Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

A self-organized sentence processing theory of gradience: The case of islands

Sandra Villata [a,*], Whitney Tabor [b]

[a] Department of Linguistics and Department of Psychological Sciences, University of Connecticut, USA
[b] Department of Psychological Sciences, University of Connecticut, USA

ARTICLE INFO

Keywords:
Gradience
Islands
Coercion
Ungrammaticality
Self-organized sentence processing model
Continuous grammar

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 effects (sentences receive gradient acceptability judgments, speakers report a gradient ability to comprehend sentences that deviate from idealized grammatical forms, and various online sentence processing measures yield gradient effects). 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. SOSP generates structures with a continuous range of grammaticality values by assuming a flexible structure-formation system in which the parses are formed even under sub-optimal circumstances by coercing elements to play roles that do not optimally suit them. We focus on islands, a family of syntactic domains out of which movement is generally prohibited. Islands are interesting because, although many linguistic theories treat them as fully ungrammatical and uninterpretable, experimental studies have revealed gradient patterns of acceptability and evidence for their interpretability. We describe the conceptual framework of SOSP, showing that it largely respects island constraints, but in certain cases, consistent with empirical data, coerces elements that block dependencies into elements that allow them.

1. Introduction

Virtually every language scientist would agree that acceptability judgments are gradient. This is manifest in the proliferation of symbols like *?,??,?*,*?,*,**,*** to express the various degrees of acceptability that a given sentence can take. Moreover, with the development of finer-grained techniques for acceptability judgment gathering, it has become clear that, when participants are asked to judge the acceptability of a sentence on a Likert scale, judgments do not cluster around the end points of the scale, but they range through the whole scale. Of course, this pattern could, in principle, be the output of a core, binary grammar system which has noise added to it, in which case judgments should distribute bimodally. But efforts to detect bimodality have produced mixed results – in some cases finding strong evidence for a lack of bimodality (see Dillon, Staub, Levy, & Clifton Jr, 2017; Dillon, Andrews, Rotello, & Wagers, 2019; Dillon & Wagers, 2019; see also Clifton Jr & Staub, 2008; Levy, 2008). Here we contribute to the effort to precisely distinguish between the options by articulating a new model of grammar (as well as processing) which produces structures with intermediate grammaticality values.

Differences in acceptability, even subtle ones, play a central role in linguistic theorizing. One well-known example is the strong/weak island distinction, discovered in the late 1960s starting from the observation that some extractions from islands are more acceptable than others (Ross, 1967). The central question about gradience pertains to its source: does gradience come from a gradient grammar or does it come from extra-grammatical factors? Chomsky may be the earliest proponent of the idea that gradience in acceptability could be the reflection of gradience in the grammar. He claimed that at least three levels of ungrammaticality can be distinguished: lexical category violations, subcategorization violations, and selectional restriction violations (Chomsky, 1965). On this view, whereas there are no degrees of grammaticality, there are degrees (although discrete) of ungrammaticality.

Despite its early recognition, gradience has not played a major role in the development of generative grammar. This is very much in line with the rule-based approach characterizing almost all work to date in this framework: grammar is a collection of rules; rules can either be followed or violated, producing a binary divide with no room for intermediacy.

* Corresponding author at: Department of Linguistics, University of Connecticut, Oak Hall, 365 Fairfield Way, Storrs, CT 06268, USA.
E-mail address: sandra.villata@uconn.edu (S. Villata).

https://doi.org/10.1016/j.cognition.2021.104943
Received 30 April 2021; Received in revised form 1 September 2021; Accepted 18 October 2021

This approach led to what we will refer to as the *traditional account*: while grammar is essentially dichotomous, thus being responsible for the divide between grammaticality and ungrammaticality, extra-grammatical factors are claimed to be responsible for the intermediate modulations observable in experimental measures. This account successfully captures a number of linguistic and psycholinguistic facts. Possibly the case par excellence is double center-embedding (e.g. *The rat the cat the dog chased ate died*, Miller & Chomsky, 1963). These sentences, although arguably grammatical, are perceived as unacceptable, although some seem better than others (Lewis, 1996). The traditional account captures these facts by claiming that the unacceptability of these sentences derives from our limited working memory capacity, which puts a bound, possibly somewhat noisy, on the number of unresolved long-distance dependencies that can be simultaneously kept in memory. Another case that can be accounted for by the traditional account is garden-path sentences: the parser is fooled by a temporary ambiguity that leads it to generate a parse that will turn out to be incorrect (e.g. *When the men hunt the birds that cheetahs eat typically scatter*). Studies that have examined the acceptability ratings of garden-path sentences report that these sentences, although fully grammatical, are rated lower than non-garden path ones (see a.o. Ferreira & Henderson, 1991; Tabor & Hutchins, 2004). The lower acceptability of garden-path sentences may be attributed to extra-grammatical, processing-related factors: even in cases where the processor successfully recovers from the garden path, reanalysis may come with a cost that mildly lowers acceptability. A third example is semantic violations in otherwise grammatical sentences (e.g. *Sheila drank a cup of porcelain*). Several studies have shown that syntactic and semantic violations have different profiles (e.g. different time courses, different timing detection, different ERPs components; see a.o. Braze, Shankweiler, Ni, & Palumbo, 2002; De Vincenzi, Job, Di Matteo, Angrilli, Penolazzi, Ciccarelli, & Vespignani, 2003; Fodor, Ni, Crain, & Shankweiler, 1996; Friederici, Pfeifer, & Hahne, 1993; McElree & Griffith, 1995; Ni, Fodor, Crain, & Shankweiler, 1998), and this has been taken to suggest that the parser has different processing routines for syntactic and semantic violations. Under this view, semantic violations are handled in the following way: the parser first generates a well-formed syntactic parse; when the meaning of the sentence is computed, however, a semantic clash occurs. We assume that in such cases, participants judge the strength of the semantic clash, and this can take on a range of values, corresponding, in many cases, to probability in the world, which clearly varies more or less continuously. This approach to intermediacy is implicit in many "semantic approaches" to constraints on extraction (e.g. Abrusán, 2011, 2014; Brown, 2015, 2017; Davies and Dubinsky, 2003; Philip & de Villiers, 1992; Szabolcsi & Zwarts, 1990, 1993; Truswell, 2017; Villata, 2017). A fourth example is drawn from the literature on similarity-based interference effects. It has been observed that the acceptability of grammatical long-distance dependencies, such as object relative clauses, is reduced as a function of the syntactic and semantic similarity between the to-be-retrieved element and other irrelevant memory representations (e.g. Friedmann, Belletti, & Rizzi, 2009; Gordon, Hendrick, & Johnson, 2001, 2004; King & Just, 1991). This reduction in acceptability has been claimed to be the result of increased processing difficulties during the retrieval of the distant element: similarly to the garden path case, the temporary disruption caused by interaction with the interfering element is hypothesized to weaken a participant's overall impression of a sentence, even if they ultimately parse it correctly (e.g. Friedmann, Belletti, & Rizzi, 2009; Gordon, Hendrick, & Johnson, 2001; Hofmeister & Vasishth, 2014; Lewis & Vasishth, 2005; McElree, Foraker, & Dyer, 2003; Van Dyke, 2002, 2007; Van Dyke & Lewis, 2003; Villata & Franck, 2019; Wagers, Lau, & Phillips, 2009). Intriguingly, the same goes for ungrammatical long-distance dependencies, the acceptability of which can be improved if the to-be-retrieved element is made more dissimilar from other elements in memory (e.g. Atkinson, Aaron, Kyle, & Omaki, 2015; Villata, Rizzi, & Franck, 2016). Similarity-based interference effects can thus be seen as processing effects that act upon a binary grammar: they can lower the acceptability of an otherwise grammatical sentence or increase the acceptability of an otherwise ungrammatical one as a function of the ease with which a given element is retrieved[1] from memory.

Notwithstanding that the traditional account successfully captures several phenomena, it also has a number of shortcomings. First, a fundamental principle in science is the principle of parsimony (Ockham's Razor): a single explanation should be preferred over a plurality of explanations. Under the traditional account, each of the phenomena described above requires a different explanation (e.g. working memory limitations, reanalysis, semantics, interference). The principle of parsimony says that if an alternative model is able to account for the same facts with a smaller number of explanations, it should be preferred. Second, while the theory of grammaticality is formalized, the theory of gradient acceptability is not. Without a specification of the scale of measurement associated with each source of acceptability adjustment, it is difficult, if not impossible, to define a linking hypothesis between the grammar and extra-grammatical factors, and, as a result, to establish how much a given extralinguistic factor is expected to degrade a grammatical sentence or improve an ungrammatical one. Relatedly, although accounts of grammar and parsing for grammatical sentences have been relatively well worked-out, the cognitive mechanisms by which extra-linguistic modules generate acceptability adjustments have not, as far as we know, been fleshed out.[2] Fourth, under the traditional account, the computation stops as soon as a syntactic violation is encountered. But if so, it is unclear how the meaning of an ungrammatical sentence is computed. Lastly, there is a rich, quantitative theory of applied language analysis which makes use of vector spaces (e.g. Bengio, Ducharme, Vincent, & Jauvin, 2003; Levy & Goldberg, 2014; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Smith & Vasishth, 2020). Gradient well-formedness is naturally related to distance in such vector spaces. At present, the relation between the traditional view and these computational approaches to sentence processing is opaque, largely because it is not clear how to interpret the vector space representations of words in relation to hierarchical analyses of sentences (though see Tabor, 2000, 2009, 2011, 2021). Given that such techniques have recently helped achieve significant advances in the structural sensitivity of applied Natural Language Processing and they also provide a theoretical link to neural mechanisms, it is desirable to have principled insight into the relationship between these methods and linguistic representations. The theory we will describe here takes a step toward providing such insight by showing how empirically justified gradient acceptability judgments arise in an interactive activation model, a close relative of vector space approaches (McClelland, 2013; McClelland, Mirman, Bolger, & Khaitan, 2014; for a related effort to link traditional and vector-space language models, see Asher, Hunter, Morey, Benamara, & Afantenos, 2016).

[1] Assuming that encountering a gap initiates a memory search to retrieve the antecedent, it is worth asking on which representation the retrieval operation occurs when the long-distance dependency is ungrammatical. If retrieval operations are constrained by the grammar, then no dependency formation should be attempted in ungrammatical sentences, not even in those that appear to get a compositional interpretation. Alternatively, the traditional view may assume an extra-grammatical repair mechanism which rescues these sentences, thus allowing dependency formation, and the resulting interpretation. This view might also assume that the repair process would be more likely to succeed if similarity-based interference is low.

[2] For the case of the production of ungrammatical utterances due to interference effects, some formal headway has been made. The Marking & Morphing theory of agreement attraction phenomena (e.g. Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001; Eberhard, Cutting, & Bock, 2005) posits the averaging of probabilities specified by grammar with probabilities specified by a pragmatic module. This theory, however, does not propose a mechanism for the derivation of these probabilities.

