A Statistical Comparison of Some Theories of NP Word Order

Richard Futrell, Roger Levy, and Matthew Dryer

June 25, 2017

Abstract

A frequent object of study in linguistic typology is the order of elements {demonstrative, adjective, numeral, noun} in the noun phrase. The goal is to predict the relative frequencies of these orders across languages. Here we use Poisson regression to statistically compare some prominent accounts of this variation. We compare feature systems derived from Cinque (2005) to feature systems given in Cysouw (2010) and Dryer (in prep). In this setting, we do not find clear reasons to prefer the model of Cinque (2005) or Dryer (in prep), but we find both of these models have substantially better fit to the typological data than the model from Cysouw (2010).

1 Introduction

A frequent object of study in linguistic typology is the variation in the order of elements inside the noun phrase (NP) across languages. In particular, much work has focused on predicting the relative frequencies across languages of orders of the elements {demonstrative, adjective, numeral, noun}. Table 1 shows the relative frequencies of different orders for these elements across languages (assuming each language exhibits only one dominant order) according to data given in Dryer (in prep). In this table, D stands for demonstrative, N stands for numeral, A stands for adjective, and n stands for noun. Genera counts are the counts of linguistic genera showing a certain order; adjusted frequencies are calculated using a methodology described in Dryer (in prep) intended to minimize any overrepresentation of some orders that may arise from areal effects.

Here we consider three proposals from the literature on how to explain these frequencies. We compare these proposals statistically in a log-linear

Order	Adjusted frequency	Genera count
$DNAn$	21.67	57
$nAND$	18.49	84
$DnAN$	16.65	38
$DNnA$	13.52	31
$NnAD$	9.00	28
$nADN$	7.94	19
$nDAN$	7.00	11
$DnNA$	7.00	10
$nNAD$	6.65	9
$NAnD$	4.00	5
$nDNA$	3.73	5
$DAnN$	3.40	8
$AnND$	3.00	3
$NnDA$	3.00	3
$NDAn$	3.00	3
$AnDN$	2.00	3
$DANn$	2.00	2
$nNDA$	1.00	1
$NADn$	0.00	0
$NDnA$	0.00	0
$ADnN$	0.00	0
$ADNn$	0.00	0
$ANDn$	0.00	0
$ANnD$	0.00	0

Table 1: NP orders and their adjusted frequencies across languages and their counts across genera, as given in Dryer (in prep).

framework based on how well they can predict the typological data given in Table 1.

We consider three proposals from the literature: those given in Dryer (in prep), Cysouw (2010) and Cinque (2005). The proposal in Dryer (in prep) is an update from the previous proposal of Dryer (2006). The first two of these theories are featural in nature: they associate each order with a set of marked features, and claim that orders with more marked features will be less frequent. The last model, that of Cinque (2005), is derivational in nature: it gives a generative model for how certain word orders arise, where certain decisions in the generative process are considered marked. Orders that require more marked operations to be generated are claimed to be less frequent. We reduce the last model to a featural model, and then compare which model provides a feature system which can best predict the typological data when the features have different degrees of markedness.

2 Method

2.1 Basics

We consider each proposal from the literature to define a feature system, and compare the ability of each feature system to predict the observed frequencies of orders. To do so, we use Poisson regression, as first used in Cysouw (2010). In Poisson regression we represent each language with a set of m binary-valued features, and say that the expected frequency λ of a language in a sample of k languages is given by:

$$\begin{aligned} \lambda &= e^V \\ V &= w_b + w_1 \cdot f_1 + w_2 \cdot f_2 + ... + w_m \cdot f_m, \end{aligned} \tag{1}$$

where f_i is an indicator variable with value 0 when the ith feature is $-$, and 1 when the ith feature is $+$, and where the weights w_b and $w_1, ..., w_m$ are those that maximize the probability of the observed counts of languages. Under the probabilistic model of Poisson regression, the probabability that a language with feature values $f_1, ..., f_m$ has frequency F is:

$$p(F|f_1, ..., f_m) = \frac{\lambda^F e^{-\lambda}}{F!}. \tag{2}$$

Fitting a Poisson regression model using a certain set of features to a set of (possibly adjusted) frequency tells us how well it is possible to predict languages in this framework given that set of features. Feature weights may

be negative, in which case they can be considered *marked*. In that case, the model embodies the claim that the presence of these features is disfavored in languages. Since features get different weights, the model implements different degrees of markedness per feature, as in Harmonic Grammar (Smolensky and Legendre, 2006). The model finds the degrees of markedness which best predict the data given the feature system.

2.2 Feature systems under comparison

Within this framework, we compare three feature systems: (1) the system in Dryer (in prep), (2) the system in Cysouw (2010), and (3) the theory of Cinque (2005). Cinque's theory is not phrased in terms of features, so we use two reductions of his theory to features: those presented in Merlo (2015) and our own, shown in Figure 5. Our featurization of Cinque's theory closely parallels the featurization given in Cysouw (2010).

2.3 Dependent variables

We apply Poisson regression to predict two quantities. First, we try to predict the *adjusted frequency* of each order, as given in Dryer (in prep) and shown in Table 1. (We round the adjusted frequencies to the nearest integer in order to satisfy the Poisson regression assumption that the dependent variable is a natural number.) Second, we try to predict the counts of genera given in the same paper.

2.4 Basis for model comparison

We compare models using log likelihood, the log probability assigned to the observed frequencies under the model. A model fits the data well when it assigns high probability to the data, so high log likelihood indicates a good fit. When log likelihood for different models is close, we can also compare them by their degrees of freedom, which is the number of free parameters in the model. In general, simpler models with fewer parameters are preferable over ones with more parameters.

