Modeling Ungrammaticality: A Self-Organizing Model of Islands

Sandra Villata (sandra.villata@uconn.edu)
Department of Linguistics, 365 Fairfield Way
Storrs, CT 06269-1145 USA

Jon Sprouse (jon.sprouse@uconn.edu)
Department of Linguistics, 365 Fairfield Way
Storrs, CT 06269-1145 USA

Whitney Tabor (whitney.tabor@uconn.edu)
Department of Psychology and Haskins Laboratories, 406 Babbidge Road
Storrs, CT 06269-1020 USA

Abstract

Formal theories of grammar and traditional parsing models, insofar as they presuppose a categorical notion of grammar, face the challenge of accounting for gradient judgments of acceptability. This challenge is traditionally met by explaining gradient effects in terms of extra-grammatical factors, positing a purely categorical core for the language system. We present a new way of accounting for gradience in a self-organized sentence processing (SOSP) model, which generates structures with a continuous range of grammaticality values. We focus on islands, a family of syntactic domains out of which movement is generally prohibited. Islands are interesting because, although most linguistic theories treat them as fully ungrammatical and uninterpretable, experimental studies have revealed gradient patterns of acceptability and evidence for their interpretability. We report simulations in which SOSP largely respects island constraints, but in certain cases, consistent with empirical data, coerces elements that block dependencies into elements that allow them.

Keywords: whether islands; subject islands; D-linking; acceptability; ungrammaticality; gradient effects; self-organized sentence processing model; SOSP

Introduction[1]

Acceptability judgments are gradient: sentences' acceptability spans from full acceptability to full unacceptability passing through a range of intermediate values which can be statistically distinguished. Grammaticality, on the contrary, is traditionally conceived of as categorical: sentences are either grammatical or ungrammatical but cannot be "partially" (un)grammatical. Degrees of acceptability have been attributed to extra-grammatical factors, such as memory limitations, plausibility etc. It is commonly assumed that this view comes with the advantage of simplicity: a grammar admitting only two states is claimed to be simpler than a grammar involving a continuous, infinite, number of states. We argue that, despite its apparent simplicity, this position is actually less parsimonious than one that accounts for graded acceptability judgments as deriving from the grammar itself. We present a self-organized sentence processing (SOSP) framework, which accounts for gradient effects through a single mechanism of structure building (e.g. Tabor & Hutchins, 2004; Smith & Tabor, 2018; Villata, Tabor, & Franck, 2018). Unlike most classical parsing and grammatical models, SOSP conceives of grammar as residing in a continuous space where fully grammaticality and fully ungrammaticality are two endpoints of a continuum (e.g. Kempen & Vosse, 1989; Cho, Goldrick, & Smolensky, 2017). As a result, gradient effects are understood as generated by the grammar itself, rather than deriving from extra-grammatical factors. To test this theory, we focus on what is arguably one of the most prototypical, and yet also most theoretically challenging syntactic phenomena: islands. Islands are encapsulated syntactic environments out of which almost nothing can be extracted (Ross, 1967). Islands come in two flavors: strong and weak. Strong islands are claimed to block all kinds of extraction. In particular, non D(iscourse)-linked (e.g. *what*, *who*) and D-linked elements (e.g. *which NP*) are equally unextractable from strong islands. This is illustrated in (1) and (2) for subject islands, where the NP (*what* or *which dissertation*) is extracted from a NP subject (*the first chapter of*)[2]:

(1) *__What__ do you think [the first chapter of _] is full of errors?
(2) *__Which dissertation__ do you think [the first chapter of_] is full of errors?

In contrast, weak islands are traditionally claimed to be selective: they prohibit the extraction of non D-linked wh-elements, but allow the extraction of D-linked wh-elements (e.g. Cinque, 1990; Rizzi, 1990). This is illustrated in (3) and (4) for whether islands, where the extraction of the NP is from a whether-clause:

(3) *__What__ do you wonder [whether the student read _]?
(4) __Which book__ do you wonder [whether the student read _]?

The sharp distinction between the examples in (1), (2) and (3) on the one hand, which are standardly deemed ungrammatical, and (4) on the other, which is typically considered grammatical, is very much in line with the traditional, categorical view of grammar, which only admits binary outcomes. However, with the development

[1]This work was supported, in part, by a grant from the Marica de Vincenzi Foundation.

[2]The island domain is in brackets, and the asterisk indicates ungrammaticality.

of finer-grained techniques for gathering acceptability judgments, experimental studies have revealed gradient patterns of acceptability for island effects. Here we focus on three empirical facts indicating gradient island effects.