Our alternative approach to empirical gradience claims that significant gradient variation derives from the grammar itself. We will refer to the class of theories that take this approach as *gradient grammar theories*. Gradient grammar theories do not claim that all gradient variation in judgment is due to gradience in the grammar – rather they attempt to explain cases in which gradient judgment values are correlated with linguistic properties of the stimuli. Several theories fall under this umbrella. For instance, Ross (1972) and Lakoff (1973) suggest that gradience derives from flexible constraints that can be violated or satisfied to different degrees (Fuzzy Grammar). Keller (2000) proposes Linear Optimality Theory (LOT), an extension of Optimality Theory (e.g. Prince & Smolensky, 2008; Smolensky, Goldrick, & Mathis, 2014), which is specifically designed to account for acceptability gradience (see also Sorace & Keller, 2005). In LOT, constraints come with different numeric weights and the grammaticalness of a given structure is proportional to the sum of the constraints it violates. Featherston (2005) proposes the *Decathlon Model*, in which constraint ranking is combined with a probabilistic selection mechanism, which has the advantage of accounting for occasional production of suboptimal candidates. A number of applied Natural Language Processing projects have developed gradient models of grammaticality, generally for the purpose of processing grammatically erroneous strings in corpora (see Foster, 2007, for review). Relatedly, several extensions of classical unification-based grammar modeling approach grammatical gradience via constraint relaxation (Douglas & Dale, 1992; Fouvry, 2003; Vogel & Cooper, 1995). Lastly, several authors have explored the ability of neural network models trained on unparsed corpus data to capture grammatical gradience (Lau, Clark, & Lappin, 2017).

In this paper, we will explore the consequences of adopting a dynamical gradient grammar model to account for gradient effects, the self-organized sentence processing–grammar derived model (SOSP-GD; see Kempen & Vosse, 1989; Smith & Tabor, 2018; Smith, Franck, & Tabor, 2021; Stevenson, 1994; Villata, Sprouse, & Tabor, 2019; Tabor & Hutchins, 2004; van der Velde & de Kamps, 2006; Vosse & Kempen, 2000, 2009). SOSP-GD, a close relative of Harmonic Grammar and Gradient Symbolic Computation theory (Cho, Goldrick, Lewis, & Smolensky, 2018; Cho, Goldrick, & Smolensky, 2017, 2020; Müller, 2019; Smolensky & Goldrick, 2016), is a self-organizing sentence processing system that is derived from a probabilistic context free grammar. The model is a kind of interactive activation (i.e. recurrent neural network) model for sentence processing (Cottrell & Small, 1983; McClelland & Rumelhart, 1981) that is built out of syntactic-semantic linguistic parts. It offers an account of gradient effects through noisy, locally-driven interactions between sentential elements (treelets). These interactions are in the service of parse formation for the purpose of sentence interpretation, the main task of the sentence processing system. SOSP-GD is not constrained to only construct perfectly grammatical structures. Instead, when no perfect attachment between treelets is available, a suboptimal structure is built by coercing bonds to form anyway (Section 3; Fodor & Inoue, 1998). This can happen in the parsing of ungrammatical sentences, and it may happen in the parsing of a grammatical sentence if the grammatical parse has a difficult time forming – for example, in garden path and local coherence structures.[3] SOSP-GD is an optimization system that strives to maximize a sentence's well-formedness in the service of parsing. Hence, a distinctive strength of SOSP-GD is that it derives gradient acceptability predictions from independently motivated assumptions about grammar and parsing.

Our account rests on two concepts, self-organization and coercion, that have rather separate scientific histories but, we suggest, are closely related. Self-organization is a rare phenomenon of spontaneous order formation that, despite its rarity, occurs in a wide range of settings (Haken, 2008). It can be defined as any situation in which multiple, small, interacting parts give rise, strictly through their interactions, to organized structure at the scale of the group. Well-studied cases include superorganisms (e.g. social insects – Gordon, 2010, slime molds – Keller & Segel, 1970; Marée & Hogeweg, 2001), far-from equilibrium chemical processes (e.g. the Belousov-Zhabotinsky reaction – Zhabotinsky, 1991), and certain energy-diffusion processes in fluid dynamics (e.g., Rayleigh-Bénard convection – Koschmieder, 1993). Self-organization is a natural place to look for insight about complex linguistic phenomena because its study has shed light on the sources of complex order in many other domains (e.g. Haken, 2004; Kauffman, 1993; Nicolis & Prigogine, 1977). Here, we describe how the (standard) linguistic parts combine to form structures that range continuously in grammaticality, and we outline the self-organizing dynamics that these structures participate in. As a step toward constructing the full dynamical model, we employ a classical parsing algorithm that builds the inventory of semi-grammatical and grammatical structures. Coercion in linguistics refers to cases of morphological or syntactic combination in which some elements play roles which stretch their normal proclivities (see Lauwers & Willems, 2011 for a review). Standard examples include cases like "Mary began the book" (interpreted to mean she began reading the book, since "began" needs an event argument), or "Harry was hiccupping" (interpreted as iterated hiccupping because the host of a progressive form must be an ongoing event, a possibility that is outside the repertoire of a single hiccup). Coercion goes hand in hand with self-organization because the process by which the small elements come together to form a structured whole is a process of adjusting their autonomous trajectories so that they mesh with the global scheme. We claim that this is the core mechanism that gives rise to gradient grammaticality. In this paper, we do not go into formal detail on the continuous temporal dynamics (but see Section 4.3.1 for an overview.) Nevertheless, self-organization has never been observed without (at least approximately) continuous feedback dynamics. In this sense, our central claim is a dynamical one. Unlike state-of-the-art neural network models for Natural Language Processing (NLP; e.g. Devlin, Chang, Lee, & Toutanova, 2019) which learn linguistic encodings from raw word sequences, our model imports a great deal of linguistic structure from generative linguistic theory. For the most part, our linguistic assumptions are uncontroversial, but one of them is related to a long-standing debate among generative linguists: we prefer slash-propagation (Gazdar, 1982) over successive cyclic movement (e.g, Chomsky, 1973) for modeling long-distance dependencies for a reason that is directly tied to the nature of self-organization – see the discussion of "particle dynamics" in the introduction to Section 4 below.

Our model has several commonalities with the gradient grammar approaches outlined above. For instance, Linear Optimality Theory (LOT – e.g. Keller, 2000; Sorace & Keller, 2005) computes the grammaticality of a structure based on the sum of the weights of the constraints it violates. This is related to our coercion approach: SOSP-GD computes the global grammaticality of a sentence as the product of its individual bond *harmonies* (a local measure of well-formedness).[4] Thus, for a given string of potential violations, the more of them occur, the worse the sentence is predicted to be under SOSP-GD; indeed, products can be turned into sums by taking logs, making the system reminiscent of LOT (Sorace & Keller, 2005). To anticipate a bit, the

[3] In SOSP-GD, suboptimal parse formation can also fail in the case of ungrammatical sentences. This can happen, for example, if there is a grammatical alternative that is semantically motivated (cf. Bailey & Ferreira, 2003; Gibson, Piantadosi, & Fedorenko, 2013).

[4] We use a product rather than a sum because we use harmony to predict degrees of acceptability. We hypothesize that adding more well-formed structure to a sentence containing a violation is powerless to overcome the effect of a single ill-formed element on degree of acceptability. In fact, adding more structure to a sentence generally decreases rather than increases its acceptability, even if the added structure is well-formed (Lau, Clark, & Lappin, 2017).

LOT distinction between soft and hard constraints (Keller, 2000; Sorace & Keller, 2005; but see also Legendre, Wilson, Smolensky, Homer, & Raymond, 1995) roughly matches our distinction between interpretable and uninterpretable coercion (see Section 3). As for the Decathlon model (Featherston, 2005), it consists of two components: a Constraints Application module, which applies constraints and assigns violation costs in a cumulative fashion, similarly to LOT; and an Output Selection module, which selects the best candidate form based on weights, but it does so in a probabilistic fashion, which ensures that sometimes sub-optimal candidates are selected. This is similar to the function that noise plays in SOSP-GD. A significant difference between our model and other models of gradience, however, lies in the fact that while the latter only model grammaticality, SOSP-GD is a model of both parsing and grammaticality. As a result, SOSP-GD can generate sub-optimal parses even when a fully grammatical parse is available. This can happen in islands (Section 4.1 on Complex NP islands), as well in cases of strong garden paths in which the parser ultimately fails to generate a coherent tree. Another distinguishing feature of SOSP-GD is that it proceeds from an assumption (self-organization) about how language fulfills what are arguably its major functions: meaning generation and meaning discernment. Combined with independent observations about the properties of meanings, this assumption generates various intermediate grammaticality values. By contrast, LOT and Decathlon posit intermediate constraint weights to generate intermediate grammaticality values, but those theories do not motivate the values of these intermediate weights. There is also a close relationship between our approach and theories of syntax based on Harmonic Grammar which use intermediate constraint weights to model syntactic entities which have a mixed character (e.g. Müller, 2019; Smolensky & Goldrick, 2016). These theories generally do not address intermediate acceptability but it seems possible that they could. Lastly, while our approach has much in common with relaxation-based approaches to intermediate grammaticality mentioned above, it diverges from them in adopting a more permissive form of relaxation — that is to say, a wider range of grammatically divergent types is possible. At the same time, our formulation constrains the predicted patterns of relaxation via its assumption of self-organization. In the General Discussion, we suggest that the greater permissivity is empirically motivated while the self-organization assumption makes the theory appealingly restrictive.

As a case study for gradient effects, we investigate constraints on the formation of long-distance dependencies - *island constraints*. Island constraints are traditionally assumed to prohibit the establishment of long-distance dependencies inside certain syntactic domains. These domains are metaphorically called *islands* to capture the idea that they are inaccessible to certain peripatetic grammatical elements that otherwise have wide range. Gradience in islands has been widely recognized in the theoretical literature since at least Ross's seminal dissertation (e.g. Ambridge, 2015; Chomsky, 1986; Cinque, 1990; Erteschik-Shir, 1973; Huang, 1982; Kalyan, 2012; Kuno, 1976; Lasnik & Saito, 1984, 1992; Pesetsky, 1987; Rizzi, 1990; Ross, 1967; Starke, 2001; a.o). With the rise in popularity of formal experiments, gradience in islands has also begun to be quantified more precisely (e.g. Abeillé, Hemforth, Winckel, & Gibson, 2020; Almeida, 2014; Ambridge & Goldberg, 2008; Atkinson, Aaron, Kyle, & Omaki, 2015; Christensen, Kizach, & Nyvad, 2013; Goodall, 2015; Keshev & Meltzer-Asscher, 2019; Kim & Goodall, 2016; Kush, Lohndal, & Sprouse, 2018; Kush, Lohndal, & Sprouse, 2019; Lu, Thompson, & Yoshida, 2020; Omaki, Fukuda, Nakao, & Polinsky, 2019; Pañeda, Lago, Vares, Veríssimo, & Felser, 2020; Sprouse, 2007; Sprouse, Wagers, & Phillips, 2012; Sprouse & Messick, 2015; Sprouse, Caponigro, Greco, & Cecchetto, 2016; Stepanov, Mušič, & Stateva, 2018; Villata, Rizzi, & Franck 2016; a.o). And yet, at present, the only precise mechanisms proposed for generating island constraints are categorical or, at best, only allow discrete levels of intermediacy – e.g. the Subjacency Principle[5] (Chomsky, 1973, 1986; see also Rizzi, 1982, Torrego, 1984). SOSP-GD accounts for gradience in island extraction judgments through coercion, the use of a word in a semantic/grammatical role for which it is not perfectly suited. Since semantic features are at least very numerous (if not actually continuous), the degrees of disharmony generated by SOSP-GD effectively lie on a continuum.[6]