2.5 Notes on Featurization of Cinque (2005)

Special care is needed when reducing the theory of Cinque (2005) to features so that it can be compared with the other theories in a regression framework.

The theory of Cinque (2005) is not featural in nature, but rather derivational. In this model, orders are built up by a generative process that makes

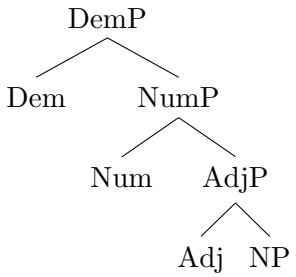

Figure 1: The universal base structure of D,N,A,n under the theory of Cinque (2005).

decisions in a certain order; whereas in featural models, orders are assigned scores based on features that have no intrinsic order. For example, the centerpiece of the theory of Cinque (2005) is the claim that the Merge order of D>N>A>n is universal, and that the Linear Correspondence Axiom (LCA) of Kayne (1994) holds, such that every word order must be generated from a base structure of the form seen in Figure 1, plus movement operations. Orders that are not derivable from this base structure are not generated at all in Cinque's theory. For such orders, the question of what if any movement operations apply never even arises under Cinque's theory, in principle.

For example, suppose we think of the Cinque model in terms of features: then `AlternativeMergeOrder` is a feature that can be + or −, and some movement operation is reflected in a feature that can be + or −. For an order that violates the specified merge order, we would give it value + for `AlternativeMergeOrder` and − for the movement feature, but this is not a completely correct reflection of Cinque's derivational theory. The reason is that under the generative process, if a word order violates the required merge order, then the model never even decides whether to perform a movement operation or not: thus the value of the movement feature should not be − or +, but undefined.

The theory of Cinque (2005) also involves theoretically-derived graded markedness values for operations in the generative process: for instance, total movement is claimed to be unmarked, "picture of who" type movement is claimed to be especially marked, and the rest of the movement features are claimed to be marked. In our methodology, we let the model decide on feature weights (markedness values) without regard to these a priori markedness values. As such, it is possible that our implementation of the Cinque model does not reflect its full intent.

The fact that our weights are derived through fits to the data and not

through a priori considerations is especially noteworthy in the case of the feature `AlternativeMergeOrder` in Cinque's system. In a literal reading of Cinque (2005), orders which violate the required merge order should occur with probability 0, and thus the feature `AlternativeMergeOrder` should be assigned infinitely negative weight. In that case the model would assign probability 0 to any data that has nonzero frequency for any such order. In our work we let the model learn that `AlternativeMergeOrder` has a large negative weight, without postulating that orders violating the required merge order must have probability 0.

Here we represent Cinque's model using features for the sake of convenience in statistical comparison, while noting that this introduces the issues above. And there is a further issue that should be noted. If Merge orders other than that seen in Figure 1 are allowed, then not only are otherwise impossible word orders generable by postulating that those word orders are generated as a base structure; additionally, word orders that were already generable under Cinque's theory also wind up with new derivations from different base structures. Taking this multiplicity of possible derivations into account increases the complexity of the problem of inferring feature weights, and takes it outside the scope of standard Poisson regression or other generalized linear models. This is a difficulty shared by the modeling approaches of Cysouw (2010) and Merlo (2015). For expediency, however, we follow previous work in treating every word order not derivable from the D>N>A>n base structure of Figure 1 as being derived by an alternative Merge order (so that `AlternativeMergeOrder =T`) with no movement, ignoring alternative derivations.

We also note that in formalizing Cinque's (2005) theory into a featural description, we noticed certain unclarities in the text which affected our formalization. These are as follows (note that Cinque uses **N** where we use **n**, and **Num** where we use **N**:

- Cinque describes order AnDN (his (k)) as involving two marked options: "derivation with raising of NP plus **pied-piping of the picture of who type** of the lowest modifier (A), followed by **raising** of [A N] **without pied-piping** around both Num and Dem" (emphasis ours). Technically speaking, it is not clear whether this latter raising (without pied-piping) should count as marked in his system, since the only relevant parameter of movement is "Movement of NP without pied piping" (his (7biii)), but this latter raising is movement of AP, not NP. Nevertheless we followed Cinque in assigning this word order a + value for movement (of NP) without pied piping. Additionally,

there is an inconsistency in Cinque (2005) between (7bv), where this order is stated to involve partial movement, and (6k), where partial movement is not listed as a type of markedness for this order. Here we went with (7bv) and listed this order as involving `partial_move=+`; we believe that this treatment is the most globally consistent overall, on analogy with orders such as NnAD which Cinque treats as involving partial movement because there are multiple types of movement and the first (raising of NP around A) is only partial.

- As with AnDN, there is inconsistency between (7bv) and the word-order-specific description (6w): in the former, this order is stated to involve partial movement of NP, but in the latter, partial movement is not mentioned as a type of markedness. As with AnDN, we listed this order as involving `partial_move=+`.

- Cinque describes order nDAN as involving "extraction of the sole NP around Dem" as the final movement in the derivation, an operation appealed to for no other word order and not unambiguously categorizable in his parameters of movement in (7b); Cinque does not list the number of marked options for this order. Cysouw (2010) dedicates a specific feature to this type of movement in his modeling effort; we treat it as a case of movement of NP without pied piping.

