set and the SVO sentence set of the two target verbs were manually checked and pruned for false hits, corpus errors and verb+aspect/complement combinations with non-alternating word order. Furthermore, since the verb 放 *fàng* can also mean "v. to release" and "v. to emit (light, electricity, etc.)", we further restricted the 放 *fàng* sentences to only those with the basic meaning "v. to put". The final dataset for 放 *fàng* includes 688 BA sentences and 320 SVO sentences, and the final dataset for 拿 *ná* includes 261 BA sentences and 727 SVO sentences.

3.2 Statistical models

All sentence tokens in the dataset were annotated for 16 properties: genre (`Genre`), language mode (`Mode`), adverbial phrase before the target verb (`AdvP_before`), another verb phrase before the target verb (`VP_before`), another verb phrase after the target verb (`VP_after`), target verb phrase embedded in a relative clause (`RelClause`), target verb phrase embedded in an adverbial phrase (`AdvClause`), target verb phrase is nominalized (`Nominalization`), a BA construction is used in previous context (`BA_before`), a BA construction is used in following context (`BA_after`), object of the target verb is mentioned in previous context (`ObjMentioned_before`), object of the target verb is mentioned in following context (`ObjMentioned_after`), object of the target verb contains a demonstrative pronoun 这 *zhé* “this” or 那 *nà* “that” (`ObjDemonstrative`), object of the target verb is animate (`ObjAnimacy`), object of the target verb is a pronoun (`ObjPronoun`), length of object NP (`ObjLen`).

Apart from `Genre` and `Mode`, which were already annotated in the corpus, all other properties were annotated manually by a trained linguist. Previous and following contexts were defined as 10 sentences (delimited by comma, full stop, exclamation mark or question mark) before or after the target verb phrase. `ObjLen` was counted by the number of Chinese characters or syllables, in case the object NP contained foreign words (e.g. code-switching). Since raw `ObjLen` was always greater than 1 and resembled a Zipfian-like distribution in our dataset, we centered and log-transformed `ObjLen` before entering in the models.

Word order variation in sentences of 放 *fàng* and 拿 *ná* were modeled separately in two generalized mixed-effects regression models. The two models had highly similar model structures. Both models contained a binary variable `SurfaceWordOrder` (BA=1; SVO=0) as the outcome variable, the set of annotated features described above as fixed effects and verb+aspect/complement as random effects, to control for individual differences of aspect markers and verb complements regarding surface word order. Since previous studies suggested that object NP weight might have a non-linear effect on word order variation and may work in interaction with information status, we also included a quadratic term of `ObjLen` as well as the interaction of `ObjLen` (both linear and quadratic terms) and `ObjMentioned_before` in the models. Table 1 shows a complete list of model terms.

Fixed-effect predictor	**Variable type**
`Genre`	Categorical
`Mode`	Categorical
`AdvP_before`	Boolean
`VP_before`	Boolean
`VP_after`	Boolean
`RelClause`	Boolean
`AdvClause`	Boolean
`Nominalization`	Boolean
`BA_before`	Boolean
`BA_after`	Boolean
`ObjMentioned_before`	Boolean
`ObjMentioned_after`	Boolean
`ObjDemonstrative`	Boolean
`ObjAnimacy`	Boolean
`ObjPronoun`	Boolean
centered(log(`ObjLen`))	Numeric
centered(log(`ObjLen`^{2})	Numeric

Table 1: Fixed-effects predictors in the initial models

After the initial construction, both models were submitted backward elimination where non-significant predictors (i.e. predictors whose elimination did not significantly affect model fit) were eliminated from the model, in order to avoid spurious effects due to the inclusion of non-significant predictor variables. Only results from the final models are reported in this paper. All models were constructed with the `lmer()` function in the `lme4` package (Bates and Maechler, 2011) of R (R Development Core Team, 2008).

4 Results

4.1 Modeling results of 放 *fàng*

The final model of 放 *fàng* contained 7 significant fixed-effect predictors. Table 2 below shows the model parameters. For simplicity, we only report the coefficient (β) of each term and the associated p value (i.e. $p(>|z|)$).