The purpose of this paper is to articulate and motivate the theory, not to describe its dynamical equations in detail, nor to argue for it on the basis of an extensive data set. The model is based on several standard generative linguistic mechanisms (drawing from a range of traditions within generative linguistics) whose utility in describing the clear-cut cases of acceptability judgments has been extensively demonstrated. It is also based on well-known and relatively simple mechanisms of dynamical systems theory. Despite its uncontroversial foundations, the model takes the non-standard positions (1) that grammaticality values lie on a continuum; (2) that the mechanism of self-organization is useful for modeling linguistic competence, and that, in connection with this, the radius of linguistic competence needs to be adjusted to include ungrammatical structures; (3) although we do not focus on the relevant cases here, the mechanism of self-organization is also useful for modeling well-established psycholinguistic phenomena (see first

[5] Subjacency states that movement cannot cross more than one bounding node at a time, where, for English, the bounding nodes are C(omplementizer) P(hrase) and I(nflectional) P(hrase). When we say that Subjacency is fundamentally a categorical notion, we mean that it is unclear how the property of "crossing a bounding" node could be understood as a gradient property. What would it mean, for example, to "70% cross a bounding node"? In *Barriers* (Chomsky, 1986), it is hypothesized that the degree of badness of a grammatical violation is an increasing function of the number of barriers crossed, with detectable differences between 0 (perfectly acceptable), 1 (marginal acceptable), and 2 (strong decrement in acceptability) crossings. However, this approach is still discrete in nature. Our attempt here is to provide a fully continuous account of gradience. Moreover, a mechanism underlying the hypothesized correlation between number of bounding nodes and degree of badness has, to our knowledge, never been specified and would seem to have to be of a very different character from virtually every other mechanism (e.g. check, delete, move, merge, pass, unify) posited in established theories of grammar.

[6] Our account is closely related to Categorical Grammar (e.g, Steedman, 2000), Head-driven Phrase Structure Grammar (Pollard & Sag, 1994), Construction Grammar (Goldberg, 1995, 2006), and the Parallel Architecture (Jackendoff, 2002) in that it posits that every syntactic construction (here referred to as a "treelet") has a semantic profile which supplements its syntactic profile. In fact, our proposal for an important role of semantically-mediated coercion in the theory of syntax is prefigured in a number of prior works – e.g., Goldberg, 1995; Michaelis, 2004; Moens & Steedman, 1988; Sag & Pollard, 1991; Piñango, Zurif, & Jackendoff, 1999. Our approach differs from these treatments in a subtle, but, we think, important way: virtually all of these accounts are concerned with handling sentences which violate what would have been very clean, formal generalizations about the set of well-formed expressions in a language, were it not for their presence in the language. Much of the discussion has focused on the question of how to formulate a generation mechanism which allows these anomalous-seeming cases but does not overgenerate the many conceivable coercions that do not produce grammatical results (e.g., **Sam encouraged Bob into the room* – Goldberg, 1995, p. 166). With the exception of discourse-based, "backgrounding" accounts (Abeillé, Hemforth, Winckel, & Gibson, 2020; Erteschik-Shir, 1973; Goldberg, 2006) such theories assume a binary divide between acceptable and unacceptable sentences. Our account, however, does not make such a stark divide: it suggests that even cases of what are deemed "licit coercions" may have subtly deficient well-formedness, and it posits a continuum in representation space between these and the severe coercions which produce clear ungrammaticality. We are partly motivated to take this approach by psycholinguistic studies showing that even licit coercions (e.g., *The author began the book*) produce a subtle, but experimentally detectable delay in processing, relative to non-coerced controls (Traxler, Pickering, & McElree, 2002) – see also section 5.5.

mention of "SOSP-GD" above). In sum, we view the approach as a novel claim about the nature of the relationship between grammar and processing.

The paper is structured as follows. In Section 2, we introduce the island phenomenon and review some of the recent empirical evidence pointing to gradient effects in islands. In section 3, we provide an overview of SOSP-GD. In Section 4, we explain how SOSP-GD handles the data, and Section 5 contains a General Discussion.

2. The case of islands

Natural languages allow for long-distance dependencies (LDDs), also called *filler-gap dependencies*, i.e. syntactic and semantic relations between two constituents that are far away in the sentence. An illustration is provided in (1): *what* is the semantic object of the verb *cooked*, but the two words are separated by two intervening words because *what* has been "moved" to the front of the sentence in order to form the question (the underscore indicates the original position of the moved constituent, also referred to as a *gap*). An interesting aspect of LDDs is that they are unbounded - there is no upper limit to the number of words by which the two elements of the dependency can be separated. Although sentences can become very cumbersome as the distance between the filler and the gap increases, distance per se does not undermine the grammaticality, as illustrated in (1–3).

(1) What did Bart cook _?
(2) What did Lisa think that Bart cooked _?
(3) What did Brian say that Lisa thought that Bart cooked _?

Although long-distance dependencies are not constrained by length, they are constrained by structure. Since Ross (1967), it has been known that certain structural domains resist the formation of a dependency between a gap inside of the domain and a filler outside of it. These are the "islands" mentioned above. Examples of islands include Whether clauses (4a), Complex noun phrases ("CNP"s) (4b), and *because*-Adjunct clauses (4c) (see Phillips, 2006 for an overview). The square brackets define the island domain, and the asterisk indicates sentence unacceptability.

(4) a. *$\textbf{What}_i$ did you wonder [whether the student solved $\__i$]?
WHETHER ISLAND
b. *$\textbf{What}_i$ did you hear the statement [that the student solved $\__i$]?
CNP ISLAND
c. *$\textbf{What}_i$ did you smile [because the student solved $\__i$]?
BECAUSE-ADJUNCT ISLAND

A key distinction in the literature on islands is the strong/weak island distinction (Szabolcsi & Lohndal, 2017). Strong islands are claimed to be *unselective*, preventing the establishment of all dependencies. For example, strong islands equally disallow the extraction of simple wh-words (e.g. *who*, *what*) and complex wh-phrases[7] (e.g. *which problem*). Weak islands, on the contrary, are claimed to be *selective*: they allow the establishment of certain dependencies inside of them, but not others. In particular, weak islands have been claimed to allow the extraction of complex wh-phrases, while disallowing the extraction of simple wh-words (e.g. Cinque, 1990; Pesetsky, 1987; Rizzi, 1990). This is illustrated by comparing sentences in (4) with the corresponding sentences in (5): while the extraction of *which problem* has been claimed to result in a fully grammatical sentence for Whether islands (5a) (and possibly CNP islands (5b) as well),[8] the same extraction is disallowed in Adjunct islands (5c). The extractability of complex wh-phrases has thus been traditionally used as a diagnostic for the nature of islands (strong vs. weak). This is known as the *D(iscourse)-linking effect* – where D-linking refers to complex wh-phrases of the form *which NP*, which refer to a set of entities that are pre-established/presupposed in the discourse (Pesetsky, 1987).

(5) a. **Which problem$_i$** did you wonder [whether the student solved $\__i$]?
b. **Which problem$_i$** did you hear the statement [that the student solved $\__i$]?
c. ***Which problem$_i$** did you smile [because the student solved $\__i$]?

2.1. Gradience in islands

The sharp distinction between strong and weak islands is in line with the traditional view of grammar, which only admits binary outcomes. However, in more recent years, with the development of refined techniques for gathering acceptability judgments, it has become increasingly clear that the strong/weak island distinction might be more nuanced than initially thought. In particular, there is a growing body of empirical evidence suggesting that the extraction of complex wh-phrases from weak islands does not result in a fully acceptable sentence, as initially supposed, but in an intermediate judgment (e.g. Kush, Lohndal, & Sprouse, 2018; Sprouse, Caponigro, Greco, & Cecchetto, 2016; Atkinson, Aaron, Kyle, & Omaki, 2015; Villata, Rizzi, & Franck 2016). In other words, the extraction of complex wh-phrases from weak islands reduces the strength of the violation, but it does not eliminate it.

Much of this empirical work uses the so-called *factorial design for island effects* (see Sprouse, 2007; Sprouse, Wagers, & Phillips, 2012; Sprouse, Caponigro, Greco, & Cecchetto, 2016), a design devised to isolate the island effect from two factors that can potentially interact with it: the presence of a complex syntactic structure in the embedded clause (an island structure, like a whether clause), and the presence of a long dependency (e.g. a wh-dependency). By crossing the factor *structure* (non-island vs. (contains) island) with the factor *dependency length* (short vs. long) the factorial design is meant to isolate the contribution of the effect of violating the island constraint. As such, the island effect will appear as a statistical interaction between the two factors. The factorial design is illustrated in (6) for Whether islands (the same logic applies to all other island types): the contrast between (6a) and (6c) isolates the cost of *structure*, while the contrast between (6a) and (6b) isolates the *dependency length* effect.

(6) Factorial design measuring the Whether island effect.

a. NON-ISLAND, SHORT.
Who _ thinks that John bought a car?
b. NON-ISLAND, LONG.

[7] Complex wh-phrases such as *which problem* are also referred to in the literature as "D(iscourse)-linked" or "lexically restricted" phrases. D-linking and lexical restriction emphasize different facets of the contribution of the lexical element (e.g. *problem* in *which problem*) in the modulation of the so-called intervention effects (Rizzi, 1990; Starke, 2001; Rizzi, 2004 and much subsequent work). The term "D-linking" emphasizes the semantic contribution of the lexical element (e.g. *which problem* presupposes the existence of a salient set of problems), while the term "lexical restriction" emphasizes its syntactic contribution (see Villata, Rizzi, & Franck 2016 for a discussion). In this work, we do not attempt to tease apart the contribution of each of these factors in the modulation of intervention effects. Hence, we use the term "complex wh-phrase" to adopt an agnostic position.

[8] The status of CNP islands is less clear cut than the status of Whether and Adjunct islands. Ross (1967) already noticed differences among complex noun phrases in their potential to block movement. For instance, he claimed that *make the claim* was not an island for movement. We will come back to the status of CNP islands in the next section.

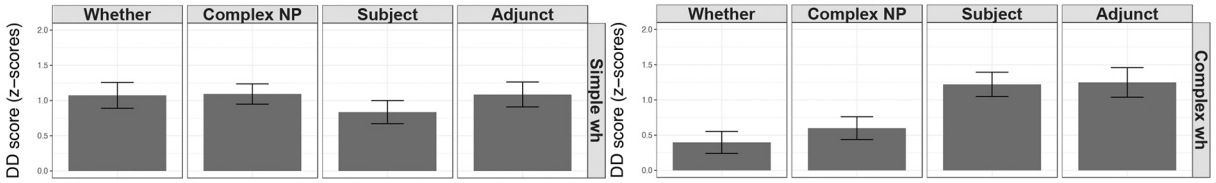

Fig. 1. The left-hand plot shows the size of the island effect with simple wh-words in DD scores for the four island types, while the right-hand plot shows the size of the island effect with complex wh-phrases (data from Sprouse & Messick, 2015). DD scores measure the interaction term. A DD score of 0 means no island effect, a positive DD score indicates an island effect, and the height of the bar indicates how strong the island effect is (the higher the bar, the stronger the island effect).