First, acceptability judgment studies have revealed that weak island acceptability is gradient (e.g. Sprouse, Wagers, & Phillips, 2012; Sprouse & Messick, 2015). In particular, D-linked whether islands (4) are more acceptable than non D-linked ones (3), and yet still ungrammatical, contra the traditional wisdom that conceives of D-linked whether islands as grammatical (see Villata, Rizzi, & Franck 2016 for similar evidence for wh-islands). These studies used a 2x2 factorial design that isolates the island effect from two processing factors that are known to interact with the effect: (i) STRUCTURE TYPE (island vs. non-island),[3] and (ii) DEPENDENCY LENGTH (long vs. short) (5). The contrast between (5a) and (5c) isolates the cost of structure, while the contrast between (5a) and (5b) isolates the dependency length effect. We define the island effect as a statistical interaction between the two factors: it is what remains after the linear sum of the two processing factors is taken into account.

Sprouse & Messick (2015) found a significant interaction for both non D-linked and D-linked whether islands, indicating an island effect in both cases. However, the island effect was stronger in the non D-linked condition as compared to the D-linked condition, providing evidence that D-linking reduces the island effect in weak islands (see Figure 5; empirical data are in black).

(5) Factorial design measuring the whether island effect

a. NON-ISLAND, SHORT
Who/Which woman _ thinks that John bought a car?
b. NON-ISLAND, LONG
What/Which car do you think that John bought _?
c. ISLAND, SHORT
Who/Which woman _ wonders whether John bought a car?
d. ISLAND, LONG
What/Which car do you wonder whether John bought _?

The second empirical fact is that D-linking interacts with island types: while D-linking ameliorates the acceptability of weak islands, it does not help strong islands — e.g., subject islands. Example (6) shows a corresponding factorial design for subject islands, and Figure 6 (black lines) shows the empirical data from Sprouse & Messick (2015).

(6) Factorial design for measuring the subject island effect

a. NON-ISLAND, SHORT
Who/Which leader _ thinks the speech interrupted the TV show?
b. NON-ISLAND, LONG
What/Which speech does the leader think _ interrupted the TV show?
c. ISLAND, SHORT
Who/Which leader _ thinks the speech by the president interrupted the TV show?
d. ISLAND, LONG
Who/Which politician does the leader think the speech by _ interrupted the TV show?

Third, D-linked whether islands with an intransitive embedded verb (e.g., *Which joke does the comedian wonder whether the audience* ***laughed****?*) are less acceptable than those with a transitive embedded verb (e.g.,*Which necklace does the detective wonder whether the thief* ***stole****?*), an effect that was significant for both D-linked and non D-linked whether islands, although it was greater for the former (Villata, Sprouse, & Tabor, 2018) (Figure 1). We take this result as evidence that whether islands, though ungrammatical, are interpreted. This suggests that the dependency between the extracted wh-phrase and the gap inside the island can, at least sometimes, be established.

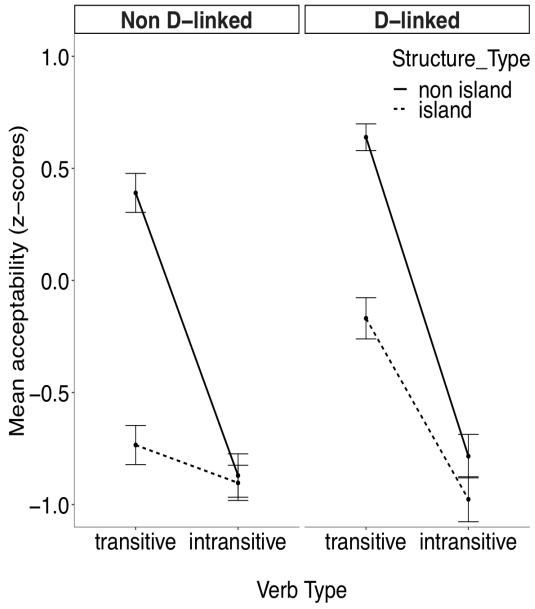

Figure 1: Acceptability proportions for weak islands (data from Villata, Sprouse, & Tabor (2018)).

Summarizing: First, weak islands are ungrammatical. Though D-linking improves their acceptability, it does not cancel the island effect. Hence, their acceptability is gradient. Second, D-linking does not improve the acceptability of strong islands. Hence, gradience is not evident in all cases. Third, the evidence suggests that weak islands are interpreted. Hence, the dependency between an extracted wh-phrase and a gap inside the island can sometimes be established. The last fact seems to point to an account of islands not couched in terms of perfectly impenetrable syntactic domains.

