A Working Report on Statistically Modeling Dative Variation in Mandarin Chinese

Yao Yao
University of California, Berkeley
Department of Linguistics
yaoyao@berkeley.edu

Feng-hsi Liu
University of Arizona
Department of East Asian Studies
fliu@u.arizona.edu

Abstract

Dative variation is a widely observed syntactic phenomenon in world languages (e.g. *I gave John a book* and *I gave a book to John*). It has been shown that which surface form will be used in a dative sentence is not a completely random choice, rather, it is conditioned by a wide range of linguistic factors. Previous work by Bresnan and colleagues adopted a statistical modeling approach to investigate the probabilistic trends in English dative alternation. In this paper, we report a similar study on Mandarin Chinese. We further developed Bresnan et al.'s models to suit the complexity of the Chinese data. Our models effectively explain away a large proportion of the variation in the data, and unveil some interesting probabilistic features of Chinese grammar. Among other things, we show that Chinese dative variation is sensitive to heavy NP shift in both left and right directions.

1 Introduction

1.1 Overview

In traditional linguistic research, the study of syntax is most concerned with grammaticality. Sentences are either grammatical or ungrammatical, and syntactic theories are proposed to explain the structural features that cause (un)grammaticality. Meanwhile, little attention has been paid to the relative acceptability of grammatical sentences. If two sentences are both grammatical and basically express the same meaning, are they equally likely to occur in the language? The answer is probably *no*. For example, in English, the sentence *I have read that book* is much more frequent than *That book I have read*. The latter topicalized sentence is only used when the entity denoted by *That book* is in focus. This indicates that the choice of surface sentence form is not entirely random, but conditioned by some factors including information status.

Thus, instead of categorizing sentences as grammatical or ungrammatical, a better way to express the degree of grammaticality would be to use a likelihood continuum, from 0 to 1, where ungrammatical sentences have zero likelihood and grammatical sentences fall somewhere between 0 and 1, with some being more likely than others. The idea of associating linguistic forms with various probabilities has been around for a while (see Jurafsky, 2003 and Manning, 2003 for an extensive review). Recent psycholinguistic research has shown that just like grammaticality, the likelihoods of sentence forms are also part of the user's linguistic knowledge. Sentences with high probabilities are in general easier to comprehend and produce, and their production is more prone to phonetic reduction (Bresnan, 2007; Gahl and Garnsey, 2004; Levy, 2008; among others). The famous example of garden path sentences also exemplifies the difficulty of comprehension in low-probability sentence forms.

If we accept the premise of probabilistic syntax, then an immediate question is what determines these probabilities. In the current work, we address this question by investigating a particular type of probabilistic phenomenon, i.e. dative variation in Chinese. We show that the probabilities of

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1236–1244,
Beijing, August 2010

various surface forms of Chinese dative sentences can be well estimated by a linear combination of a set of formal and semantic features.

The remainder of this paper is organized as follows. Section 1.2 briefly reviews previous work on English dative variation. Section 1.3 introduces dative variation in Chinese. Section 2 describes the dataset and the statistical models used in the current study. Section 3 presents modeling results, followed by a discussion in Section 4. Section 5 concludes the paper with a short summary. To preview the results, we show that dative variation in Chinese is more complicated than in English, in that it features two levels of variation, which exhibit different (sometimes even opposite) probabilistic patterns.

1.2 Dative variation in English

A dative sentence is a sentence that encodes a transfer event. Typical verbs of transfer in English include *give, send, mail*, etc. A characterizing property of transfer events is that they often involve two objects. In addition to the direct object (DO), the verb also takes an indirect object (IO) which usually denotes the recipient of the transfer action. For instance, in sentence 1a, the direct object is *a book* and the indirect object is *John*.

Cross-linguistically, it has been documented that many languages in the world have multiple syntactic forms for encoding the same transfer event (Margetts and Austin, 2007 , among others). In English, both 1a and 1b describe the same event, but 1a is a double object form (V IO DO) while 1b takes a prepositional phrase (V DO *to* IO).