Regarding the first two cases, it should be emphasized that Cinque (2005) is far from totally clear about what does and does not count as partial (and thus marked) movement. For example, NnAD is described as involving partial (and thus marked) raising of NP around A, followed by a second raising that gets the raised constituent all the way to the left edge, but though nNAD likewise starts with a partial raising of NP around A (and N) followed by a second raising that gets the raised constituent all the way to the left edge, it is not considered to involve partial movement.

Additionally, there are two cases of what we believe are coding errors by Cysouw (2010) in his implementation of Cinque's model (see his Appendix on page 284): Cysouw encodes NnAD and nNAD as involving NP movement with pied piping of the *whose picture* variety, but Cinque describes these orders (his (6s) and (6t)) as involving *picture of who* pied piping instead, which seems correct to us.

2.6 Comparison with Merlo (2015)

Merlo (2015) conducts a study with similar aims to ours and uses featurizations more or less the same as what we've discussed above. She uses

features to predict frequency classes using a Naive Bayes estimator and an Weighted Averaged One-Dependence (WAODE) estimator, rather than Poisson regression as we use here and as was proposed by Cysouw (2010). As a summary of how this work: Merlo (2015) first discretizes the integer-valued word order frequency counts (by language or by genus) into 2, 4, or 7 categories; then she learns a model that categorizes language classes by their features according to the classic Naive Bayes formula:

$$P(\text{class}|\text{features}) \propto P(\text{features}|\text{class})P(\text{class})$$
$$P(\text{features}|\text{class}) = \prod_{f_i \in \text{features}} P(f_i|\text{class}),$$

where f_i is the value of the ith feature in the featurization scheme under consideration; our f_i here are Merlo's a_i in her Equations 1–4, p. 334. A word order is assigned to the frequency class that maximizes the probability $P(\text{class}|\text{features})$ above for that word order's features. Note that technically these "features" are attribute-value pairs, such as `Symmetry1=T` for the Dryer model or `Partial=whose-pp` for the Cinque model. The WAODE model is a bit more complex than the Naive Bayes model but is fundamentally similar.

Our work differs from Merlo's approach on two points:

- In predicting typological data, we use Poisson regression, which is a discriminative log-linear predictor, rather than Naive Bayes and WAODE, which are generative models.
- Our models predict integer-valued typological counts, whereas the models in Merlo (2015) predict unordered categorically-valued frequency classes.

We favor Poisson regression (and more generally log-linear models) over the Naive Bayes/WAODE approach because it allows us to predict more fine-grained typological data and to model the strong intuition that the effects of features on typological frequencies should be monotonic. In a model where the goal is to classify each language into the categories (e.g.) {Very Frequent, Frequent, Rare, None}, there is nothing to prevent a feature from getting weights that favor Very Frequent and None while disfavoring Frequent and Rare. Examples of this non-monotonicity in feature weights can be seen in Merlo's (2015) Table 11, our Table 2: the feature `Harmony=Y` favors a language to be either Very Frequent or None, while favoring Frequent and Rare less. The monotonicity in weights means that the weights from this framework cannot be considered markedness values, which either penalize

Probability	Value
P(`Harmony=Y`|Very Frequent)	= 0.99
P(`Harmony=Y`|Frequent)	= 0.16
P(`Harmony=Y`|Rare)	= 0.51
P(`Harmony=Y`|None)	= 0.55

Table 2: Table of feature weights (conditional probabilities under the Naive Bayes assumption) from Merlo (2015), Table 11

an order (make it less frequent) or do not. In addition to making the model weights less interpretable, this non-monotonicity means that the model has the flexibility to take advantage of artifacts of the discretization of word order frequencies into bins.

2.7 Comparison with Cysouw (2010)

As stated in Section 2.5, our approach here is very similar to that of Cysouw (2010): we use the same statistical model class and the same theory comparison (Cinque/Cysouw/Dryer). The differences are as follows:

- We use the more recent data of Dryer (in prep) rather than the earlier data of Dryer (2006);
- We use the feature set of Dryer (in prep) rather than the earlier feature set of Dryer (in prep);
- We correct what we believe are two featurization errors made by Cysouw in featurizing Cinque's theory, and we featurize order nDAN differently than Cysouw does (see above).

3 Results

The results do not give clear grounds for deciding between the Dryer model and the Cinque model, but both of these models come out better than the Cysouw (2010) model. Whether or not the Dryer model comes out better depends on whether we use the model to predict adjusted frequencies or genera counts.

3.1 Predicting Adjusted Frequencies

Table 3 shows log likelihoods for models predicting adjusted frequency (rounded to the nearest integer). It also shows the number of parameters (d.f.) in each

Model	Log likelihood	d.f.
Dryer (in prep)	-46.6	6
Cysouw (2010)	-60.1	5
Cinque (2005) (our features)	-44.2	7
Cinque (2005) (tied markedness)	-44.6	5
Cinque (2005) (Merlo's features)	-46.0	8

Table 3: Log likelihoods of *adjusted frequency* data under various models, and the degrees of freedom (d.f.) of those models.

model. The table shows that Cinque's model slightly outperforms Dryer (in prep) on fitting the data.

For a more detailed comparison of model performance, we compared model predictions to observed adjusted frequencies from Dryer (in prep). Figure 2 shows model predictions compared against adjusted frequency.

