The trace of categorical structure in gradient judgments

Introduction. In acceptability judgment experiments, the distribution of responses to a particular sentence type tends to lie along a continuum—e.g., different instances of the same sentence type generally receive slightly different average ratings. This gradience is unsurprising, since acceptability is likely sensitive to processes that may be fundamentally gradient, like processing and task effects (Sprouse 2007); but it remains an open question in many areas of syntax whether this gradience furthermore implies a gradient grammar (Keller 2000, Featherston 2005) or whether a traditional categorical account is tenable. This question has been particularly contentious in the domain of islands, where both categorical and gradient accounts exist.

Goals. We present computational modeling evidence that, once gradient processing and task effects are controlled for, there is evidence *against* a gradient grammar in the domain of islands (cf. Gorman 2013, 2014 on categorical phonotactic models). To do this, we construct four Bayesian hierarchical models that differ only in their representation of the grammar: (a) one with no grammatical representation, (b) one with a gradient grammar, (c) one with a categorical grammar, and (d) a hybrid categorical model incorporating a minimal amount of gradience. We fit these models to Sprouse's (2015) large-scale islands dataset and show that, using a standard measure of data fit and model complexity, (i) the models with a grammar outperform the one without and (ii) the purely categorical model outperforms both the gradient and hybrid models.

Background. For current purposes, accounts of islands (and other syntactic phenomena) fall into two high-level categories: gradient grammar accounts (Keller 2000, Featherston 2005) and categorical grammar accounts (Ross 1967, Chomsky 1973, 1986). The first step in comparing these accounts is to empirically isolate the size of island effects. Because island violations also entail several potential processing costs, such as the cost of completing a long distance dependency and the cost of building the complex syntactic structure that typifies islands, Sprouse et al. 2012 suggest using a 2 × 2 factorial design (ISLANDHOOD × DISTANCE) to control for the effects of long distance dependencies and island structures, as in (1).

(1) a. Who thinks that John bought a car? *non-island* | *short*
b. Who wonders whether John bought a car? *island* | *short*
c. What do you think that John bought _ ? *non-island* | *long*
d. What do you wonder whether John bought _? *island* | *long*

Under this design, if the only effects impacting acceptability were independent costs of long distance dependencies and island structures, the difference in acceptability between (1a) and (1b) and that between (1a) and (1c) should sum to the difference between (1a) and (1d). However, if there is an additional island effect present, there should be a superadditive interaction that in effect leads the difference between (1a) and (1d) to be even greater than that between (1a) and (1b). This superadditive interaction term provides an estimate of the size of the island effect itself. (Purely processing-based explanations of the interaction also exist. We assume a grammatical source here in order to probe the possibility of a gradient grammar.)

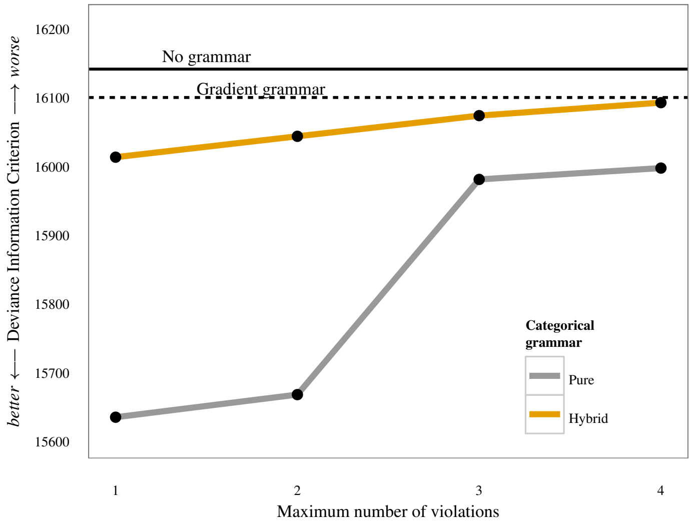
Once the island effects themselves are isolated, the next question is whether the grammatical violation that gives rise to the effect is gradient or categorical. There are two possibilities: the interaction might be due to a gradient violation with magnitude equal to the size of the interaction, or it might be indicative of some constant acceptability decrement found for categorical violations. One way to distinguish these two accounts is to construct multiple factorial experiments of the type described above with potentially different kinds of grammatical violations. In a gradient grammar approach, one expects to see (i) acceptability decrements of various magnitudes (possibly as many as there are different structural configurations and dependencies) and (ii) no constraints on the relationship among these magnitudes. On the categorical approach, (i) there should only be as many magnitudes as there are potential violations and (ii) if there is the possibility of multiple violations affecting acceptability (cf. Chomsky 1986 on barrier counting) those magnitudes should come in equal "hops"—i.e., if one violation causes an acceptability decrement d (above and beyond the processing and task effects), two violations would cause a decrement $2d$. Sprouse (2015) deploys an experiment series of this type, though he does not formally compare the categorical and gradient models, which we do here.