(1) a. I gave John a book. → V IO DO
b. I gave a book to John. → V DO *to* IO

A number of conditioning factors have been identified for the alternation between the two surface forms. For instance, when the indirect object is a pronoun (e.g. *him*), it is more likely to have the double object form (i.e. *I gave him a book*) than the PP form (i.e. *I gave a book to him*). On the other hand, if the indirect object is a complex NP (with relative clauses), it tends to occur at the end of the sentence. Since most of these effects are subtle and often correlated with each other (e.g. definiteness, pronominality and syntactic complexity), investigating individual factors can give convoluted and unreliable results. To avoid this problem, many recent works in the field adopted a statistical modeling approach (Bresnan et al., 2007; Wasow and Arnold, 2003, among others). Instead of investigating separate factors, statistical models are built on large-scale datasets, using all potential conditioning factors to predict the surface form. In Bresnan et al. (2007), a dozen predictors relating to the verb (type of transfer event), the two object NPs (accessibility, pronominality, definiteness, syntactic complexity, etc), and the discourse (presence of parallel structures) were used to make the prediction. Using data input from 2,360 dative sentences from the Switchboard corpus, the model correctly predicted surface form in 97% of the sentences, which was a great improvement over the baseline prediction accuracy of 79% (i.e. the percentage of correct responses if the model knows nothing but which variant is more frequently used). It also showed that dative variation in English was indeed sensitive to all the predictors in the model.

1.3 Dative variation in Chinese

Dative variation in Chinese is much more complicated than in English. In addition to the two word orders that exist in English (2a, 2b), it is also common for direct object to appear before the verb, as in a BA construction or a topicalized sentence (2c). Besides, indirect object can also precede the verb, as shown in 2d. Another dimension of variation is in the use of coverbs *gei* and *ba*, both of which can be optional (2b, 2c; see Li and Thompson, 1981 for a detailed discussion on this), or replaced by other morphemes (*zhu, yu, jiang*, etc).

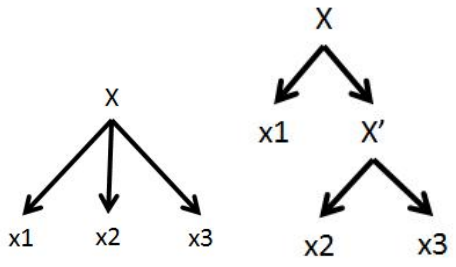
(2) a. John song-le shu gei Mary.
John give-ASP book to Mary
John gave one/some book(s) to Mary.
→ V DO IO
b. John song (gei) Mary yiben shu.
John gave (to) Mary one book
John gave Mary a book.
→ V IO DO
c. John ba shu song (gei) Mary, (ba)
John BA book gave (to) Mary (BA)

jiu song (gei) Kate.
wine gave (to) Kate
John gave the book(s) to Mary and gave the wine to Kate.
→ DO V IO

d. Ta meiren fa-le yiben shu.
He everyone allocated one book
He gave everyone a book.
→ IO V DO

For the purpose of the current study, we will ignore the existence (hence also the variation) of *gei* and *ba*, and concentrate on the variation in the relative order of V, DO and IO. In addition, our corpus search shows that sentences in the form of IO V DO are the least frequent (<9%) and mostly limited to a small set of verbs (mostly *fa* "to allocate" and *banfa* "to award"), so we drop this category from the current study. Thus the three remaining word order variants are: DO V IO, V DO IO, and V IO DO.

Generally speaking, there are two ways of modeling a variation phenomenon involving three variants. One way is to assume that the three variants are equally dissimilar from one another and the selection process is just to pick one out of three (Fig. 1a). The other approach is to assume a hierarchical structure: two of the variants are more similar to each other than they are to the third one and thus form a subcategory first before they join the third variant (Fig. 1b). In the selection process, the user first selects the subcategory (i.e. **x1** or **x'** in Fig 1b), and depending on which subcategory is chosen, they might need to make a second choice between two end nodes (i.e. **x2** and **x3**).



(a) (b)

Figure 1: Two possible schemas

We argue that the variation among the three word order variants in the current study is better modeled by a schema like Fig 1b, for both theoretical and methodological reasons. First, V DO IO and V IO DO are structurally more similar to each other than they are to DO V IO. Both V DO IO and V IO DO are in canonical word order of Chinese but the form DO V IO features the preposing (or topicalization) of the DO, whether or not the BA morpheme is present. Object preposing also exists outside ditransitive sentences (e.g. 3). Previous research has associated object preposing with the disposal meaning of the verb phrase, and the definiteness, givenness and weight of the object NP (Li and Thompson, 1981; Liu, 2007).