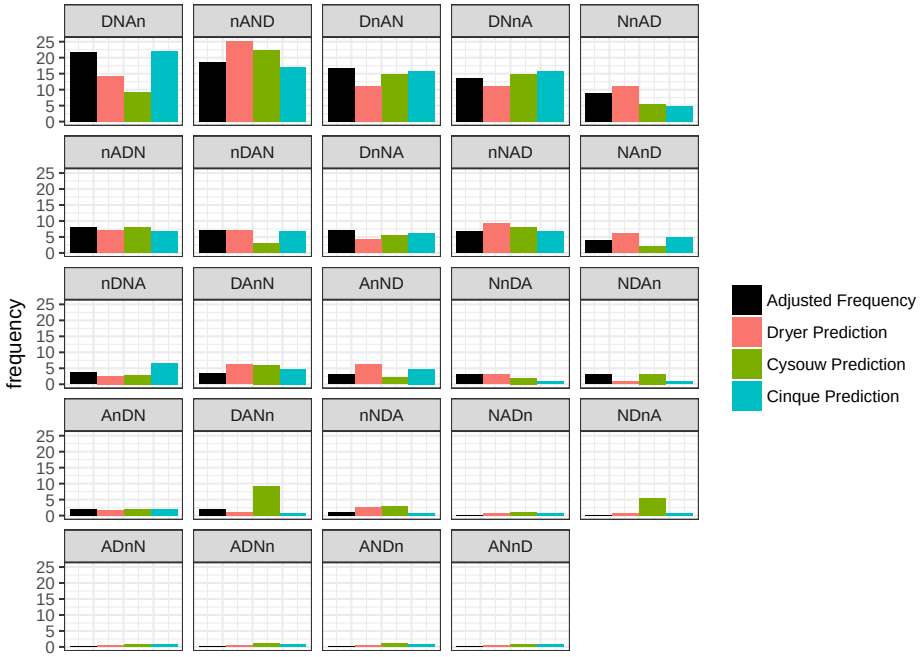

Figure 2: Adjusted frequency of word orders compared to model predictions. In this and all plots, the "Cinque" model refers to our featurization of Cinque's theory, including all features in the model without tying weights.

We wanted to know how much each order contributed to model fit, so in Figure 3 we show signed χ^2-discrepancies between model predictions and adjusted frequency. The χ^2 discrepancy measures how much the prediction error for each word order contributes to the overall discrepancy between data and model fit; signed χ^2 discrepancy presents this discrepancy in the direction of the discrepancy for each order (whether it under-predicts or over-predicts). If a model predicts a count of E_i for the ith word order and the observed count is O_i, then the signed χ^2 discrepancy is:

$$\frac{(O_i - E_i) \times |O_i - E_i|}{E_i}.$$

The magnitude of the discrepancy corresponds to how much a model is penalized for failing to predict a certain order.

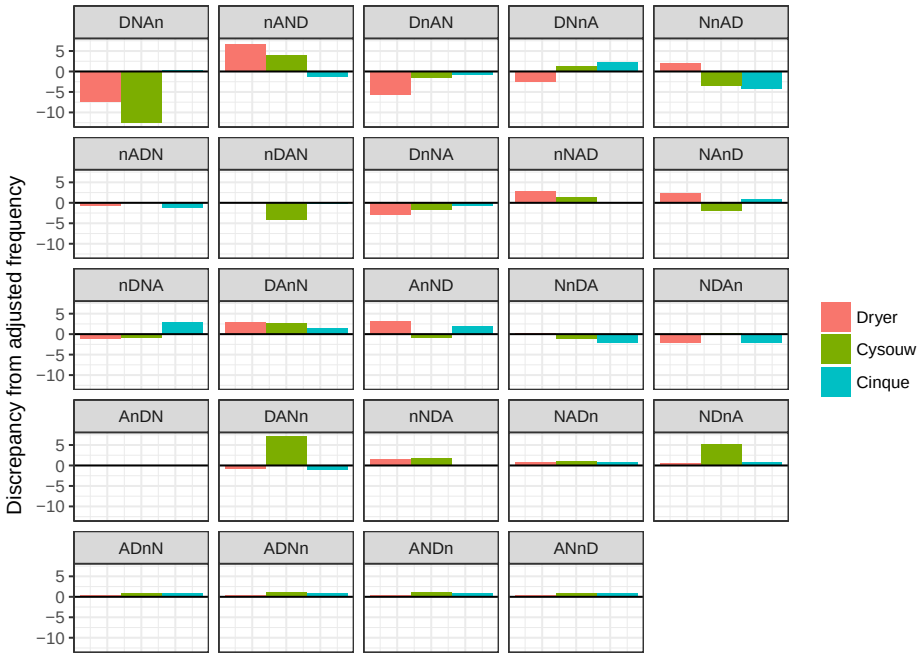

Figure 3: Signed chi-squared discrepancies between model predictions and adjusted frequency. Scores below zero mean that a model underpredicts frequency; scores above zero means that the model overpredicts frequency.

As another way to analyze the results, we show in Figures 4 and 5 the weights assigned to features for different word unders under the Dryer model and the Cinque model. Here we see that Cinque's model works by strongly

Model	Log likelihood	d.f.
Dryer (in prep)	-61.3	6
Cysouw (2010)	-106.0	5
Cinque (our features)	-67.7	7
Cinque (tied markedness)	-70.1	5
Cinque (Merlo's features)	-71.1	8

Table 4: Log likelihoods of *genera count* data under various models, and the degrees of freedom (d.f.) of those models.

penalizing low-frequency orders using the `AlternativeMergeOrder` feature, and then the differences among the remaining orders are handled by the rest of the features. The figure also highlights why the Dryer model underpredicts the frequency of $DNAn$ orders: these orders are positive for the `nadj` feature, which must have a negative weight in order to penalize various low-frequency orders.