As shown in Table 2, everything else being equal, a verb phrase with 放 *fàng* is more likely

Predictor	β	p
(Intercept)	-0.27	.74
`RelClause` =T	-2.34	< .001
`BA_after`=T	0.81	.006
`ObjMentioned_before` =T	1.31	< .001
`ObjMentioned_after`=T	0.78	.015
`ObjDemonstrative`=T	1.22	.059
center(log(`ObjLen`))	-0.44	.031
center(log(`ObjLen`))^2	0.65	.0019

Table 2: Summary of fixed effects in the final model of 放 *fàng*.

to be used in a BA construction (than a SVO construction) when (1) the target verb is *not* used in a relative clause; (2) a BA construction is used in the following context; (3) the object NP is mentioned in the surrounding context (either before or after the current sentence); (4) the object NP contains a demonstrative (marginally significant); (5) the object NP is short. Overall the model correctly predicts (635+277)/1008 = 90.5% of the use of BA construction in context, a significant improvement compared with the baseline accuracy at (635+53)/1008 = 68.3%. Table 3 below shows the number of correct and incorrect predictions.

	Surface BA	Surface SVO
Predicted BA	635	43
Predicted SVO	53	277

Table 3: 放 *fàng* sentence counts by surface word order and predicted word order.

All the above effects were in the expected directions except for maybe the weight effects. As suggested in previous literature, the use of BA construction is promoted when the object NP is highly prominent, which is the case when the object NP contains a demonstrative - an explicit marker for definiteness in Mandarin Chinese - and when the NP is given in previous context or repeated in the following context (i.e. likely to be a discourse topic). Meanwhile, language users are also more likely to produce a BA construction when at least one BA construction has been produced in the near context, suggesting an influence of structural parallelism in word order variation. The effect of relativization might be due to elevated processing difficulty associated with BA construction in relative clause.

What is intriguing is the effect of object NP weight. The model of 放 *fàng* shows that object weight has a negative effect on the use of BA construction. The longer the object is, the **less** likely to observe a BA construction. Furthermore, the magnitude (i.e. the steepness) of this negative effect reduces as `ObjLen` increases, as suggested by the positive coefficient of the quadratic term `ObjLen`^{2}. But the model found no significant interaction between `ObjMentioned_before` and `ObjLen` or `ObjLen`^{2}.

4.2 Modeling results of 拿 *ná*

Results of the final model of 拿 *ná* sentences are shown in Table 4. Everything else being equal, a verb phrase with 拿 *ná* is more likely to be expressed in a BA construction (than a SVO construction) when (1) the target VP is used in an adverbial phrase (marginally significant); (2) a BA construction is used in the following context; (3) the object NP has been mentioned in previous context; (4) the object NP is long. Again, the model indicated effects of object NP prominence (`ObjMention_before`) and structural parallelism (`BA_after`). However, contrary to the model of 放 *fàng*, the model of 拿 *ná* shows a positive effect of weight. A BA construction is more likely to be used when the object NP is **long**. This finding is apparently more expected given the previous literature (Yao and Liu 2010), although still no interaction of weight and information status is found.

Predictor	β	p
(Intercept)	-0.80	.021
`AdvClause` =T	1.10	.055
`BA_after`=T	0.61	.002
`ObjMentioned_before` =T	1.60	< .001
center(log(`ObjLen`))	0.32	.012

Table 4: Summary of fixed effects in the final model of 拿 *ná*.

Overall model accuracy is (669+125)/988 =

80.4%, compared to a baseline accuracy at (58+669)/988 = 73.6% (see Table 5. The improvement is less significant than the model of 放, probably due to a higher baseline prediction accuracy in 拿 *ná* sentences.

	Surface BA	Surface SVO
Predicted BA	125	58
Predicted SVO	136	669

Table 5: 拿 *ná* sentence counts by surface word order and predicted word order.

4.3 Discussion

In this study, we built two statistical models for predicting word order variation between BA and SVO constructions, one for the verb 放 *fàng* and the other 拿 *ná*. The two models exhibit both similarity and differences. First of all, both models show significant effects of object NP prominence. The more prominent the object is (definite, given, topic, etc), the more likely it is to use a BA construction. Structural parallelism is another common effect shown in both models. When BA construction has already been used in the context, it is more likely to used again.