(3) a. Wo ba fan chi wan le.
I BA rice eat finish SEP
I have finished the rice.

b. Ta zhe dianying kan-le henduo bian.
he this movie saw many time
He has watched this movie for many times.

There is also a methodological motivation for adopting a hierarchical schema. Though it is not impossible to model a categorical variation with more than two variants (using multinomial logistic regression), binary variation is much easier to model and the interpretation of the results is more straightforward (this is especially true when random effects are present).

In view of the above, we propose the schema in Fig 2 for modeling the current variation phenomenon. We refer to sentences in the form of DO V IO as preverbal ditransitive sentences (since DO is before the verb), while both V DO IO and V IO DO are postverbal ditransitives. The distinction between the latter two forms regards whether DO is before or after IO, therefore one is termed as pre-IO and the other post-IO. Compared with the upper-level preverbal-postverbal distinction, the lower-level variation is much less studied in the literature (though see Liu, 2006 for a relevant discussion).

Corresponding to the schema in Fig 2, we constructed two separate models, one for the upper-level variation ("upper model") and the other for the lower-level variation ("lower model").

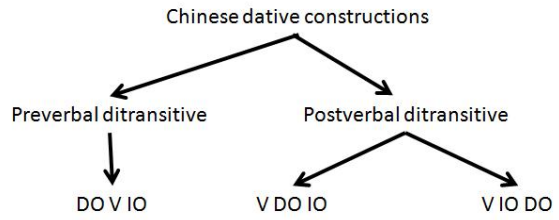


Figure 2: A two-level schema for Chinese dative variation

2 Methodology

2.1 Corpus and dataset

The data we use are from the Sinica Corpus of Modern Chinese (v3.1; Huang et al., 1995). We first compiled a list of 36 verbs that could be used ditransitively (see Appendix A) and then extracted from the corpus all sentences containing these words (n= 48,825 sentences). We then manually went through the sentences and selected those that (a) featured the ditransitive sense of the target verb, with both object NPs being overt, and (b) were in the form of any of the three form variants. 1,574 sentences remained after step (a) [1] and 1,433 after step (b) [2].

Further removal was conducted on verbs that were too sparse in the dataset. In each variation model, we removed verbs with fewer than two occurrences under either form variant. The final dataset for the upper model has 1149 sentences (of 20 verb types) while the dataset for the lower model has 801 sentences (of 14 verb types). The latter dataset is largely but not fully contained in the former due to the elimination of low-frequency verbs.

2.2 Data annotation

Similar to Bresnan et al.'s work on English, we annotated each data sentence for a wide range of features pertaining to the verb and the two NPs (see Appendix B for a complete list of annotated factors). Specifically, the verb was coded either as expressing a canonical transfer event, such as *ji* "to mail", or an extended transfer event, such as *jieshao* "to introduce". Semantic annotation of the two NPs is much trickier in Chinese than in English due to the lack of morphology. In practice, we used Bresnan et al.'s criteria for English, whenever applicable (e.g. accessibility, person, concreteness, animacy). In cases where the English rules did not apply (e.g. definiteness and number of bare NPs in Chinese), we developed working principles based on phrasal substitution. For example, if a bare NP can take a specifier like *yige/yizhi* "one" without changing sentence meaning, it is considered to be indefinite. Conversely, if a bare NP is better replaced with a full NP with a demonstrative *zhege* "this" or *nage* "that", it is coded as definite. Similar rules were used to assist annotating the number feature, using specifiers *yige/nage* "one"/"that" and *yixie/naxie* "some"/"those".

In addition to the factors in the English model, we also coded a set of structural features, including the presence of a following verb after the ditransitive construction, the presence of quantifiers/numerals in the NPs, and whether or not the ditransitive structure is embedded, nominalized, or relativized, etc. We suspect that since semantic features are often covert in Chinese words, it is possible that overt marking (e.g. the use of quantifiers/numerals) plays a more important role in conditioning surface form variation.

Finally we also included genre in the model. Sentences listed under the categories of dialogue and speech in the Sinica corpus were coded as "spoken" and the rest are coded as "written".

Altogether 24 factors were annotated and included in the statistical models as predictor variables. All variables are categorical except for the (log) length difference between DO and IO, which is numerical.

2.3 Statistical models

The statistical tool we use is mixed-effects logistic regression models. Compared with regular logistic regression models, mixed-effects models are more sophisticated in that they allow the user to specify factors that might introduce ran-

[1] A vast number of sentences were removed because the target verb was not used as a verb, or used with a different sense, or used as part of a different verb phrase, e.g. *fa* to allocate could also mean to bloom or be used in *fazhan* to develop, *faxian* to discover, etc.