One limitation of applying Poisson regression to the adjusted frequency data is that Dryer (in prep)'s method of computing adjusted frequencies in general compresses high frequency counts more than low frequency counts. This means that adjusted frequency may overly penalize models that perform best at predicting the counts of common orders. For this reason, it is also important to evaluate the performance of the models under consideration in predicting genera counts. We turn to this matter in the next section.

3.2 Predicting Genera Counts

Now we turn to models that were trained to predict the genera count data given in Dryer (in prep). When we use the various feature systems to predict genera counts, we get the following data log-likelihoods, shown in Table 4:

So when predicting genera, we get the best fit to the data using the set of features from Dryer (in prep), followed by Cinque's (2005) features, followed by Cysouw's (2010) features.

We think Cinque's model comes out worse when predicting genera primarily because it underpredicts $NnAD$ orders, whereas the Dryer model gets that order exactly correct. This can be seen in Figure 6, which shows model predictions, and Figure 7, which shows signed χ^2 discrepancies compared to genera counts. Figures 8 and 9 show the optimal feature weights for the Dryer and Cinque models, respectively, when predicting genera counts.

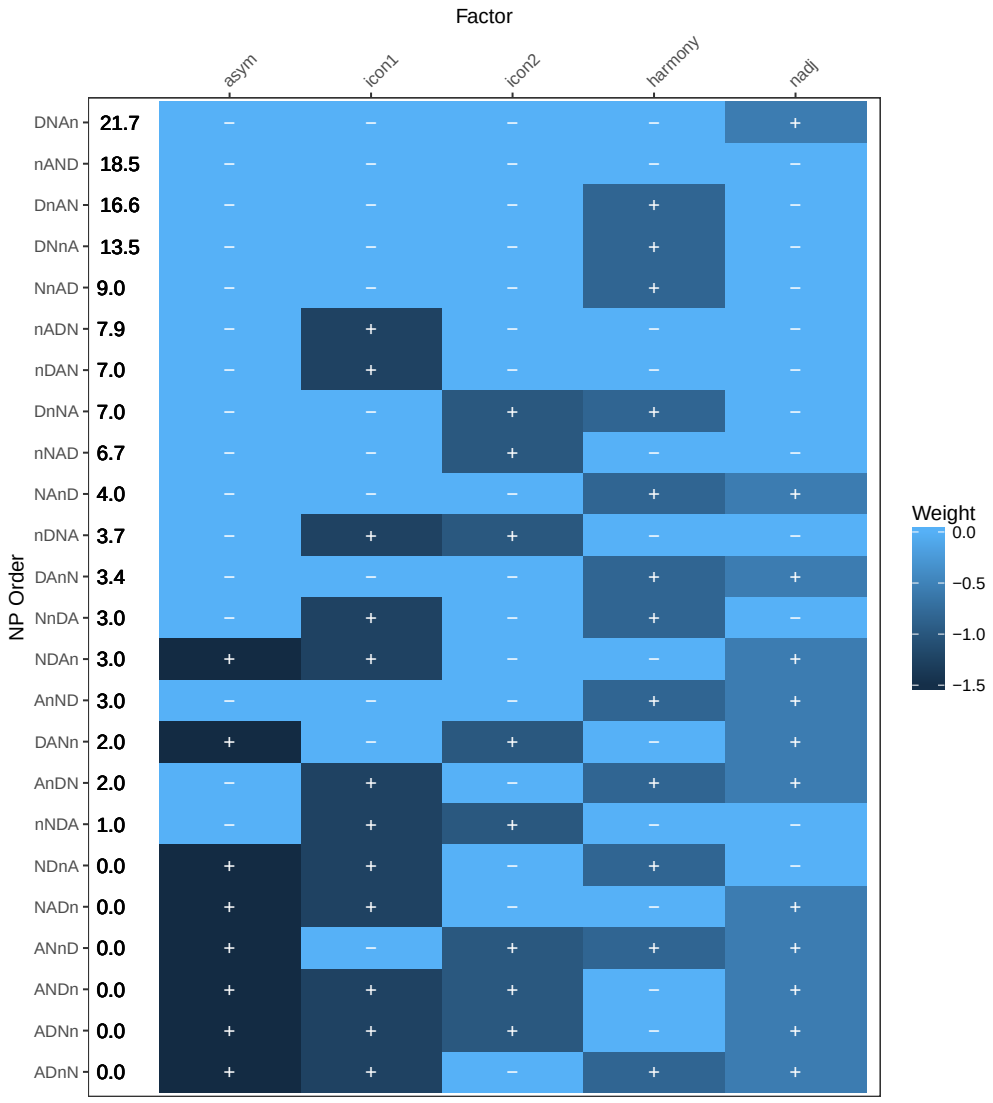

Figure 4: Feature weights from the Dryer (in prep) model when predicting adjusted frequency (first column).