What do you think that John bought _?
c. ISLAND, SHORT.
Who _ wonders whether John bought a car?
d. ISLAND, LONG.
What do you wonder whether John bought _?

Sprouse and Messick (2015) tested 4 island types (Whether, CNP, *because*-Adjunct, and Subject islands) and two dependency types (simplex wh vs. complex wh) on a 7-point Likert scale in English. They observed an island effect for all island and dependency types. Critically, however, this effect was reduced (but not absent) in both Whether and CNP islands with complex wh-phrases. This is illustrated in Fig. 1. The y axis reports the island effect in D(ifference in) D(ifference) scores (Maxwell & Delaney, 2003), which are calculated by subtracting the difference between two conditions related by one factor from the difference between the two conditions related by the other factor (7):

(7) DD = ([Non-island Short] - [Island Short]) – ([Non-Island Long] – [Island Long]) = [6a-6c]-[6b-6d]

Several points are worth noting. First, according to the D-linking diagnostic, both Whether and CNP islands are weak islands, as the size of the island effect is reduced when a complex wh-phrase is extracted as compared to when a simple wh-word is extracted (the reduction in acceptability was confirmed by the results of the statistical analyses reported in Sprouse & Messick, 2015; see also Sprouse, Caponigro, Greco, & Cecchetto, 2016). Second, no difference in the island effect size as a function of the type of the extracted element (simple vs. complex) was attested either for Adjunct or for Subject islands, confirming the strong status of these islands. Third, the extraction of a complex wh-phrase from Whether and CNP islands does not result in a fully grammatical sentence, but in a semi-grammatical one, contrary to the traditional wisdom. Hence, the strong/weak distinction appears to be less sharp than initially thought: rather than being a divide between islands that always result in full ungrammaticality and islands that sometimes result in full grammaticality, it appears that weak islands show intermediate judgments within a single construction.

These results have been recently replicated for Whether, CNP, and *because*-Adjunct islands in a series of acceptability judgments experiments (Villata, Tabor, & Sprouse, 2020a, 2020b). This replication effort largely reproduces the results from Sprouse and Messick (2015), but the reduced island effect in CNP islands with complex wh failed to replicate. Villata, Tabor, & Sprouse (2020a, 2020b) also conducted another set of experiments using a different procedure, the maze task (e.g. Forster, Guerrera, & Elliot, 2009). The maze task is similar to the moving-window word-by-word self-paced reading, except that participants are asked to make a choice between two possible continuations of the sentence at each word in the sentence (e.g. The | gone / dog | chased / sink | our /hosed | into. / cat.). The goal was to investigate whether higher acceptability rates in islands corresponds to an increased willingness to posit a gap inside of the island domain: if islands with higher acceptability rates turn out to be those for which participants are more willing to establish an island-violating dependency, this would suggest that higher acceptability rates are the result of comprehenders' ability to interpret the sentence. To test this hypothesis, they designed an adapted version of the maze task in which participants read the sentence word-by-word by pressing a button to make the next word appear up to the critical potential gap position where they are asked to make a choice between two possible continuations. For instance, for a Whether island this will happen right after the embedded verb *solved* in *What do you wonder whether the candidate solved…*, as illustrated in Fig. 2. At this point, participants were shown two choices for how to continue the sentence: they could either select a determiner (*the*), which is compatible with not forming a dependency inside of the island, generating the sequence, *What do you wonder whether the candidate solved* ***the*** *problem before the interview?*, or they could select a preposition (e.g. *before*), which is compatible with establishing a dependency inside of the island, generating the sequence *What do you wonder whether the candidate solved* ***before*** *the interview in Paris?*. After making the determiner/preposition choice, the participants continued to choose between the words instantiating the two continuations up until the end of the sentence. The results supported the hypothesis: island types that receive higher acceptability rates under extraction are also the types for which participants are more prone to choose the prepositional continuation, suggesting that a gap inside the island has been posited. In particular: (i) all islands showed fewer "gap continuations" than grammatical long-distance dependency controls for which a gap is expected (e.g. *What do you think that the candidate solved before the interview?*); however, (ii) all island types (except for *because*-Adjunct islands with simple wh) showed more gap continuations than yes/no grammatical control conditions where no gap is expected (e.g. *Do you think that the candidate solved the problem?)*; (iii) more gap continuations in Whether islands were observed with complex wh than with simple wh, thus showing signs of intermediacy, but this effect was not attested for Complex NP islands nor for *because*-Adjunct islands; (iv) Whether and CNP islands with complex wh exhibit intermediate probabilities of gap continuation – lower than grammatical controls but higher than *because*-Adjunct islands.

Taken together, the empirical results summarized above show indications of gradience. The question that arises next is: which factor is responsible for the reduction of the island effect in the places in which we observe it? And, more generally, where does gradience come from?

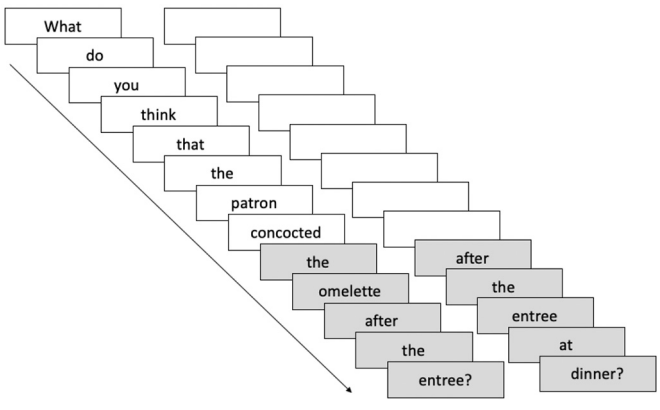

Fig. 2. Schema illustrating the maze task procedure for a Whether island.

In the reminder of this paper, we illustrate an independently motivated and parsimonious account of these gradient island phenomena in a self-organized sentence processing (SOSP) model.

2.2. Theories of islands

A "theory of islands" is a theory that says what properties of a constituent (e.g. semantic, syntactic, pragmatic) make it an island. The theory we present here is not such a theory. Instead it is a preliminary claim to such a theory, about the topology of linguistic space. Nevertheless, the topological claim – that the space is continuous rather than discrete – interfaces with actual theories of islands. Therefore, we briefly review these theories here. There are four main species: syntax-based, intervention-based, discourse-based, and processing-based. We focus on two questions that bear on our central themes: Do they predict the strong/weak distinction? Do they generate gradience in the grammar?

2.2.1. Syntax-based theories

These include Subjacency, Barriers, the Minimal Link Condition, and, most recently, Phases (Chomsky, 1973, 1986, 1995, 2000, 2001, 2008). These accounts have in common the claim that the ungrammaticality of islands is due to the violation of a formal property of the island boundary which prevents chain formation across the boundary (see Boeckx, 2012 for a review). None of these theories addresses the strong/weak distinction: a constituent is either an island or it is not. With the partial exception of Barriers (see fn. 5), these theories do not generate gradience in the grammar – instead, gradience is claimed to derive from extra-grammatical processes.

2.2.2. Intervention-based theories

Intervention-based theories claim that island effects are the result of the intervention of a blocking element between the head and the tail of a dependency. Intervention theories can be syntax-based or semantics-based. Relativized Minimality (RM; Rizzi, 1990, 2004) is syntax-based: an element counts as an intervener if it c-commands the trace of the dependency, it is c-commanded by the head of the dependency, and is of the same syntactic type of the head of the dependency (e.g. both *what* and *who* are of syntactic type Q in ****What*** *did you wonder* ***who*** *solved_?*). RM is not a gradient theory but its refinement to featural RM (e.g. Friedmann, Belletti, & Rizzi, 2009; Rizzi, 2004; Starke, 2001) made it more gradient: if the feature overlap between the intervener and the head of the dependency is partial, the detrimental effect is claimed to be milder (see Atkinson, Aaron, Kyle, & Omaki, 2015, Villata, Rizzi, & Franck 2016 for an experimental investigation). Relativized Minimality thus predicts gradient effects, but the space of gradience is discrete (a maximum of three levels of acceptability are admitted, see Villata, Rizzi, & Franck 2016). Inspired by both featural RM and unification-based approaches to gradient grammaticality (Douglas & Dale, 1992; Fouvry, 2003; Vogel & Cooper, 1995), we employ a rich feature space, positing gradient grammaticality that is proportional to the degree of match between feature vectors.

Scope Theory (e.g. Szabolcsi & Zwarts, 1993; Szabolcsi & Zwarts, 1997; Szabolcsi & Lohndal, 2017) claims that intervention is, in fact, a form of semantic failure. For instance, in negative islands (e.g. **How didn't Johnny say that Roar fixed the car?)* the intervening scope-taker (the negation) requires a particular Boolean operation (complementation) to be performed. However, this operation is undefined in the domain denoted by the *how* wh-phrase, producing a semantic clash. Our account puts important weight on semantic features in determining possible loci of interpretable coercion, and thus is broadly in line with Scope Theory. However, it remains to be determined whether the claims align in detail.

2.2.3. Discourse-based theories

Erteschik-Shir (1973), Goldberg (2006), and Abeillé, Hemforth, Winckel, & Gibson, (2020) hypothesize that extraction can only occur out of constituents that are in the focus of attention. As a result, backgrounded constructions are islands (see also Ambridge, 2015; Ambridge & Goldberg, 2008; Kalyan, 2012). Inasmuch as attention is arguably a continuum mental property,[9] these theories, predict full gradience, although they do not offer a model of continuum attention.[10] Kalyan (2012) links backgroundedness to similarity among phrases (similarity was measured experimentally). Our approach takes significant inspiration from Kalyan's work, and extends it to a parsing model which defines similarity semantically. It is less clear whether discourse-based theories can account for the strong/weak distinction (but see Erteschik-Shir, 1973, who distinguishes between two principles, the Subject Constraint and the Principle I, responsible, respectively, for ruling out strong and weak islands).

2.2.4. Processing-based theories

Processing-based theories reduce island effects to extra-grammatical factors such as working memory load (e.g. Deane, 1991; Hofmeister & Sag, 2010; Kluender, 1991; Kluender, 1998; Kluender, 2004; Kluender & Kutas, 1993). The central claim is that the ungrammaticality of island-violating sentences is not due to the violation of a syntactic constraint, but to processing costs that, summed together, exceed the capacity of our working memory system, thus resulting in the perception of ungrammaticality. These accounts naturally predict gradience because the processing costs are continuous quantities. We do not know of an application of this approach to the strong/weak distinction. In any case, we are skeptical of these theories because a growing body of empirical work employing a factorial design for island effects suggests that island effects persist even when processing costs are controlled for (e.g. Sprouse, 2007; Sprouse, Wagers, & Phillips, 2012; Sprouse, Caponigro, Greco, & Cecchetto, 2016; Kush, Lohndal, & Sprouse, 2019 a.o.).