In the next section we introduce the SOSP model. In the section *Model Implementation* we describe SOSP's implementation and, in the section *Simulations*, we describe

[3] "Island" here does not refer to an island-violating structure, but to the mere presence of a structural domain that does not tolerate extractions, like a whether embedded clause or a complex subject.

how the model accounts for the data at hand.

The SOSP Model

In SOSP, structures are formed through continuous dynamical interaction among their constituent elements. Building on several linguistic foundations (Fillmore et al., 1988; Fodor, 1998; Gazdar, 1981) and following the psycholinguistic formulation of Kempen & Vosse (1989), we take the constituent elements to be treelets. Treelets are subtrees formed by a mother node and a finite number of daughter nodes that become active when a word is encountered. Each treelet is associated with a vector of syntactic and semantic features that specifies the properties of the word and its expected dependents. Treelets interact in all possible ways to form structure, creating competition for attachment. Attachments between treelets with a good feature match generally outcompete attachments with a poor feature match, which leads the system to stabilize, most of the time, on a grammatical structure. Structures in which all attachments perfectly satisfy the requirements of the feature vectors of the treelets receive the maximum *harmony value* of 1. Harmony is a formal measure of the degree of coherence in a set of interacting treelets — details below (e.g. Smolensky, 1986).

Importantly, SOSP also allows the generation of intermediate structures, i.e. structures with a harmony value strictly between 0 and 1 (0 harmony = no structure). This happens when an attachment is made between treelets whose features only partially match. This can happen in two ways. First, due to noise in the system, attachments between treelets with a poor feature match can sometimes outcompete attachments between treelets with a good feature match. However, this will happen in a small proportion of the cases, for attachments with a good feature match tend to win competitions. Second, when no optimal bond is available, as in ungrammatical sentences and difficult garden paths, the system forces the attachments to form anyway, generating (sub-optimal) structures. This leads to a variety of differently-valued outcomes which are *internally* generated (i.e. generated by the functioning of the system itself), rather being the result of factors that are external to the system.

Model Implementation

The implemented model (Smith & Tabor, 2018) consists of sets of differential equations that converge on fixed points corresponding to locally optimal structures. Treelets are encoded as banks of feature vectors (all of the same dimensionality, $nfeat$) with one bank for each attachment site (mother/daughters). The general implementation is achieved by first determining all the possible structures (both fully and partially grammatical) that can be formed from the vocabulary of the language, treating the concatenated banks of features and link values as forming a single vector space, and identifying the location of each locally optimal structure in this space. The local harmony, h_i, associated with such a point in the feature space is given by (1):

$$h_i = \prod_{l \in links} \left(1 - \frac{dist(\mathbf{f}_{l,daughter}, \mathbf{f}_{l,mother})}{nfeat}\right) \quad (1)$$

where $dist(\vec{x} \cdot \vec{y})$ is Hamming distance between $\vec{x}$ and $\vec{y}$. In other words, the local harmony is a product, across links, of a measure of similarity between the daughter feature vector $\mathbf{f}_{l,daughter}$ and the mother feature vector $\mathbf{f}_{l,mother}$ on the end of each link. Thus, if every link has a perfect match, then h_i is maximal and equals 1. The minimum possible value is 0, and various degrees of mismatch give intermediate values.

For each such structural locus, we specify a radial basis function (RBF), ϕ_i (Muezzinoglu & Zurada, 2006):

$$\phi_i(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{c}_i)^\intercal(\mathbf{x}-\mathbf{c}_i)}{\gamma}\right)$$

Here, $\mathbf{x}$ is the state of the system encoding the values of all features on all activated treelets and all possible links between them, $\mathbf{c}_i$ is the location of the ith (partial) parse, $^\intercal$ denotes the vector transpose, and γ (a free parameter) specifies the width of the RBFs. We define the harmony function $H(\mathbf{x})$ as the height of that RBF among n RBFs that is maximal at x, where n is the number of optimal and partially-optimal structures (harmony peaks) that can be formed with the currently activated elements, and h_i is the height of the i'th mode:

$$H(\mathbf{x}) = \max_{i \in 1...n} h_i \phi_i(\mathbf{x}) \quad (2)$$

This equation interpolates a harmony landscape between the structural loci, c_i, associated with the local harmony peaks.[4]