4 Discussion

The results give clear evidence that the Dryer (in prep) and Cinque (2005) model provide feature systems that have better predictive power than the model of Cysouw (2010). But in our opinion they do not give strong reason

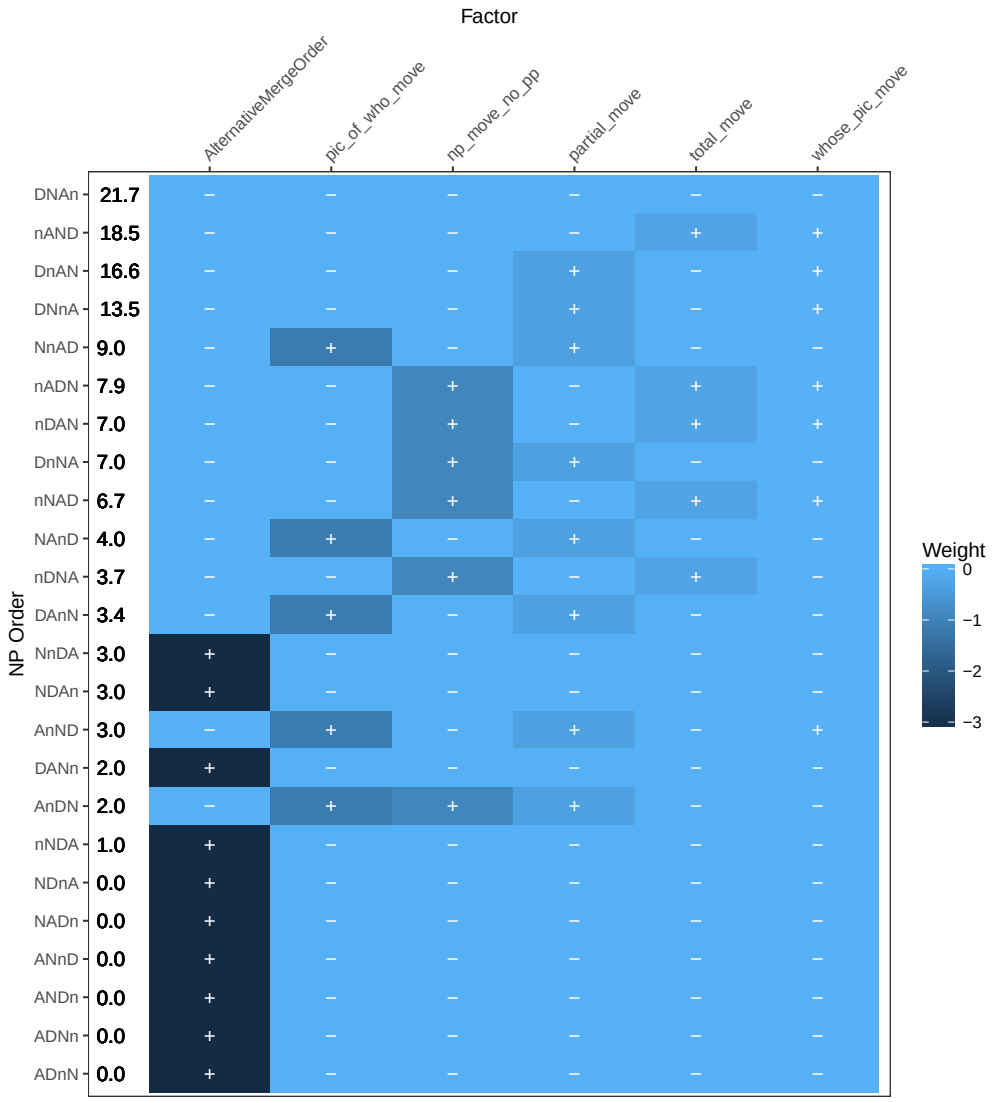

Figure 5: Feature weights from the Cinque (2005) model when predicting adjusted frequency (first column). Note that the feature `whose_pic_move` comes out to have a (non-significant) positive weight.

to favor Cinque's model over Dryer's model or vice versa. Although under a certain interpretation Cinque's model can a slightly higher fit to the data, this only holds under one featurization, and it does not hold when predicting

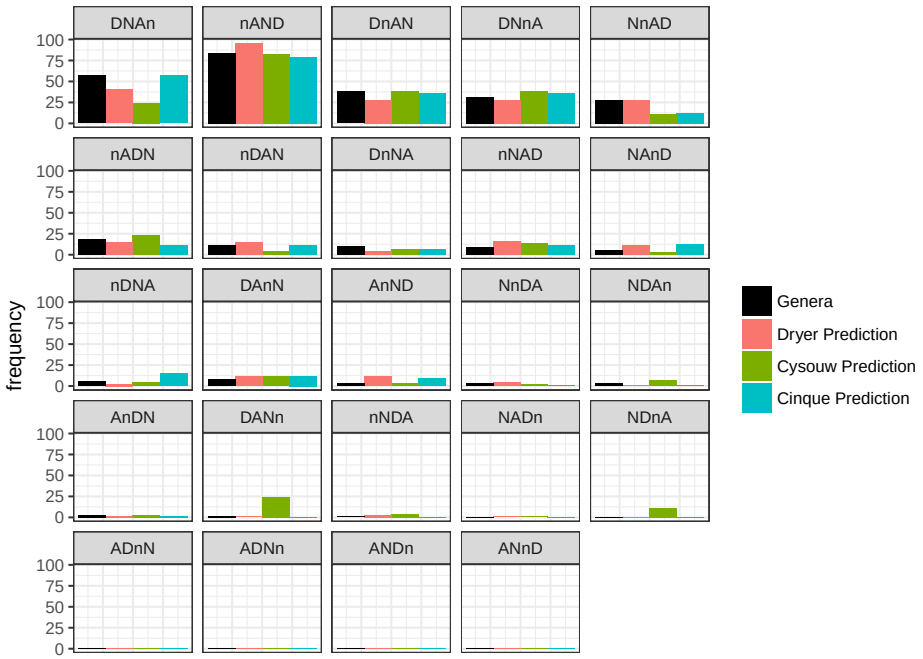

Figure 6: Genera counts of word orders compared to model predictions. In this and all plots, the "Cinque" model refers to our featurization of Cinque's theory, including all features in the model without tying weights.

genera counts. The discrepancy in results between adjusted frequency and genera counts may be due to the particular distributional characteristics of adjusted frequency as discussed above. The analysis suggests overall that the Dryer model and the Cinque model have roughly similar predictive power, and the current data do not discriminate between them.