[2] 141 sentences were removed because they were in the form of IO V DO.

dom variation in the dataset. In the current study, the datasets in both models contain sentences with different verbs. It is possible that different verbs have different intrinsic tendencies toward a certain word order variant.[3] Incorporating this piece of information into the model makes it more powerful and less affected by the unbalanced distribution of verb types. The mathematical formula of the mixed-effects logistic regression model is given below.

(4) Probability(V DO IO) = $\frac{1}{1+e^{-(\alpha_i + x\beta)}}$, where α_i is the verb-specific intercept of the verb v_i , x is a vector of predictors and β is a vector of corresponding coefficients.

Using the annotated datasets described in 2.2, we built an upper model and a lower model, corresponding to the schema in Fig 2. The general procedure of statistical analysis (which is the same for both models) is described as follows.

We first run the model with all 24 predictors, which will generate a coefficient and a p value for each predictor. Then we refit the model with only significant predictors (i.e. $p < 0.05$). The purpose of doing so is to filter out the noise in the model fit created by the large number of insignificant predictors. Only predictors that remain significant in the simplified model with largely unchanged coefficients are considered to be reliably significant.

Two model evaluation techniques are used to check the model results: cross-validation and separate analysis of high-frequency verbs. A potential problem in any statistical model is that it might overfit the data. After all, what we are interested in is the general probabilistic trends in dative variation, not the trends in a particular set of sentences featuring a particular set of verbs. A cross-validation test helps us evaluate the generalizability of model results by running the same model on a randomly sampled subset of the data. In doing so, it simulates the effect of having different datasets. In practice, we use two types of cross-validation procedures: one randomly samples sentences and the other samples verbs. Each procedure is executed on 100 randomly sampled subset of half the sentences/verbs. Only predictors with consistent performance over all iterations in both tests will be considered as stable.

Another concern in the model design is the effect of verb frequency. In the current dataset, one verb, i.e. *tigong* "to provide", is extremely frequent. 37.3% of the sentences in the upper model and 50.9% in the lower model come from this verb. Though in theory, verb frequency is already taken care of by using mixed-effects models and running cross-validation on samples of the verb set, it is still necessary to test *tigong* separately from the rest of the verbs, due to its extremely high frequency. In the next section, we will report in detail the results from the two regression models.

3 Results

3.1 Upper model: predicting preverbal and postverbal variation

In the upper model, the distinction is between preverbal (DO V IO; coded as 1) and postverbal ditransitives (V DO IO and V IO DO; both coded as 0). The dataset in this model contains 1,149 sentences (of 20 verb types), with 379 preverbal and 770 postverbal. The distribution of the verbs is highly skewed. The most frequent verb is *tigong* "to provide" (n=428 tokens), followed by *song* "to send" (135) and *jiao* "to hand; to transfer" (117). The remaining 17 verbs have between 5 and 54 occurrences in the dataset.

10 out of 24 predictors in the full model are significant and most of them remain significant when the other 14 predictors are removed from the model. Table 1 below summarizes the results of the simplified model.

Judging from the signs of the coefficients in Table 1, a dative sentence is more likely to take the preverbal form (as opposed to the postverbal form) when (a) the verb expresses canonical transfer event, (b) DO is definite, plural, abstract and given in the previous context, with no quantifiers or numerals, (c) IO is not a pronoun and is not given in the previous context, and (d) DO is longer

[3] The same can be said about individual speakers, as some speakers might be more inclined to use certain forms than other speakers. However, since the sentences in the current datasets were sampled from a vast pool of speakers/writers (given the way the corpus is developed), individual differences among speakers is not considered in the current model.

Predictor	β	p
verb is canonical	1.71	0.03
DO is given	1.22	<0.001
DO is definite	4.89	<0.001
DO is plural	1.4	<0.001
DO is concrete	-1.13	0.004
quan/num in DO	-0.99	0.005
IO is pronoun	-1.64	<0.001
IO is given	-0.9	0.007
quan/num in IO	1.32	0.07 (n.s.)
Len(DO)-Len(IO)	0.53	0.002

Table 1: Fixed effects in the simplified upper model

than IO.