Parsing starts with all features equal to 0. The perception of the first word of a sentence causes features of a lexical treelet associated with that word to be turned on. This, in turn, causes links and additional treelet feature banks corresponding to the most viable parse of just that word to be turned on. In SOSP, treelets are interacting subsystems that attempt to assemble themselves through local interactions that locally maximize harmony. This is implemented as noisy gradient ascent on the harmony surface, $H(\mathbf{x})$:

$$\frac{d\mathbf{x}}{dt} = \nabla_{\mathbf{x}} H(\mathbf{x}) = -\tfrac{2}{\gamma} h_{i_{max}}(\mathbf{x} - \mathbf{c}_{i_{max}})\phi_{i_{max}}(\mathbf{x}) + \sqrt{2D}\, dW \quad (3)$$

where $D > 0$ scales the magnitude of the Gaussian noise process dW. In other words, the system moves approximately uphill on the harmony landscape as it processes each word. Moving uphill is equivalent to growing the link structures and adjusting values of unspecified or conflicting features to

[4]This definition differs from the form specified in Smith & Tabor (2018) who summed the RBFs to form H. We have found that, in systems with many harmony peaks, if a summation is used, there are often ganging effects that influence the structure of the gradient and flummox effective parsing: many proximal ungrammatical structures gang together to pull the state toward their mean and away from a lone worthy grammatical candidate. Humans seem to be strongly influenced by the presence of a good candidate even if there are also many bad ones around, so the max method yields more plausible parsing than does the summing method when the language model is realistically rich.

reach a locally optimal parse state. After a local optimum is reached, new features specified by the next word are turned on (moving the system off its current hilltop and into a nearby valley). The gradient ascent process then begins anew and a new harmony maximum is reached, corresponding to the next step of the parse. Across multiple trials, the noise produces a distribution over the harmony maxima, generally favoring those that correspond to plausible parses of the input seen up to the present moment. At the end of parsing a sentence, the system will be at a particular harmony peak that has a value between 0 and 1. We take this harmony value to correspond to the model's assessment of the acceptability of the sentence.

Simulations

In the terms of the model, based on the empirical results reviewed above, we identify the following desiderata: (i) non D-linked whether islands should receive a low harmony value, and D-linked whether islands should receive a higher, but not maximal, harmony value; (ii) the high-but-not-maximal harmony for D-linked whether islands should be generated by linking the gap to the filler inside the island with some strain, in line with experimental findings suggesting that these structures are interpreted (Villata, Sprouse, & Tabor, 2018); (iii) subject islands should receive a low harmony value irrespective of the presence of D-linking (comparable to non D-linked whether islands) .

Figure 2 portrays the model's processing of a non-D-linked whether island. The model considers, in parallel, all conceivable parses of the input string. However, since many of these parses have extremely low harmony and do not have much influence on the processing, the figure only shows those that play a significant role in the parsing dynamics. One reads the figure from left to right and bottom-up.[5] Typically, when a word is perceived, bonds between treelets form. For example, when "you" is perceived, a bond between "NP_{you}", the mother of "you", and "NP_{you}", the daughter of "S/NP_{what}", typically forms. Bonds between treelets that are formed by the system are illustrated with dashed lines, while straight lines indicate the treelet's structure as it is defined in the lexicon based on phrase structure rules (e.g. S $\rightarrow$ NP VP). Crucially, the treelet feature vectors are mutable within a range of values corresponding to the syntactic/semantic range that the treelet affords. For example, in the present case, the mother of the "S $\rightarrow$ NP VP" treelet has mutated to acquire a slash buffer that specifies the syntax and semantics of the fronted element "what". Due to this mutation, links can often achieve a perfect feature match, causing the relevant term in Equation (1) to take on the value, 1.

The crucial developments in the case of the non-D-linked whether sentences occur when the words "wonder" and "whether" are perceived. As shown in Figure 2, the system is deciding between two possible, not-fully-grammatical structures at "wonder". The first, shown by the left