3. The self-organized sentence processing - grammar derived (SOSP-GD) model

In this section, we describe in detailed conceptual terms how SOSP-GD works, clarifying the framework through which it generates both grammatical and semi-grammatical structures. We examine its predictions for the three island cases above (Whether, CNP, and *because*-Adjunct) focusing on the contrast between weak and strong islands. In Section 4, we describe a computational analysis which formalizes one component of the full SOSP-GD system, using traditional computational linguistic tools. This is related to previous implementations of the model (Smith, Franck, & Tabor, 2018; Smith & Tabor, 2018; Smith & Vasishth, 2020; Villata, Sprouse, & Tabor, 2019), but it is a significant advance in that it deals with entire grammatical systems rather than local processing events. The results of this analysis provide support for the claim that the theory has a principled insight into the reason that some islands (weak ones) show mixed behaviors and other islands (strong ones) do not.

In SOSP-GD, the constituent elements are treelets, small pieces of syntactic information stored in memory (see also Fodor & Inoue, 1998; Lewis & Vasishth, 2005; Marcus, 2001). Treelets are subtrees consisting of a mother and a finite number of daughter nodes. Whenever a word is perceived, a lexically anchored treelet becomes active. For instance, encountering the determiner *the* activates the lexically-anchored *D* → *the* treelet (D = determiner) (Fig. 3a). Each treelet is associated with feature vectors that specify the syntactic, semantic, and phonological features of the corresponding word and its expected dependents (Fig. 3b, the features are represented by shaded circles). Feature values are continuous and able to change within limits specified by the lexical type (e.g. the plural feature

[9] See, inter alia, Franck & Wagers (2018).

[10] Transformer neural networks (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, & Polosukhin, 2017) model attention as a learned, contextually conditioned vector which gradiently amplifies certain parts of the mental encoding and attenuates others. Our approach is closely aligned with such models.

can range over the continuum from 0 [= singular] to 1 [= plural]). If the word *the* is encountered at the beginning of a sentence, a treelet for this word gets activated, including features that specify all of its syntactic and semantic attributes (e.g. its determiner status, its definiteness, etc.). At this moment, the plural feature has an intermediate value reflecting the fact that *the* can be either singular or plural ([±PL]). With the activation of the lexically-anchored *D ➔ the* treelet, also the *DP➔ D NP* treelet becomes active.[11] The *D ➔ the* treelet will start to form a link with the D daughter of the DP treelet due to the perfect feature match between these two nodes. When the word *boy* is encountered, a new lexically-anchored treelet becomes active (*N ➔ boy*). Since *boy* is singular, the feature vector of this treelet is specified as [-PL]. As soon as the link between *N ➔ boy* and the NP daughter of the DP treelet starts to form, the number feature of the NP will gradually mutate to [-PL]. This will also cause the number feature of the determiner to gradually mutate to [-PL] (the mutation of features' values is represented in a discrete fashion in Fig. 3a, but it actually happens gradually).

A key property of SOSP-GD concerns the dynamics of link formation between treelets. Treelets interact in all ways that afford tree-fragment formation. Link formation happens continuously and competitively with small-magnitude noise, based on feature match between attachment sites: links between treelets with a good feature match grow strong faster than links between treelets with a poor feature match. This usually allows the former attachments to outcompete the latter. As a result, most of the time, for input that is well-formed in the traditional sense, a well-formed structure is generated. When all attachments perfectly satisfy the requirements of the feature vectors of all treelets, the structure attains the maximum harmony value of 1 (harmony is a measure of coherence in the system; see Smolensky, 1986). This is schematically illustrated in Fig. 4 for the sentence *The boy leaps*. The left-hand plot shows the all-to-all treelets interaction (dashed colored lines). For instance, the determiner *the* will not only try to attach to the D daughter node of the DP treelet, but also to the N daughter node of the DP treelet, the DP daughter node of the S treelet, the V node of the VP treelet, and the VP daughter node of the S treelet (see orange lines in the left-hand graph of Fig. 4). However, these latter attachments will be outcompeted by the first one because of their poor feature match. A similar dynamic happens for the *N ➔ boy* and the *V ➔ leaps* treelets. Furthermore, there is noise in the feature and link activations so there is some variation in which configuration the system will eventually settle in (sometimes, a poorer feature match can outcompete a better one because of noise).

When no perfect feature match is available, as is the case for sentences containing a violation, sub-optimal links are formed instead. This is because SOSP-GD is an *optimization system*. On average, it moves in a direction that increases the current harmony value (the harmony value is calculated as the product of the harmonies of the bonds, where the harmony of a bond is the percentage of feaures that match). When it is not possible to reach a harmony value of 1, the system tends to approach the proximal structure with the highest possible harmony value. This is achieved by forcing one or more elements to play a role that does not fully fit it. To illustrate this, consider the agreement-violating sentence *The boy sleep*, where the subject of the sentence is singular and the verb is plural (Fig. 5). The formation of the first chunk of the tree (*The boy*) will happen in the same manner illustrated above. When the verb *sleep* is reached, a *V ➔ sleep* treelet becomes active, as well as a *S ➔ DP VP* abstract treelet. The DP mother node, which is a combination of the *D ➔ the* and *N ➔ boy* treelets (from now on referred to as *DP ➔ the boy* for ease of exposition) will start to form an attachment with the DP node of the S treelet, as well as with the VP node of the S treelet. The first attachment, however, will outcompete the second because of the better feature match. At this point, since the *DP ➔ the boy* treelet is specified as [-PL], the DP daughter of the S node will also start to gravitate toward a [-PL] feature, and so will the VP daughter of the S node, since a singular DP requires a singular VP. In the meantime, the *V ➔ sleep* treelet will also start to form attachments with other treelets: it will try to form an attachment with the VP node of the S treelet, as well as with the DP node of the S treelet. Since the number feature of the VP daughter node of the S treelet has started to gravitate toward a singular number feature because of the attachment with the singular DP *the boy*, no perfect attachment is available for the *V ➔ sleep* treelet. Nonetheless, even if the features of the two V nodes do not perfectly match, this attachment will still outcompete all other possible attachments for which the feature match is even poorer. This will put some strain into the system (illustrated by the red squiggly line in Fig. 5), corresponding to a harmony value less than 1.

This example illustrates *coercion*: an element is forced to play a role that is not fully equipped to play (the plural verb *sleep* that ends up being attached as the main verb of a singular subject).[12] In some coerced structures, as the one illustrated in Fig. 5, all role-players get roles and all roles get role-players (i.e. the Theta Criterion is met). This accounts for the fact that, although the sentence is perceived as degraded to some extent, it is fully interpretable. The mismatch between the coerced treelet and the featural environment that it ends up being attached to is responsible for the fact that coerced structures are not judged as acceptable as fully grammatical ones.

We refer to the coercion phenomenon just illustrated as *interpretable coercion*: it occurs when the system forms a thematically coherent tree, despite the feature mismatch on some nodes. Coercion can also be *uninterpretable*. This happens when the coerced treelets are not able to mutate enough to match the thematic requirements of their environment. Take for instance a sentence that is fully grammatical up to a certain point and then turns into a word salad – e.g. *The boy sleeps fell*. In this example, the system still builds a coherent parse up to *sleeps* along the lines discussed above. However, there is no way in which *fell* can be coerced to form a coherent, interpretable tree. For interpretable coercion to occur, the thematic requirements of all words involved must be satisfied. For instance, in the agreement error example *The boy sleep*, *sleep* [V, +PL, ...] can be easily coerced to play the role of *sleeps* [V, -PL, …]: the only clash occurs at the level of the number feature, which, by itself, does not alter the thematic coherence of the sentence. In the word salad example, however, for a coherent parse to be built, *fell* or one of the other words must suffer a thematic violation. For example, Fig. 6 shows a case in which *fell* is coerced into playing the role of adverb (this role would work well for a word like *often* or *peacefully*). In the resulting stable structure, the experiencer role of *fell* remains unassigned, rendering the sentence ungrammatical and uninterpretable.[13] Note that,

[11] The DP ➔ D NP treelet is "abstract" (i.e. not lexically-anchored). Abstract treelets encode phrase-structure rules (e.g. S ➔ DP VP, where S = sentence, DP = determiner phrase, and VP = verb phrase).

[12] There are, in fact, multiple stable configurations associated with the sentence *The boy sleep*. For example, the VP and S treelets could inhabit their plural state, agreeing with the verb, and there could be a coerced bond between the DP[+PL] daughter above and the DP[−PL] mother below. We have focused on the stable state shown in Figure 5 because it is the most likely asymptotic state under normal, left-to-right detection of the words, but the other structures may also occur (e.g., due to noise or an anomalous word ordering experience).

[13] We acknowledge that our use of "coercion" for the cases we here call "uninterpretable coercion" takes the meaning of the term nontrivially beyond its current orbit in the literature. Up to now, as far as we know, the term has been restricted to the types of cases we call "interpretable coercion". Moreover, for some researchers, it may be restricted to the subset of interpretable coercion cases which produce grammatical sentences. We use the term in this new, broad way on the grounds that, in our formal system, a structure is formed even in cases of "uninterpretable coercion" and that structure involves forcing one or more words to play a structural role for which they are (very) badly suited. We believe that this approach is ultimately more parsimonious than theories which claim that interpretable coercion and syntactic failure lead to different kinds of mental states because it models both perfect grammaticality and all degrees of ungrammaticality in one representational universe in which all states are reached by the same process.

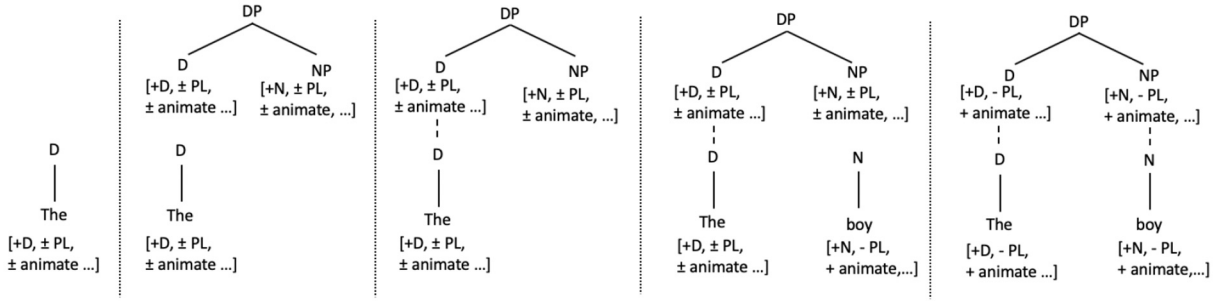

Fig. 3a. Each quadrant (from left to right) represents the successive activation of treelets as new words are perceived for the segment *The boy*. Dotted lines represent attachments that are formed between treelets, while straight lines signal fixed relations within treelets. Feature values are in square brackets.