The differences between the two models are obvious, too. While both models show significant effects of object NP weight, the directions of the effects are opposite of each other. The model of 放 *fàng* shows a negative effect of object weight on the use of BA construction, with shorter object NPs being more likely to be preposed before the verb (or in other words, longer object NPs are more likely to appear with SVO word order). The model of 拿 *ná*, on the other hand, features a positive effect of object weight, with longer object NPs being more likely to be preposed.

The co-existence of heavy NP shift in both directions is not unheard of. Yao and Liu (2010) reported in their study of the three-way Chinese dative variation that longer direct object NPs were more likely to be preposed before the verb, yet also more likely to appear at the end of the sentence after both the verb and the indirect object NP. Liu (2007) associated object weight and information status and claimed that if the object NP is given in the context, it tends to be preposed if it is longer whereas if the object NP is new, it tends to be preposed if it is shorter. Although the weight×givennes effect has not found direct evidence in either Yao and Liu (2010) or the current study, it remains to be checked whether the lack of the interaction effect is due to unbalanced datasets or the presence of confounding factors.

Last but not the least, the current study reveals significant cross-verb differences in the use of BA construction. Apart from the opposite weight effects as discussed above, the two models are also different in the presence/absence of other significant predictors and the magnitude of the effects. To our best knowledge, such verb-specific variation patterns have rarely been reported in the study of word order variation. An important takeaway message for future studies on Chinese BA-SVO variation is to be more aware of the vast cross-verb differences, which may hold a key to the perplexing nature of the variation phenomenon.

Acknowledgments

Research reported in this paper was supported by the Early Career Scheme of Hong Kong RGC (Grant No. 558913).

References

Bates, D., and M. Maechler. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. <http://CRAN.R-project.org/ package=lme4>.

Bresnan, Joan. (2007) Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld (eds) *Roots: Linguistics in Search of its Evidential Base*, pp 77-96. Series: Studies in Generative Grammar. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni and Tatiana Nikitina and Harald Baayen. (2007) Predicting the dative alternation. In G. Boume et al. (eds) *Cognitive Foundations of Interpretation*, pp 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.

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. (1996). Sinica Corpus: Design Methodology for Balanced Corpora. In B.-S. Park and J.B. Kim (eds) *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*. Seoul:Kyung Hee University. pp.167-176.

Givón, Talmy. (1978) Definiteness and referentiality. In Joseph H. Greenberg (eds) Universals of Human Language, Vol. 4: Syntax, pp291-330. Stanford: Stanford University Press.

Hopper, Paul J. and Sandra Thompson. (1980) Transitivity in grammar and discourse. *Language*, 56, 251-299.

Li, Charles N. and Sandra A. Thompson. (1974) Historical change of word order: A case study in Chinese and its implications. In John Anderson and Charles Jones (eds) *Historical Linguistics*, pp199-217. Amsterdam: North-Holland.

Li, Charles N. and Sandra A. Thompson. (1975) The semantics function of word order: A case study in Mandarin. In Charles Li (eds) *Word Order and Word Order Change*, pp163-195. Austin: University of Texas press.

Li, Charles N. and Sandra A. Thompson. (1981) *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Liu, Feng-hsi. (1999) Transitivity and structure preservation. In Michael Darnell et al. (eds) *Functionalism and Formalism in Linguistics II*, Case Studies, pp175-202. John Benjamins Publishers.

Liu, Feng-hsi. (2007) Word order variation and ba sentences in Chinese. *Studies in Language*, 31(3), 649 - 682.

R Development Core Team (2008). R: A language and environment for statistical computing. Vienna, Austria. ISBN: 3-900051-07-0. <http://www.R-project.org>.

Sun, Chaofen. (1995) Transitivity, the ba construction and its history. *Journal of Chinese Linguistics*, 23, 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.

Thompson, Sandra A. (1973) Transitivity and the ba construction in mandarin Chinese. *Journal of Chinese Linguistics*, 1, 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, 1-54.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach and Benedikt Szmrecsányi. (2011) 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.

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)*. Beijing, China.