Acknowledgments

This work was supported by NSF DDRI grant #1551543 to R.F.

References

Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic inquiry*, 36(3):315–332.

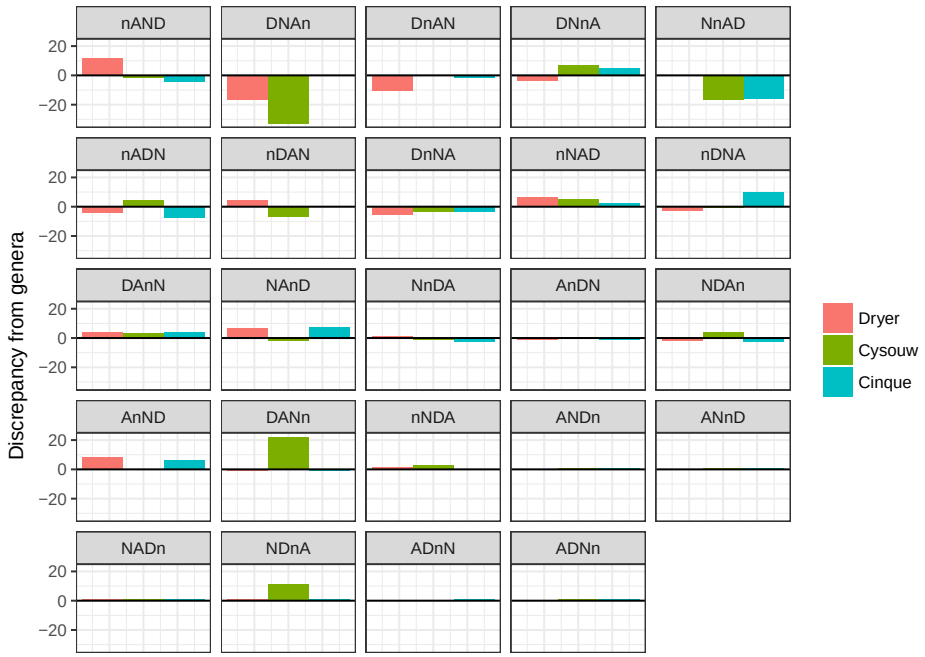

Figure 7: Signed chi-squared discrepancies between model predictions and genera counts. Scores below zero mean that a model underpredicts frequency; scores above zero means that the model overpredicts frequency.

Cysouw, M. (2010). Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology*, 14(2-3):253–286.

Dryer, M. S. (2006). On Cinque on Greenberg's Universal 20.

Dryer, M. S. (in prep). On the order of demonstrative, numeral, adjective and noun.

Kayne, R. S. (1994). *The Antisymmetry of Syntax*. MIT Press, Cambridge, MA.

Merlo, P. (2015). Predicting word order universals. *Journal of Language Modelling*, 3(2):317–344.

Smolensky, P. and Legendre, G. (2006). *The Harmonic Mind*. MIT Press, Cambridge, MA.