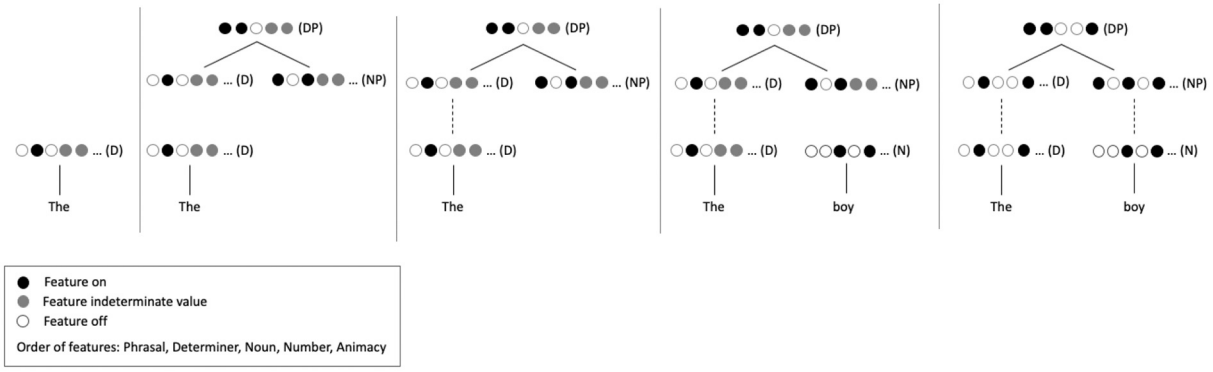

Fig. 3b. Illustration of treelets activation, attachments formations, and feature convergence for the segment *The boy*. This illustration is fully analogous to Fig. 3a, but here treelets' features are represented with vectors. Each dot represents a feature. The order of features is: Phrasal, Determiner, Noun, Number, Animacy. White dots represent the absence of the (marked) feature value (e.g.-D, -PL), black dots represent its presence (e.g. +D, +PL), and gray dots indicate that the value of the feature is underspecified (can be anywhere along the continuum). The syntactic category is also reported in parentheses for convenience.

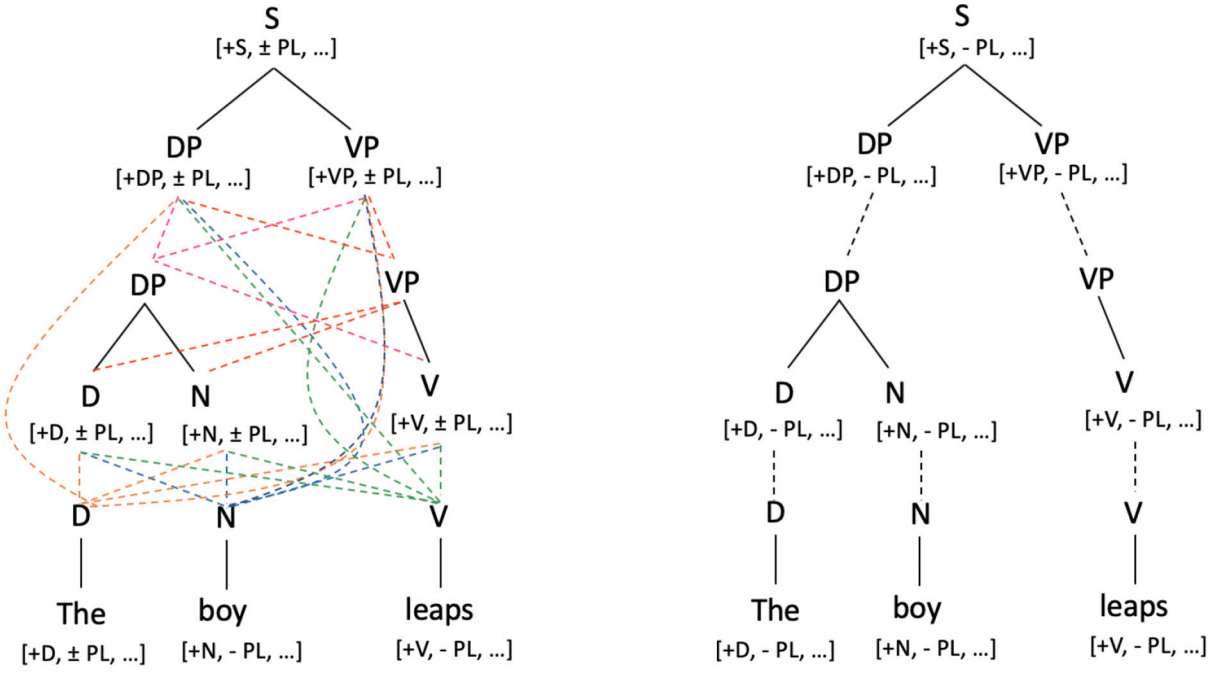

Fig. 4. The left-hand graph shows treelets interactions for the sentence *The boy leaps*. Solid lines indicate fixed links, while dashed lines indicate (nearly) all-to-all interactions during link formation between treelets. Each colour refers to the possible interactions of a given treelet's mother with the other treelets' daughters (e.g. orange lines indicates all possible attachments that the system explores for the treelet *the*), with the proviso that only undirected acyclic graphs (i.e. trees or collections of trees) are entertained. The right-hand plot shows the links that typically stabilize for *The boy leaps* due to optimal feature match. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

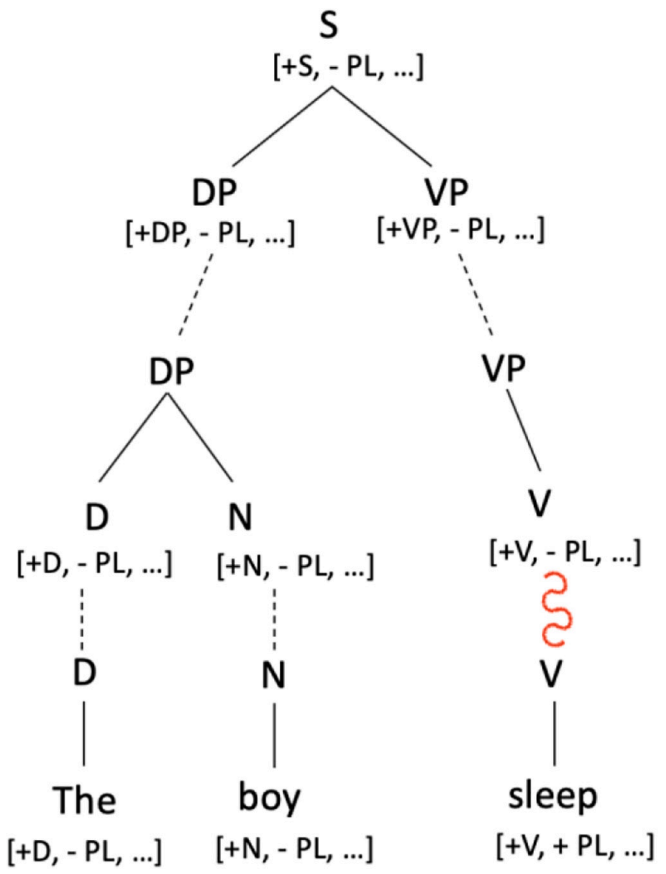

Fig. 5. Illustration of sub-optimal attachment (represented by the red squiggly line) for the ungrammatical sentence **The boy sleep*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

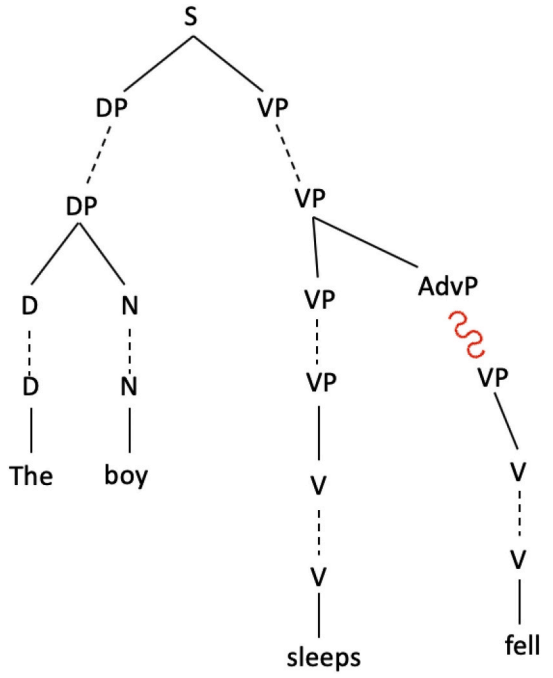

Fig. 6. Illustration of uninterpretable coercion for the sentence "The boy sleeps fell". The red squiggly line indicates a coerced attachment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

although coercion in SOSP-GD involves the formation of bonds between treelets that would be prohibited in a classical model, there is no feature-distortion within treelets – the range of states that a treelet can occupy is defined by the grammar and even in coerced structures, the treelets do not stray outside of their ranges.

4. Islands and SOSP-GD

Like other unification-based theories, SOSP-GD handles long-distance dependencies through slash-propagation (Gazdar, 1982, pp. 167–174). Slash feature bundles[14] (e.g. /NP) encode long-distance dependencies by keeping track of extracted elements. The slash feature bundle is propagated down the tree until a gap position is found. Under standard grammatical models, islands are ungrammatical because the only available gap is inside an island and the grammar does not permit the slash features to propagate across the island boundary. If the gap is not propagated inside the island, the wh-element does not receive its thematic role from the verb, producing an ungrammatical result called *vacuous quantification*. The key insight offered by SOSP-GD is that slash-propagation inside what appears to be an island *can* occur if some treelets in the island-violating sentence are interpretably coerced to produce a structure that allows slash propagation from the filler to the gap. Specifically, it is through the distinction between interpretable and uninterpretable coercion that SOSP-GD captures the strong/weak island distinction: strong islands are strong because they cannot be interpretably coerced, while weak islands are weak because they can be interpretably coerced.

We commented above that the use of slash-propagation to handle long-distance dependencies is motivated by our self-organization premise. This is because self-organization is most easily conceptualized as a property of particle-dynamics: through feedback interactions, small particles organize themselves into larger units. This is easy to make sense of if the only elements in the system are the particles, as in unification grammar models (e.g. Categorical Grammar, HPSG, Construction Grammar, TAG, the Parallel Architecture); it is harder to conceptualize if there are additional structural elements like chains. Nevertheless, for any grammatical framework, if there is a reasonable way to encode parses as points in a metric space, then there is a natural way of building a self-organizing parser (Smith & Tabor, 2018), so our preference here for particle-grammars may be more of a conceptual convenience than a formal requirement. These observations clarify that, in our formulation, the only constraints at work in an implemented model are the featural specifications of the treelets (as in all other particle-based grammar theories) and the self-organizing dynamics. Although we are not focused here on principles of Universal Grammar, if one were to express them in this framework, they would take the form of generalizations about the possible featural specifications and restrictions on the form of the self-organization equations.

[14] Such feature bundles are called "slash features" because they are traditionally tagged with a "slash", e.g. /NP. As with other treelet labels, the label after the slash is a shorthand for a bundle of features which specifies the semantic and syntactic properties of the extracted element. In unification formalisms, the slash bundle is typically implemented as an element in a hierarchically structured attribute-value matrix. In SOSP-GD, it is implemented as a fixed subset of the elements of a vector. This is important because all vectors on all nodes of SOSP treelets must have the same number of dimensions and each dimension must have a fixed syntactic/semantic interpretation so that all the pairwise comparisons that mediate self-organized convergence are meaningful.