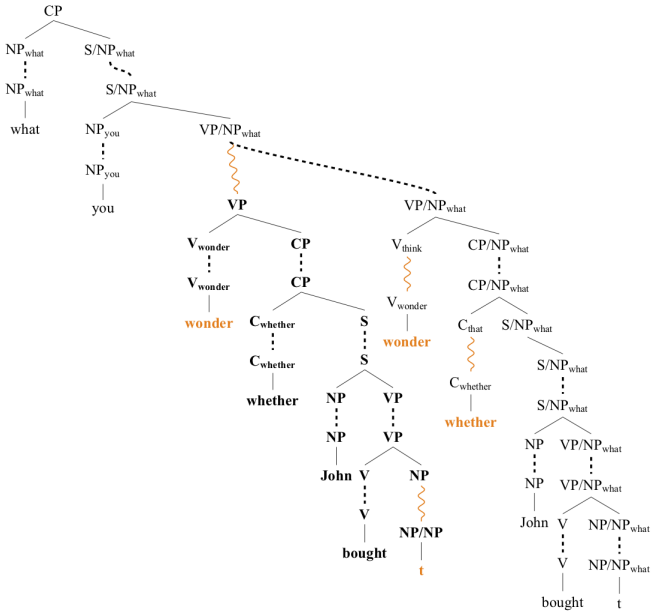

Figure 2: Simplified tree for non D-linked whether island. Subscripts indicate which feature has been transmitted to the node (e.g. S/NP_{what} means that *what* has been propagated to the S node). Words in orange are those that trigger coercion. Wavy orange lines illustrate the dynamic of the coercion. Here the parsing that ultimately wins out in most trials is the one on the left branch (bold font).

VP-branch, respects the constraint imposed by the verb "wonder", which cannot take as a complement an element with a slash feature. This implements the "islandhood" of the CP-complement of "wonder". This parse makes it possible for "$V_{wonder} \rightarrow$ wonder" to attach with perfect harmony to its CP-complement, but at the cost of failing to propagate the slash buffer ("$/NP_{what}$") onto the VP node below. We assume this failure has a cost, but not a severe cost because "what" is a very abstract element, so its encoding plausibly contains only a few features — that is to say, the difference between the "NP_{what}" slash buffer and an empty slash buffer is a small difference. This mild penalty is indicated by the orange color of the link between "VP/NP_{what}" and "VP". The second parse in Figure 2, shown by the right-branch, takes an opposite approach: it propagates the slash buffer, "$/NP_{what}$", onto the VP node below, but it can only do this by coercing "wonder" into a verb that licenses slash propagation. We have taken the verb "think" as a canonical example of such a verb. The orange squiggly line from "V_{think}" to "V_{wonder}" indicates this penalty. Again, this penalty is not extreme because the semantics and syntax of "wonder" and "think" are fairly similar. At the next word, "whether", the system undergoes a second coercion, of "whether" into "that". This coercion, which is triggered by the requirements of the verb "think", allows the slash buffer to propagate down the tree, for "that", unlike "whether", does not act as a slash propagation blocker. As before, although this coercion comes at a cost, the cost is mild, because of the similarity of the two complementizers.

[5]The model employs slash-propagation (Gazdar, 1981) to implement long-distance dependencies.

We now consider the case of the D-linked whether-island, illustrated in Figure 3. In this case, not propagating the slash feature onto the VP node (the parse on the left-branch) comes with a strong penalty (illustrated by the red squiggly line in the figure). This is because, D-linked words, unlike non D-linked ones, are associated with a rich bundle of semantic and syntactic features. As a result, failing to propagate the features associated with D-linked NPs incurs a strong penalty. The system therefore tends to prefer the second parse (right-branch): the close analogy between "think" and "wonder", and "that" and "whether" makes the mild coercions option better than any other parsing option, leading the parser to stabilize on the option that propagates the slash buffer inside the whether island.

In subject islands, on the contrary, there is no such close analogy between the words in the sentences and alternatives in the lexicon. As a result, the parser systematically fails to consider the possibility of propagating the slash feature down the subject branch and then is caught up short when a gap appears in the subject, and no gap appears in the main verb phrase (see Figure 4).

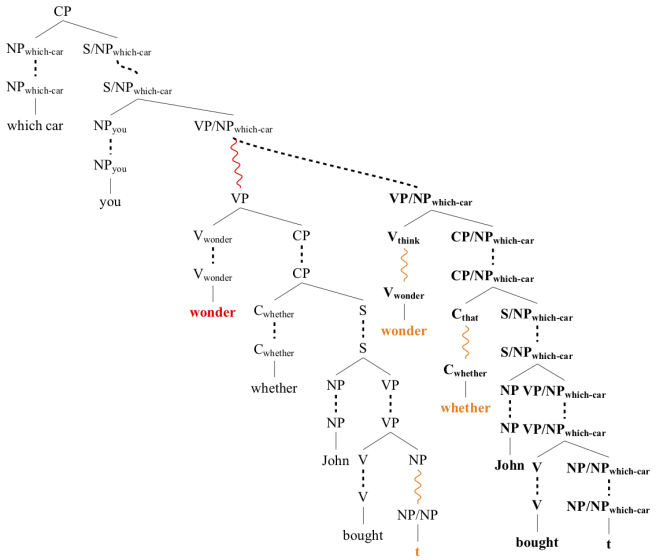

Figure 3: Simplified tree for D-linked whether island. Words in orange trigger a mild coercion, while words in red trigger a severe coercion. Here the parsing that ultimately wins out in most trials is the right branch (bold font).