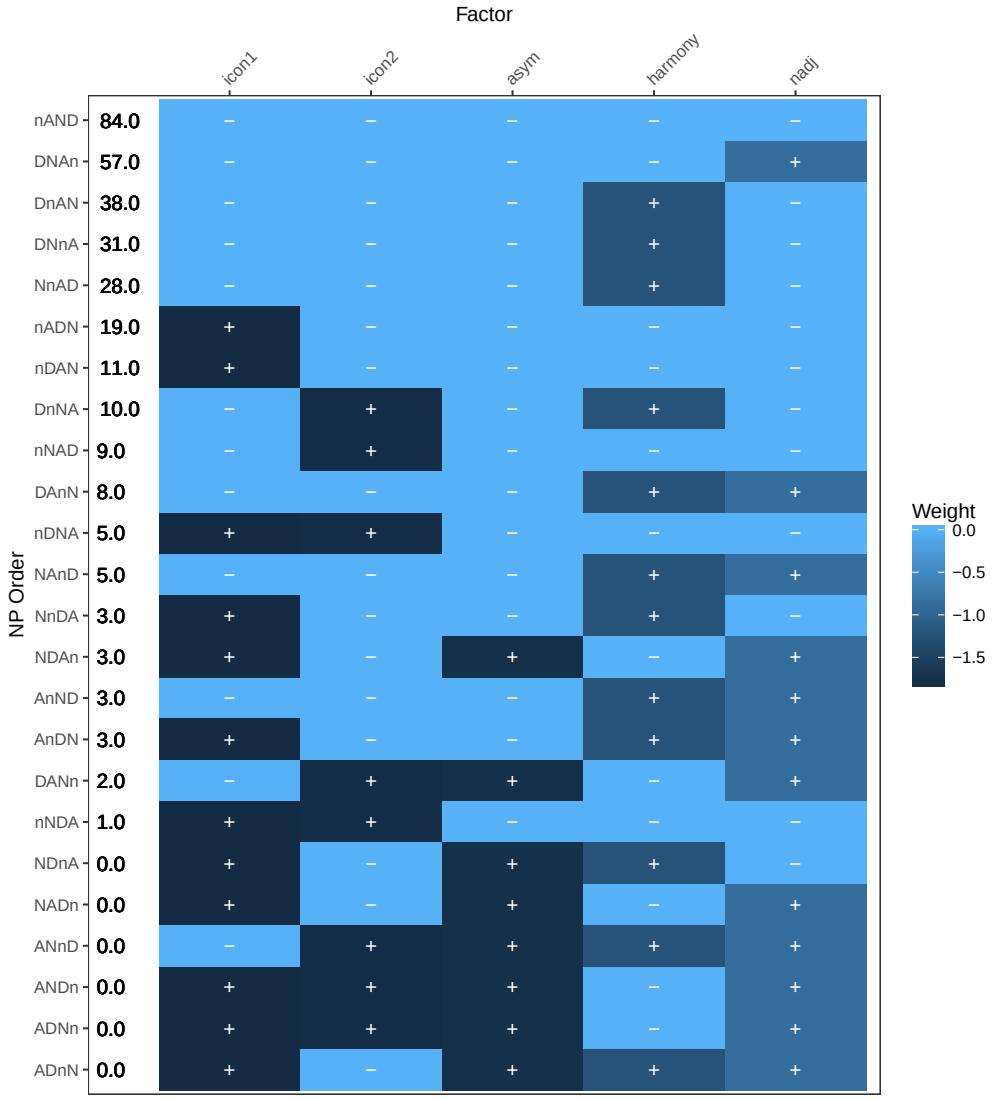

Figure 8: Feature weights from the Dryer (in prep) model when predicting genera counts (first column).

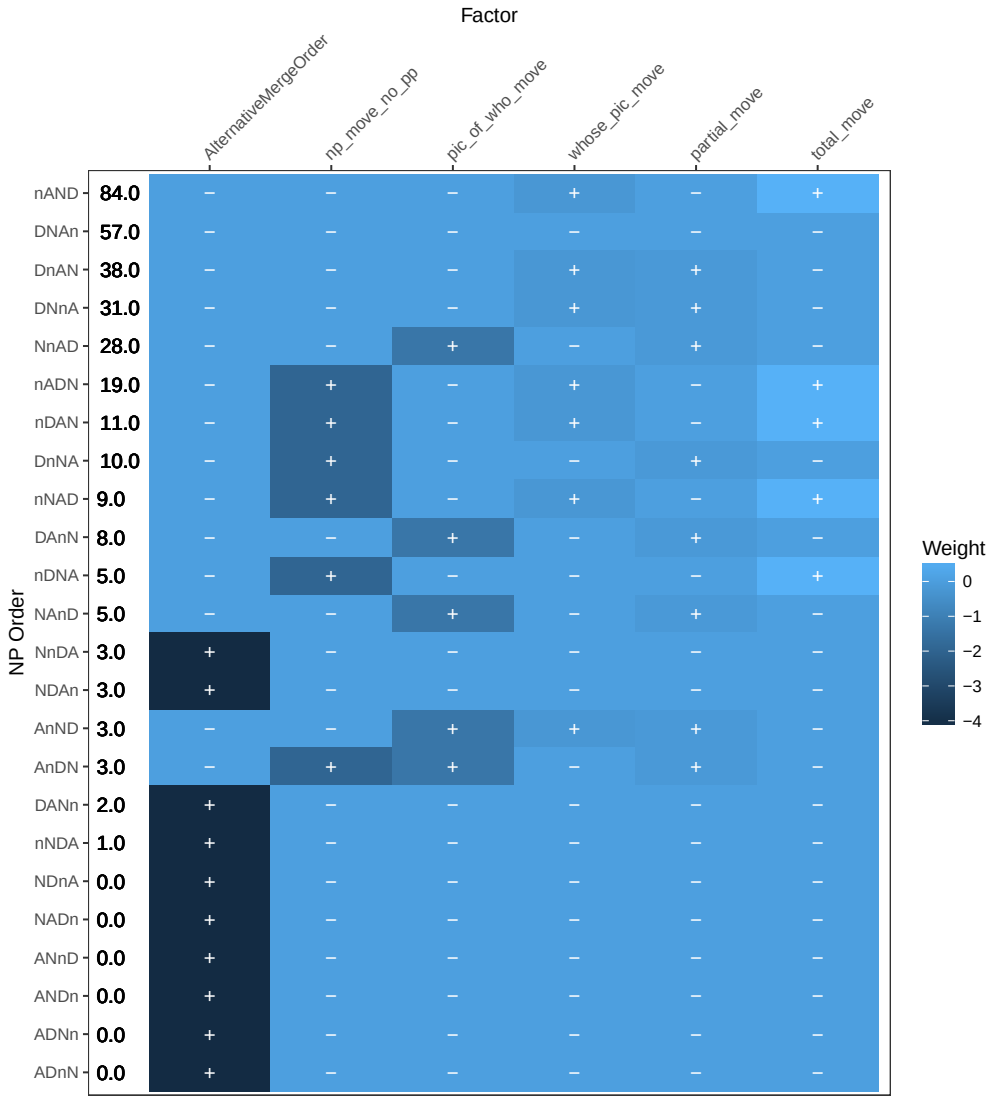

Figure 9: Feature weights from the Cinque (2005) model when predicting genera counts (first column). Note that the feature `total_move` comes out to have a (non-significant) positive weight.