4.1. Coercion in islands

4.1.1. Whether islands

SOSP-GD predicts that Whether islands (e.g. *What do you wonder whether the candidate solved?*) undergo interpretable coercion, both with simple and complex wh, but particularly so when the extracted wh is complex. Whether islands undergo interpretable coercion because the featural profiles of their licensors (*wonder* and *whether*) are close (though not identical) to the profiles demanded by slash-propagating treelets. This closeness can be established by comparing the profiles of the island licensors to the profiles of verbs (like *think*) and complementizers (like *that*) that freely participate in wh-extraction structures – cf. *What do you think that the candidate solved?*.[15] In broad syntactic terms, both *wonder* and *think* are mental-process verbs that subcategorize for a propositional complement. Each is followed by a complementizer that introduces the complement. So, although the complementizers have different semantic values, they themselves are in syntactically similar roles. Moreover, considering the construction semantics, a prominent dimension of the meanings of both *wonder whether* and *think that* is the subject's degree of certainty about the embedded proposition. *Wonder* indicates maximal uncertainty about the truth/falsehood of its complement: *John wonders whether the candidate solved the puzzle* entails that John is uncertain about the truth/falsehood of the proposition *the candidate solved the puzzle* (i.e. each scenario might have approximately a 50% chance of occurring). *Think*, in one of its senses, also indicates a degree of uncertainty about the truth of its complement, but there is a bias toward the truth of the proposition. The utterance *John THINKS that the candidate solved the puzzle, but he can't swear to it*, where the capital letters indicate focus, entails that John believes that the proposition *The candidate solved the puzzle* is likely to be true (say, perhaps 70% chance that the candidate solved the puzzle vs. 30% that she did not). This percentage can be shifted even more when *think* is not focused: *John thinks that the candidate solved the puzzle* entails that John is strongly biased toward the truth of the proposition (perhaps more like 90% vs. 10%). Because of their proximal semantic structure, it is possible for *wonder* to appear in the same environment as *think*, and likewise for *whether* to appear in the same environment as *that* without any thematic requirements being unsatisfied. Thus, SOSP-GD expects extraction from Whether islands to produce interpretable coercion.[16] Fig. 7 depicts the interpretably coerced structure the theory assigns to whether islands.

We noted above that self-organization formally implements a kind of analogical structure formation. Specifically, *wonder whether* supports extraction from its complement because *wonder whether* can play an analogous role to *think that*. However, this does not imply that the meaning of a coerced wonder-whether sentence is the same as that of its analogous think-that sentence. Specifically, we suggest that the semantics of the interpretably coerced wonder-whether sentence is more or less the semantics of *For which problem, did John engage in wondering whether or not the candidate solved it?* not the semantics of *Which problem did John think that the candidate solved?*. The reason is that, even though *wonder whether* is being forced to play a role unnatural to it by supporting propagation of slash features, this role-forcing does not disturb the equal-alternative-pondering meaning that *wonder whether* conveys, so the construction ends up having the meaning of object-questioning combined with equal-alternative-pondering.

Next, we ask: why are Whether islands with complex wh more coercible than their simple counterparts? Non-harmonious trees (i.e. trees with harmony less than 1) are generally more unstable than fully harmonious ones. The lesser stability renders non-harmonious trees more vulnerable to noise than fully harmonious ones. Noise can have the effect of knocking the system off the track, diminishing the chances for interpretable coercion to occur. Interpretable coercion, however, can be favored by several factors, the most important of which is semantics. Let us consider the following sentences:

(8) What do you wonder whether the candidate solved _?
(9) Which problem do you wonder whether the candidate solved _?

In (8), when *what* is reached, a slash-bundle is activated. The bundle only contains very general features, such as [+nominal], because *what* has a very generic meaning. When *solved* is encountered, its theme only roughly matches the slash bundle: it is not only specified as [+nominal], but also as [+operable-by-willful-agent], [+abstract], [+hurdle] etc. The featural specification of the two elements eventually converges, since *what*, being semantically underspecified, activates additional features to match the semantics of the theme of *solve*. However, while this change is happening, there is ample time for noise to knock the system into a different state. In fact, due to the relatively low featural specification of *what*, the slash bundle /NP_{what} is relatively close in its featural make-up to an empty slash bundle. Thus, there is a good chance that the noise will simply cause the slash bundle to be dropped, making the system expect a no-gap continuation when it gets to the verb in the Whether clause. Since the featural specification is more fleshed out and fits the verb well in the complex wh case, the model is much less likely to generate a no-gap continuation in that condition.

Fig. 8 illustrates the two options just described for Whether islands with simple wh[17]: the left branch of the tree (in blue) illustrates the parse that results when the system simply loses track of the slash feature bundle and, as a result, the slash feature fails to propagate inside the island domain. This is a case of uninterpretable coercion: the verb *solved* is forced to behave like an intransitive verb – this loss of a theta role results in a very low final harmony because of the large number of feature violations that occur when a thematic element is completely dropped. The right branch (in green) illustrates the case in which the system coerces *wonder* to behave like *think* and *whether* to behave like *that*, allowing the propagation of the slash feature down the tree, inside the island domain. This is a case of interpretable coercion: although it has some compromised grammaticality, the thematic requirements of all elements, including the verbs *wonder* and *solved* are satisfied, so the final harmony is much higher than in the uninterpretable case. It is important to clarify that the resemblance of the island-violating sentence and the corresponding grammatical one is not discovered by the system through a "search" in some sort of mental repertoire containing both the Whether island and the corresponding extraction from a declarative sentence. Instead, the formation of the parse is achieved by the dynamics of treelet interaction: the system forces attachments between treelets to form even in suboptimal conditions; the syntactic and semantic closeness of *wonder*

[15] Our proposal is closely related to that of Kalyan (2012) who argues that similarity to fully grammatical extraction structures is a key factor that determines how acceptable an island-extraction will be. Our account differs subtly from Kalyan's which is an exemplar-based model, while ours is a self-organization model – so rather than claiming that comparison to other exemplars is the mechanism by which a mind arrives at an opinion about the acceptability of a sentence, we claim that treelet self-organization optimizes the constellation of bonds, which is closely related to matching instances to prototypes.

[16] We have just offered several tokens of evidence in support of our claim that featural similarity between the mother nodes of Whether island licensors and relevant daughter nodes of certain slash-propagating treelets is relatively high in Whether island constructions. In making these arguments, we have mentioned syntactic and semantic features that are shared between *wonder* and *think* as well as between *whether* and *that*. In focusing on syntax and semantics, we do not mean to either rule in or rule out the possibility, advocated in a number of island treatments (e.g., Abeillé, Hemforth, Winckel, & Gibson, 2020; Ambridge, 2015; Erteschik-Shir, 1973; Goldberg, 2006) that the constraints on extraction are (also) significantly discourse-based.

[17] These same two options are, in fact, also available for Whether islands with complex wh. In the case of Whether islands with complex wh, however, most of the time, the system discovers the coercion of *wonder whether* into *think that*.

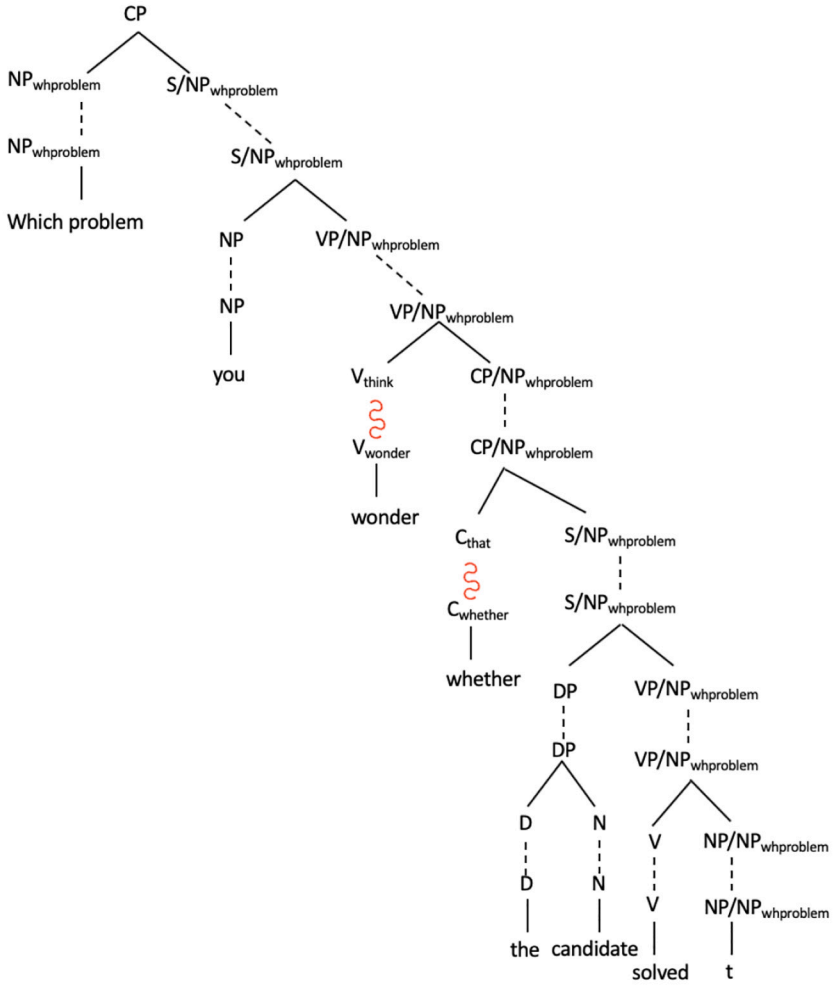

Fig. 7. Simplified tree illustrating interpretable coercion for a Whether island (*Which problem did you wonder whether the candidate solved?*). Do-support has been omitted for simplicity. Red squiggly lines indicate coerced bonds. Subscripts indicate which feature has been transmitted (e.g. $S/NP_{whproblem}$ means that the slash feature bundle for *which problem* has been propagated down inside the S node). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

whether and *think that* has the consequence that sentences that differ only in these elements tend to stabilize on the same syntactic form, but SOSP-GD does not claim that the processing of either sentence involves participation by the other as a computational unit.

4.1.2. Complex NP (CNP) islands

SOSP-GD predicts that CNP islands will also undergo interpretable coercion in some proportion of the cases. The complex VP (e.g. *hear the rumor*) is coerced into a simple VP (e.g. *hear*), as illustrated in Fig. 9. This is a case of interpretable coercion because the meaning of the sentence is preserved: both VPs denote the same hearing event. Moreover, and along the lines discussed for Whether islands, coercion is predicted to be favored in CNP islands with complex wh due to the stronger semantic link that complex wh-elements can establish with the embedded verb, which fosters the generation of a dependency inside the island domain. This prediction, although borne out in the acceptability study by Sprouse and Messick (2015), failed to replicate in subsequent acceptability studies (Villata, Tabor, & Sprouse, 2020a, 2020b). More investigation is needed to understand the status of complex Complex NP islands. We leave this question for future research.