We ran 20 runs of the model on simplified versions of each sentence in examples 5 and 6 (no determiners, ignoring English do-support). The model, somewhat revised from the one described in the first, reviewed version of this paper, is both an elaboration and a simplification of the model described in Smith & Tabor (2018).[6] It used 45

[6]We describe the revised model here rather than the original one because its assumptions are more plausible and easier to describe, as requested by several anonymous reviewers, and the causal dynamics by which it produces the data points reported here—specified in the analyses above—are the same as those previously described.

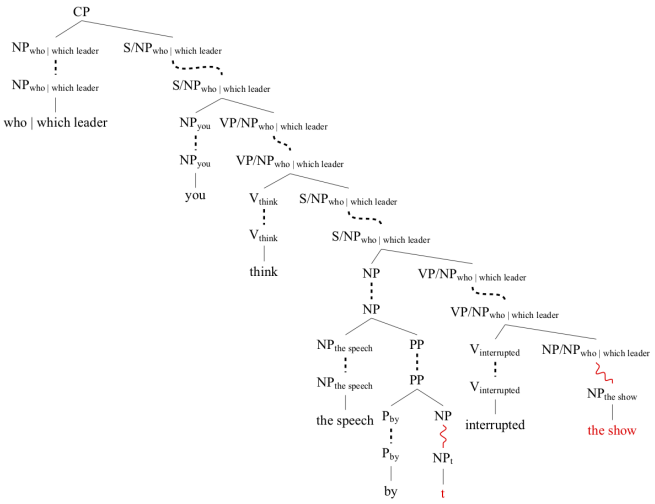

Figure 4: The low-harmony structure that stabilizes when the model is presented with a subject island.

distinct feature vectors for coding the lexical and syntactic nodes needed for the tree configurations described in the analyses above (as well as variants needed for all the stimulus sentences listed in (5) and (6)). Whereas previous versions of the model were hand-coded with roughly plausible linguistic features, the current version started by generating random bit vectors in 20 space for each feature vector (which was either a mother or a daughter of a treelet). This made all the feature vectors relatively distant from one another. Then, in keeping with the hypotheses described above, the vector for "wonder" was made to be equal to the vector for "think" except in two dimensions where it had contrasting bits; the vector for "whether" was analogously made similar to the vector for "that"; and the vector for "CP/What" was analogously made similar to the vector for "CP" (i.e., to CP with an empty slash buffer). SOSP entertains a plethora of possible ways of combing the treelets, most of which give rise to very low harmony structures. In the simulations reported here, motivated by the assumption that many of the low-harmony variants have little effect on the parse trajectory and to simplify implementation, we only considered the variants that we have mentioned as alternatives in the analyses above. An earlier version of the model had trouble telling sentences apart if many were included in the stimulus set. Here, we introduced a two-fold magnification of the dimension coding the features for the lexical elements. This effectively moved the harmony peaks for sentences with different word forms farther apart from one another, causing the system to prefer parses that are faithful to the input, though not rigidly—see Levy (2008). To allow the model to detect harmony maximization upon processing of each word, we allowed the dynamics to settle through a quadratic velocity profile: the model had to speed up (associated with reaching the steep section of one of the RBF humps) and then slow down (indicating that it was topping out on a harmony maximum) before moving on to the next word.

In addition to the number of feature dimensions (20) and the degree of lexical isolation (2 x) mentioned above, important free parameters are γ, specifying the width of the RBFs, D, specifying the magnitude of the noise, and ρ which takes its values in [0, 1] and determines how far over the velocity "hump" the model must travel before moving to the next word ($\rho = 1$ implies immediate transition, $\rho = 0$ implies infinite processing time per word), and Δt which specifies the step size in the Euler Integration that we used to approximate the dynamics. We explored these parameters by hand finding a way to roughly optimize behavior in a test grammatical sentence and the D-linked whether island extractions (D-linked, whether, island, long) to establish the settings $\gamma = 4$, $D = 7 \times 10^{-1}$, $\rho = 0.4$, $\Delta t = 0.5$ and then examined the results in the other fourteen conditions.