Table 2 shows the accuracy of the simplified model. If 0.5 is used as the cut-off probability, the model correctly predicts for (737+338)/1149=93.6% of the sentences. For comparison, the baseline accuracy is only 770/1049=-67% (i.e. by guessing postverbal every time). In other words, the model only needs to include 10 predictors to achieve an increase of around 39% (93.6-67)/67) in model accuracy.

		Predicted	
		preverbal	postverbal
observed	preverbal	338	41
	postverbal	33	737

Table 2: Prediction accuracy of the simplified upper model

Results from the two cross-validation tests confirm all the predictors regarding DO in Table 1, as well as the pronominality of IO and the length difference between DO and IO. Verb category and the givenness of IO do not survive the cross-validation tests.

Separate analysis of tigong shows that indeed, the extremely high-frequency verb exhibits vastly different patterns than other verbs. Only one predictor turns out to be significant for *tigong* sentences, that is, the definiteness of DO (β = 6.17, $p < 0.001$). A closer look at these sentences suggests that they are strongly biased toward postverbal word order, in that 400 out of 428 (95.4%) *tigong* sentences are postverbal (compared with the average level of 67% in all sentences). In other words, just by guessing postverbal every time, one is able to make the correct prediction for *tigong* over 95% of the time. Not surprisingly, there is little need for additional predictors. For non-*tigong* sentences, all factors in Table 1 are significant except for verb category and the presence of quantifiers/numerals in IO. Overall, the non-*tigong* model has an accuracy of 91.5% (baseline = 50.6%).

To sum up, we are confident to say that the semantic features of DO, as well as pronominality of IO and the length difference between the two objects, play important roles in conditioning the preverbal-postverbal variation. Knowing these factors boosts the model s predicting power by a great deal.

3.2 Lower model: predicting pre-IO and post-IO variation

In the lower model, the distinction is between pre-IO sentences (i.e. V DO IO; coded as 1) and post-IO sentences (i.e. V IO DO; coded as 0). The dataset consists of 801 sentences of 14 verb types, among which 161 are pre-IO and 640 are post-IO. The most frequent verb is again, *tigong* (n=408 tokens), followed by *dai* "to bring" (137) and *song* "to send" (89).

Table 3 below summarizes the results of the simplified version of the lower model (constructed in the same fashion as described in Section 3.1).

Predictor	β	p
DO is definite	1.59	0.006
DO is concrete	1.06	<0.001
DO is plural	-0.57	0.04
followed by a verb	2.29	<0.001
normalized verb phrase	1.36	0.13 (n.s.)
Len(DO) - Len(IO)	-1.37	<0.001

Table 3: Fixed effects in the simplified lower model

Compared to the upper model, fewer predictors are significant in the lower model. Everything else

being equal, a postverbal ditransitive sentence is more likely to take the pre-IO form (V DO IO) if (a) DO is definite and concrete, (b) IO is singular, (c) DO is shorter than IO, and (d) the ditransitive construction is followed by another verb. The last point is illustrated in sentence 5a, which is adapted from a real sentence in the corpus. In 5a, the NP *women* "we" is both the recipient of the first verb *song* "to send" and the agent of the second verb *chi* "to eat". Thus, by using a pre-IO form, the NP *women* is in effect adjacent to the second verb chi, which might give an advantage in sentence processing. Notice though, if the other form (V IO DO) is used, the sentence is still grammatical (see 5b).

(5) a. Ta hai song xiaoye gei wo chi.
he also sent snacks to me eat
He also sent snacks for me to eat.

b. Ta hai song (gei) wo xiaoye chi.
he also sent (to) me snacks eat
He also sent me snacks to eat.

Overall the lower model is not as successful as the upper model. The prediction accuracy is 87.7% (baseline accuracy is 79.9%; see Table 4).

		Predicted	
		pre-IO	post-IO
observed	pre-IO	85	76
	post-IO	22	618

Table 4: Prediction accuracy of the simplified lower model

Moreover, cross-validation and the analysis of *tigong* show that only two factors, the presence of the following verb and length difference, are stable across subsets of the data. In fact, with length difference alone, the model generates correct predictions for 86.8% of the sentences (only 1% less than the accuracy reported in Table 4).