Data. Sprouse tested four distinct ISLAND types—*whether* islands, complex NP islands, subject islands, and adjunct islands—across three potentially distinct DEPENDENCY types—bare WH-word movement, Dlinked WH-phrase movement, and relative clause displacement. Because relative clause dependencies are necessarily embedded, Sprouse also tested the two WH dependencies in both matrix and embedded structures. The full cross of these factors results in 20 factor pairs, which were fully crossed with DISTANCE and ISLANDHOOD factors—see (1) for example bare WH + *whether* island items—to create 80 conditions.

Model. We construct four Bayesian hierarchical models, which model processing and task effects identically and differ in only their grammatical representation. Each of these models is an ordinal mixed model with participant and item random effects and at least the following fixed effects structure: DEPENDENCY*ISLAND*DISTANCE + DEPENDENCY*ISLAND*ISLANDHOOD. This fixed effects structure implements controls on processing and task effects in the same way suggested in Sprouse et al. 2012. (The DEPENDENCY × ISLAND interactions reflect the fact that Sprouse 2015 had 20 ISLAND-DEPENDENCY pairs instead of one.) And since it does not admit DISTANCE × ISLANDHOOD interactions, it has no way of representing island violations (see Background). Thus, we call this baseline model our *no grammar* model.

We implement our *gradient grammar* model by adding to the *no grammar* model fixed effects for DISTANCE × ISLANDHOOD interactions and all possible interactions with ISLAND and DEPENDENCY: DEPENDENCY*ISLAND*DISTANCE*ISLANDHOOD. This addition adds a real-valued term for each ISLAND-DEPENDENCY pair representing a gradient violation penalty associated with that dependency in that structure. These penalties are unconstrained in their size and their relationship to each other and thus fulfill our two criteria for a *gradient grammar* model discussed above. To implement our *pure categorical grammar* model, we take our *gradient model* and force the interactions (i) to be integer-valued and (ii) to not exceed a specific maximum. The first property satisfies the criterion that violations produce the same decrement in acceptability—unlike the gradient model, where decrements can be of any size. (The maximum is necessary, since a categorical model would approximate a gradient model using many small hops were it not present.) We implement our *hybrid categorical grammar* model by adding to the *pure* model the minimal amount of gradience possible: a single real value added to the integer-valued representation of the island violation.

Results. We fit all four models to Sprouse's data using Markov Chain Monte Carlo (MCMC). (The two categorical models were fit with 4 different violations maxima.) MCMC is useful since it allows us to compute the *Deviance Information Criterion* (DIC), which is a measure of model optimality that trades off model fit with model complexity in a statistically principled way.



More complex models—e.g., the *gradient grammar*—always fit data better, since they are strictly more expressive than less complex models—e.g., the *categorical grammar*. Model complexity as quantified by DIC can mean having more parameters (the *no grammar* v. *categorical* and *gradient grammars*) or allowing more possible values for the same number of parameters (*categorical* v. *gradient grammars*). The graph plots the DIC for each model. We see that the grammar models beat the *no grammar* model and that the pure categorical model with only 1 or 2 violations handily beats all others. This suggests that the best model of the data is a very constrained pure categorical model. This suggestion is further corroborated by the fact that the *hybrid categorical* model is not much better than the gradient model even under heavy constraint on the maximum; and it converges very quickly in DIC to the *gradient grammar*.

Discussion. We have shown evidence against a gradient grammar account of islands. In future work, we will show how our computational models can be deployed to further investigate violation stacking phenomena.

Keywords: islands, gradience, computational models