One other point about the implementation is particularly important. The current versions of SOSP add dimensions to the state space with every new word (these dimensions correspond to the feature banks in treelets that the word introduces, and to the links this treelet can potentially form with other activated treelets). The behavior of the dynamical equations is sensitive to the dimensionality, so to achieve reasonable parsing, such an implementation needs to change the dynamical parameters (γ, D, ρ, Δt) as the sentence grows. We do not think this is very plausible. Instead, we think the dimensionality of human processing is kept roughly constant via a focusing mechanism (possibly related to what is called "Working Memory" in other work). We suspect that the form of this focusing involves fractal scaling as has been proposed in work on neural encoding of arbitrary dependency languages (Plate, 2003; Tabor, 2000). However, we do not know how to apply such scaling techniques to the SOSP encodings, so we have used a kind of Poor Man's focusing method: run the dynamics on just the vectors associated with the current word and the previous word. Coupled with slash propagation, this technique is capable of tracking of all the dependencies needed for the current stimuli.

Figures 5 and 6 present a comparison of the predicted island effects by the SOSP model (in red) and the observed island effects (in black)—the model exhibited very little variance within trials so no model error bars are shown.[7] Indeed the qualitative behavior of the model matched the desiderata we have mentioned, often succeeding in linking the gap to the fronted element in D-linked whether islands, and in extractions from non-islands, but not in the subject islands, and rarely in the non-D-linked whether islands.

[7]For subject islands, Sprouse & Messick (2015) report a reverse effect of D-linking, with D-linked subject islands showing a stronger interaction than non D-linked ones. However, this reverse D-linking effect appears to be driven by the non-island/long condition, which exhibited lower ratings in the non D-linked than the D-linked condition. Although it is unclear what might have driven this effect, for current purposes it is sufficient to observe that the reverse D-linking effect is not driven by the island condition itself. Moreover, the ratings for the island condition are comparable with those obtained by the non D-linked whether island, and also by the Complex NP and adjunct islands tested by Sprouse & Messick (not reported here), for which no reverse D-linking effect was observed.

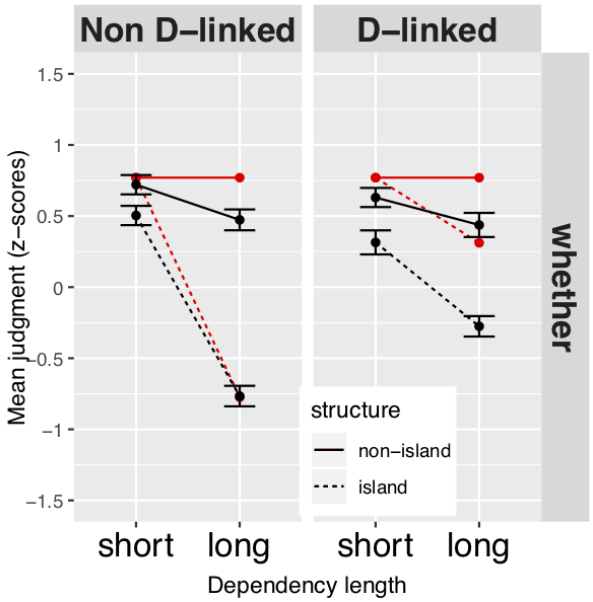

Figure 5: Interaction plots for whether island. The points correspond to the 4 conditions in (5). Empirical results are in black (data from Sprouse & Messick, 2015) and results from the model's simulation are in red.

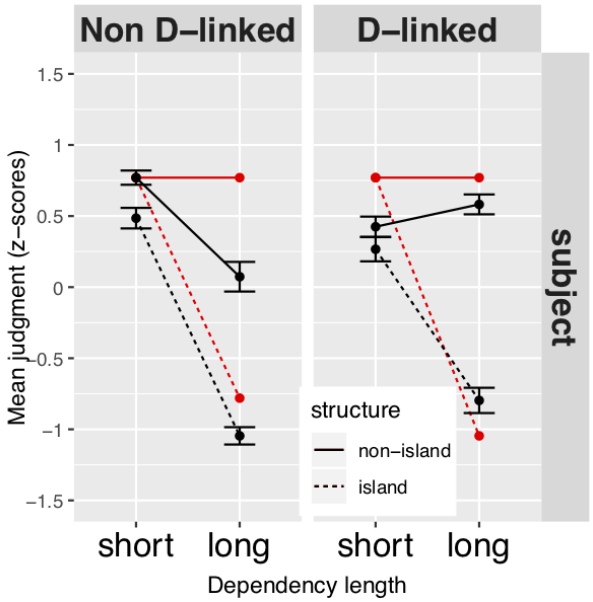

Figure 6: Interaction plots for subject island. The points correspond to the 4 conditions in (6). Empirical results are in black (data from Sprouse & Messick, 2015) and results from the model's simulation are in red.