However, before we hastily conclude that length difference is the only thing that matters in the lower-level variation, it is important to point out that when the length factor is removed from the model, some predictors (such as the accessibility of DO) turn out to be significant and the model still manages to achieve an accuracy of 85.3%. Therefore, a more plausible explanation is that length difference is the strongest predictors for lower-level dative variation. Though the part of variation it accounts for can also be explained by other predictors, it is more effective in doing so. Therefore the existence of this variable tends to mask other predictors in the model.

4 Discussion

4.1 Comparing he two models

In the current study, we propose a two-level hierarchical schema for modeling the variation among three major word orders of Chinese dative sentences. On the upper level, there is a distinction between sentences with preverbal DOs and those with postverbal DOs. On the lower level, among postverbal sentences, there is a further distinction between pre-IO sentences (i.e. with prepositional phrases), and post-IO sentences (i.e. double object forms). This schema is promoted by structural as well as methodological concerns.

Our modeling results show that the two levels of variation are indeed characterized by different probabilistic patterns, which in turn provide evidence for our original proposal. As presented in Section 3, the upper-level distinction is mostly conditioned by the semantic features of the DO. However, in the lower-level variation, the two best predictors are length difference and the presence of a following verb. Overall, the upper-level model is more successful (accuracy = 93.6%, baseline = 67%) than the lower-level model (accuracy = 87.7%, baseline = 79.9%).

A more striking difference between the two models is that they exhibit weight effects in opposite directions. In both models, length difference between DO and IO plays an important role. Nevertheless, in the upper model, length difference has a positive sign (β = 0.53), meaning that the longer the DO is (compared to the IO), the more likely it is to prepose DO before the verb. Conversely, in the lower-level model, this factor has a negative sign (β = - 1.37), which means that the longer the DO is (compared to the IO), the less likely it is for DO to be before IO. That is to say, everything else being equal, if a DO is long, it will probably be preposed before the verb, but if it is

already after the verb, then it will more likely be placed after IO, at the end of the construction.

The difference in directionality explains why it is only in the lower-level model that the weight effect overshadows other predictors. Features like pronominality, definiteness, and accessibility are inherently correlated with weight. Pronouns are shorter than full NPs; definite NPs tend to be shorter than indefinite NPs (which often take quantifiers and numerals); NPs that have appeared before tend to be in shorter forms than their first occurrences. In both models, a general trend is that NPs that are more prominent in the context (e.g. pronouns, definite NPs, NPs with antecedents) tend to occur earlier in the construction. Thus, in the lower model, the general trend of prominence is confluent with the short before long weight effect, but in the upper model, it is pulling away from the long before short weight effect. As a result, weight effect only masks semantic predictors in the lower model, not in the upper model.

4.2 Comparing with English dative variation

Compared with Bresnan et al.s models, the current results reveal a number of interesting differences between Chinese and English dative variation.

First, the variation phenomenon in Chinese involves at least one more major variant, that is, the preverbal word order, which significantly increases the complexity of the phenomenon. The fact that overall the English model has greater prediction accuracy than the Chinese models might have to do with the fact that the variation phenomenon is more complicated and harder to model in Chinese.

Second, dative variation in Chinese seems to be less sensitive to semantic features. If we only consider the lower-level variation in Chinese, which involves the same form variants as in English (i.e. V DO IO and V IO DO), the Chinese model is best predicted by the length difference between DO and IO and most other predictors are muted by the presence of this factor. In the English model, semantic features are still significant even when length difference is controlled.

Last but not least, as discussed at length in the previous section, the two levels of dative variation in Chinese exhibit weight effects in opposite directions. The English variation is also sensitive to weight, but only in the short before long direction, which is the same as the lower-level variation in Chinese.

5 Conclusion

In this work, we present a corpus-based statistical modeling study on Chinese dative variation. In doing so, we show that this new methodology, which combines corpus data and statistical modeling, is a powerful tool for studying complex variation phenomena in Chinese. The statistical models built in the current study achieve high accuracy in predicting surface forms in Chinese dative sentences. More importantly, the models unveil probabilistic tendencies in Chinese grammar that are otherwise hard to notice.

A remaining question in the current study is *why would Chinese dative variation exhibit weight effects in both directions*. The answer to this question awaits further investigation.

Acknowledgement

We would like to thank three anonymous reviewers for helpful comments on an earlier version of the paper. We owe special thanks to Joan Bresnan and her colleagues in the Spoken Syntax Lab at Stanford University, for sharing working manuals and for valuable discussions.