Discussion

We reported three empirical findings from the literature pointing to gradient effects in island acceptability. We presented a new way to account for islands' ungrammaticality in a self-organized sentence processing (SOSP) framework.

SOSP's key novelty lies in its conception of grammatical states as lying in a continuum of grammaticality values. As a consequence, and unlike traditional theories of grammar, SOSP treats degrees of acceptability as deriving from the grammar itself, rather from extra-grammatical factors. This occurs because self-organizing treelets, not being under the control of a central coordinator, build whatever structure they can, sometimes achieving only partial coherence. This is the case of D-linked whether islands: the system succeeds in coercing them into a non-island structure, leading to the propagation of the slash feature inside the island, thus rendering these structures interpretable, in line with empirical findings. However, coercion comes at a cost, which is what causes the sentence to be given a suboptimal harmony value by the model, thus accounting for the fact that D-linked whether islands, although improved as compared to non D-linked ones, are still degraded. Importantly, the system is also able to generate extreme grammaticality values, in line with classical models. On the ungrammatical side, this happens when no grammatical parse is available and no coercion can take place, either because no grammatical structure is similar-enough to the to-be-parsed structure or because the system is not sufficiently prompted in undergoing the coercion. The first case is illustrated by subject islands, where no alternative (coerced) parse is available. The second case is illustrated by non D-linked whether islands: here the non D-linked wh-phrase is not powerful enough to cause the system to discover the coercion, resulting in failure to propagatied the slash feature inside the island, and very low harmony value.

Shortcomings of the current model are that the treelet forms are based on linguistic theorizing, not on a machine-learning method. A machine learning approach would make the method more completely formalized. Also, the feature vector composition, which ends up determining the harmony values, was mainly random. It will be valuable to explore more realistic feature analyses motivated by linguistic theory. Finally, as noted above, it is desirable to find a more principled method of keeping the state space finite.

All in all, we argue that SOSP offers a valuable new way of approaching the relationship between grammar and processing. It is closely related to generative linguistic theory. Nevertheless, it differs in non-trivial ways from traditional assumptions, notably continuity, and a central role for processing in grammatical explanation. We hope our results will spur new discussion on these topics.

References

Cho, P. W., Goldrick, M. A., & Smolensky, P. (2017). Incremental parsing in a continuous dynamical system: Sentence processing in dradient symbolic computation. Linguistic Vanguard, 3, 76-96.

Cinque, G. (1990). Types of Ā-dependencies. MIT press.

Fillmore, C., Kay, P., & O'Conner, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. Language, 64, 501–538.

Fodor, J. D. (1998). Unambiguous triggers. Linguistic Inquiry, 29, 1–36.

Gazdar, G. (1981). On syntactic categories. Philosophical Transactions (Series B) of the Royal Society, 295, 267-83.

Kempen, G., & Vosse, T. (1989). Incremental syntactic tree formation in human sentence processing: a cognitive architecture based on activation decay and simulated annealing. Connection Science, 1, 273–290.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In Proceedings of the 13th conference on empirical methods in natural language processing (p. 234-243).

Muezzinoglu, M. K., & Zurada, J. M. (2006). RBF-based neurodynamic nearest neighbor classification in real pattern space. Pattern Recognition, 39, 747–760.

Plate, T. A. (2003). Holographic reduced representation: Distributed representation for cognitive structures. Stanford, CA: CSLI Publications.

Rizzi, L. (1990). Relativized minimality. MIT press.

Ross, J. R. (1967). Constraints on variables in syntax. (Ph.D. thesis, MIT)

Smith, G., & Tabor, W. (2018). Toward a theory of timing effects in self organized sentence processing. In Proceedings of the 16th international conference on cognitive modeling (p. 138-143).

Smolensky, P. (1986). Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & the PDP Group (Eds.), Parallel distributed processing, volume I (pp. 194–281). MIT Press.

Sprouse, J., & Messick, T. (2015). How gradient are island effects?. (Poster presented at NELS 46)

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. Language, 82-123.

Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. Expert Systems: The International Journal of Knowledge Engineering and Neural Networks, 17(1), 41-56.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging in effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(2), 431-450.

Villata, S., Rizzi, L., & Franck, J. (2016). Intervention effects and relativized minimality: New experimental evidence from graded judgments. Lingua, 179, 76-96.

Villata, S., Sprouse, J., & Tabor, W. (2018, March). Modeling ungrammaticality: A self-organizing model of islands. (Poster presented at the CUNY conference on sentence processing)

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. Frontiers in psychology, 9(2).