References

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Featherston, S. and Sternefeld, W., editors, *Roots: Linguistics in search of its evidential base*, Studies in generative gramar, pages 77–96. Mouton de Gruyter, Berlin.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, H. (2007). Predicting the dative alternation. In Boume, G., Kraemer, I., and Zwarts, J., editors, *Cognitive foundations of interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.

Gahl, S. and Garnsey, S. (2004). Knowledge of grammar, knowledge of usage: Syntactic prob-

abilities affect pronunciation variation. *Language*, 80(4):748–775.

Huang, C., Chen, K., Chang, L., and Hsu, H. (1995). An introduction to Academia Sinica Balanced Corpus. [in chinese]. In *Proceedings of ROCLING VIII*, pages 81–99.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, J. H. and Jannedy, S., editors, *Probabilistic Linguistics*, pages 39–96. MIT Press, Cambridge, Massachusetts.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Li, C. N. and Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press, Berkeley.

Liu, F. (2006). Dative constructions in Chinese. *Language and Linguistics*, 7(4):863–904.

Liu, F. H. (2007). Word order variation and ba sentences in Chinese. *Studies in Language*, 31(3):649 – 682.

Manning, C. D. (2003). Probabilistic syntax. In Rens Bod, J. H. and Jannedy, S., editors, *Probabilistic Linguistics*, pages 289–341. MIT Press, Cambridge, Massachusetts.

Margetts, A. and Austin, P. (2007). Three participant events in the languages of the world: toward a cross-linguistic typology. *Linguistics*, 45(3):393–451.

Wasow, T. and Arnold, J. (2003). Post-verbal constituent ordering in english. In Rohdenburg, G. and Mondorf, B., editors, *Determinants of Grammatical Variation in English*, pages 119–154. Mouton.

Appendices

A Complete verb list [4]

song “to send”, *tigong* “to provide”, *jie* “to lend (to)”, *fu* “to pay”, *ban* “to award”, *banfa* “to award”, *zengsong* “to send (as a gift)”, *shang* “to award”, *jieshao* “to introduce”, *huan* “to return”, *fa* “to distribute/allocate”,*jiao* “to transfer”, *ji* “to mail”, *liu* “to leave (behind)”, *liuxia* “to leave (behind)”,*reng* “to throw”, *diu* “to throw”, *diuxia* “to throw (behind)”, *juan* “to donate”, *juanzeng* “to donate”, *juanxian* “to donate”, *bo* “to allocate”, *di* “to hand (to)”, *zu* “to rent (to)”, *fen* “to distribute”, *na* “to hand (to)”, *dai* “to bring”, *dailai* “to bring”, *jiao* “to teach”, *chuan* “to deliver”, *chuanran* “to pass around (a disease)”, *chuanda* “to deliver (a message)”, *chuansong* “to deliver” , *chuanshou* “to deliver (knowledge)”,*ci* “to give (as a reward)”, *pei* “to pay (compensation)”

[4] The verb *gei* “to give” is not included in the list, because it has the same form as the coverb gei and therefore has different properties than other ditransitive verbs. Among other things, the verb *gei* cannot take the V DO IO form in Mandarin (e.g. **gei yiben shu gei wo* “give a book to me”).

B Predictors in the full model

Predictor	**Coding**
genre	1=spoken; 0=written
verb category	1=canonical transfer; 0=otherwise
definiteness of DO	1=definite; 0=indefinite
pronominality of DO	1=pronoun; 0=otherwise
number of DO	1=plural; 0=singular
person of DO	1=1st and 2nd person; 0=otherwise
concreteness of DO	1=concrete; 0=abstract
givenness of DO	1=given; 0=otherwise
quan/num in DO	1=yes; 0=no
definiteness of IO	1=definite; 0=indefinite
pronominality of IO	1=pronoun; 0=otherwise
number of IO	1=plural; 0=singular
person of IO	1=1st and 2nd person; 0=otherwise
concreteness of IO	1=concrete; 0=abstract
givenness of IO	1=given; 0=otherwise
followed by another verb	1=yes; 0=no
embedded under another verb	1=yes; 0=no
part of a copular sentence	1=yes; 0=no
adverbial phrase after the verb	1=yes; 0=no
particle after the verb	1=yes; 0=no
question form	1=yes; 0=no
sentence negation	1=yes; 0=no
relativization	1=yes; 0=no
nominalization	1=yes; 0=no
log(len(DO)- log(len(IO))	numerical