Just as in the Whether island case, the extractability from CNP islands depends on a semantic coincidence: the meaning of *hear the rumor that* is very similar to the meaning of *hear that*. Although the nominal object of *hear* is not mentioned in the latter case, whenever there is an event of <hearing that S>, there must be a corresponding proposition with a nominal present that conveys almost the same meaning. Correspondingly, whenever there is a nominal object (like *the rumor*) referring to the linguistic entity that expresses the relevant propositional content, then the description *hear that* is also true of the situation (if one has *heard the rumor that* X is true, then one has also *heard that* X is true). It is true that the meanings of *hear the rumor that* and *hear that* are not identical – this is why coercion is involved – but because the two forms have isomorphic thematic structures the coercion is interpretable.[18]

A natural question to ask is: does coercion only occur when no grammatical continuation is available? If coercion only occurs when no grammatical continuation is possible, then it would be important to consider an alternative theory in which all the semi-grammatical extractions under consideration here are generated by an extra-grammatical mechanism that kicks in when grammatical processing fails (see, for example, Wagers, Lau, & Phillips, 2009). Indeed, in the case of Whether islands, no licit gap position is possible outside of the island domain. As a result, coercion is the only way to reach an interpretation of the sentence. The case of the CNP island, however, is interesting as it appears that CNP islands allow a grammatical continuation. Consider the sentences below:

[18] Similar semantic arguments apply to all the other cases investigated by Villata, Tabor, & Sprouse, 2020a, 2020b: {hear/repeat/believe} {the rumor/report/claim/news}. While those authors did not investigate *make the claim*, it is notable that *make the claim*, which Ross felt provided no barrier at all to extraction, has a very similar meaning to the verb *claim*. Our claim is that because the words in the Complex NP extractions can participate in a fairly well-formed tree, the parser is likely to build that tree on a significant portion of trials.

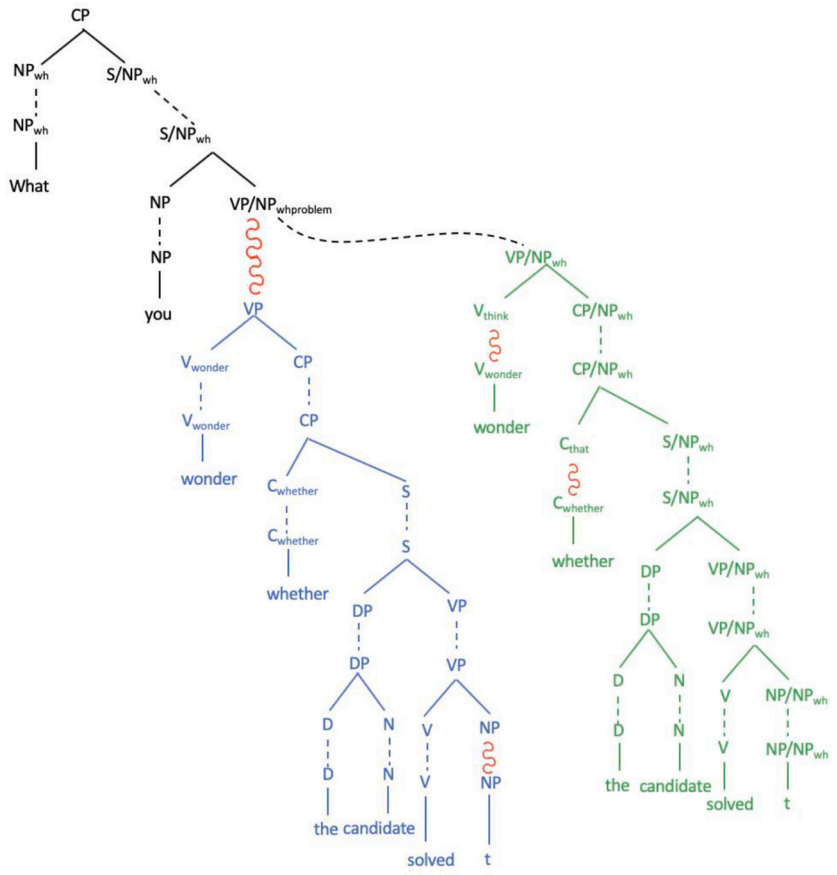

Fig. 8. Simplified tree for Whether island with simple wh (*what*). The two options (uninterpretable coercion vs. interpretable coercion) are illustrated with different colors: uninterpretable coercion is in blue, while interpretable coercion is in green. Subscripts indicate which feature has been transmitted (e.g. S/NP$_{what}$ means that the slash feature for *what* has been propagated to the S node). Squiggly lines indicate loci of coercion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(10) ? Which person did you hear the rumor that the cake delighted _?
(11) Which person did you hear the cake delighted _?
(12) Which person did you hear the rumor that the cake delighted Mary surprised _?

Sentence (10) is a Complex NP island violation, (11) is the grammatical extraction of a noun phrase from the sentential complement of *hear* with *that* deleted at the start of the sentential complement, and (12), building on the possibility indicated by (11), illustrates a possible grammatical continuation for (10) (see Fig. 10). The question is, which structure the parser is going to opt for upon processing the words in (10) up to *delighted*. Does it opt for interpretably coercing (10) into a non-island, thus licensing the gap shown in (10) at the cost of putting the parser in a semi-grammatical state? Or does it opt for expecting a downstream gap after the island domain (12), which would lead to a fully grammatical parse? The data of Villata, Tabor, & Sprouse (2020a, 2020b) provide evidence that English speakers opt for (10), with gap as shown, over any other option (including (12)) in around 58% of the cases when the extracted wh is complex. These data thus provide evidence that the choice to adopt a coerced interpretation is not determined by whether a grammatical option is available. In some cases (e.g. Whether islands) the parser opts for coercion when there is no plausible grammatical continuation available, but in other cases (e.g. CNP islands), it opts for coercion even though a plausible grammatical continuation is available. In the General Discussion, we explain under what conditions SOSP-GD forms a coerced structure even though a perfectly well-formed one is available.

The claim that the parser might opt for coercion even when a grammatical parse is available is intriguing as it suggests that semi-grammaticality can be preferred over full grammaticality.[19] This claim is in line with several works from the literature reporting that the parser can entertain representations that are not licensed by the grammar even when a grammatical option is available. This is the case for local coherence effects, i.e. locally coherent but globally ungrammatical parses (e.g. Konieczny, 2005; Paape & Vasishth, 2016; Tabor & Hutchins, 2004; Villata & Lorusso, 2020), and for misanalysis persistence in garden path sentences, i.e. the perseverance of the initial (wrong) analysis even after reanalysis has occurred (e.g. Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira & Bailey, 2004; Ferreira & Henderson, 1991).

4.1.3. *Because-Adjunct islands*

Unlike Whether and Complex NP islands, *because*-Adjunct islands

[19] It is important to notice that results from the maze task from Villata, Tabor, & Sprouse (2020a, 2020b) do not allow us to establish whether participants who opted for the filled-gap continuation over the gap continuation did so because they expected a downstream gap in CNP islands. However, we believe that this is unlikely. First, none of the experimental sentences for CNP islands contained a downstream gap, which means that when participants went for filling the gap inside the island, the sentence turned out to be a case of vacuous quantification. Second, downstream gaps are not likely with inanimate extractees, since inanimate NPs are rarely subjects, and all the extractees in Villata, Tabor, & Sprouse (2020a, 2020b) were inanimate.

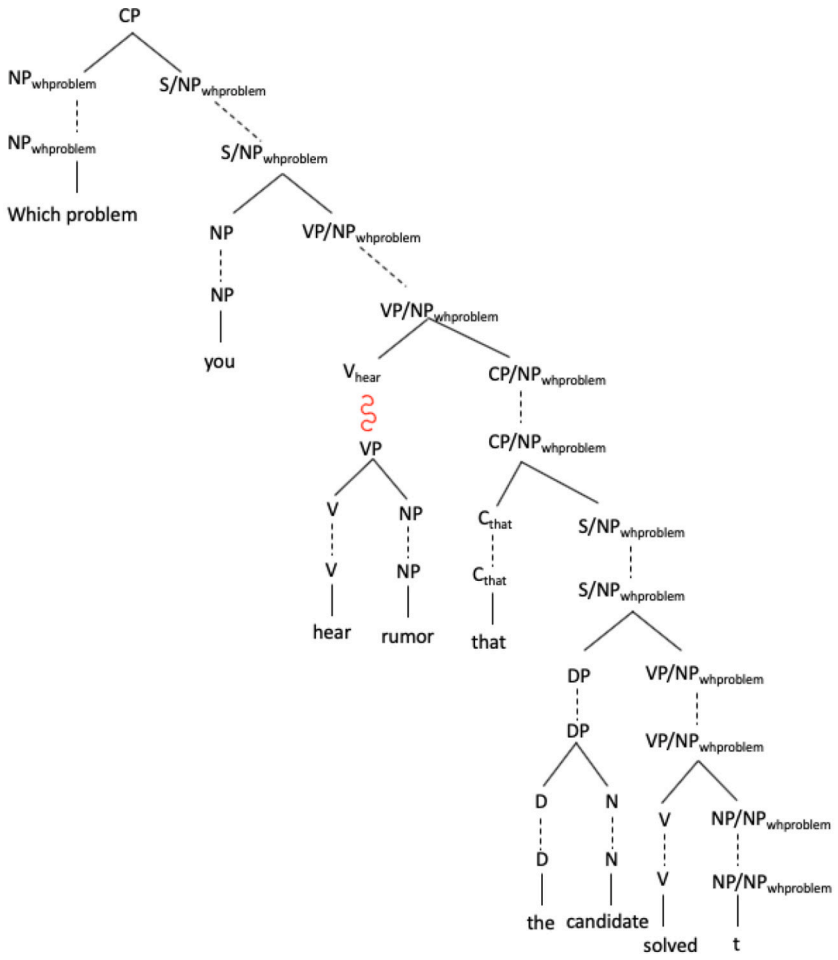

Fig. 9. Simplified tree illustrating the coerced tree for CNP islands with complex wh (*Which problem did you hear the rumor that the candidate solved?*). The red squiggly line indicates the coercion of *hear the rumor* into *hear* which enables the parser to propagate the slash feature down the tree. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cannot undergo interpretable coercion – they can only undergo uninterpretable coercion. Fig. 11 illustrates one of the uninterpretable coercion possibilities for parsing *because*-Adjunct extractions: intransitive *sneeze* has been coerced into a transitive verb like *bring* and mandatorily transitive *adores* has been coerced to an intransitive. The first coercion is uninterpretable because *sneeze* is assigned a direct argument which it does not license. The second coercion is also uninterpretable because *adores* is a strictly transitive verb and the internal argument cannot be omitted. In Section 4.3 below, we report a computational investigation which strengthens this claim: the algorithm systematically searches for possible coercions and only comes up with uninterpretable parses for *because*-Adjunct extractions.

4.2. Summary of SOSP-GD and empirical findings on islands

To recapitulate, judgment studies have shown that weak islands with complex wh are rated higher in acceptability than their simple counterparts. This has been attested for Whether islands in English (e.g. Sprouse, Caponigro, Greco, & Cecchetto, 2016; Sprouse & Messick, 2015), wh-islands (e.g. Atkinson, Aaron, Kyle, & Omaki, 2015 in English; Villata, Rizzi, & Franck 2016 in French), and it has also been reported for Complex NP islands (Sprouse and Messick, 2015), although Villata, Tabor, & Sprouse (2020a, 2020b) failed to replicate this effect. No increase in acceptability ratings as a function of the type of the extracted wh (simple vs. complex) is observed for *because*-Adjunct islands. Turning to the maze task, Villata, Tabor, & Sprouse (2020a, 2020b) find that both types of weak islands (Whether and CNP islands) with complex wh exhibit intermediate probabilities of gap continuation – lower than grammatical controls but higher than *because*-Adjunct islands. Moreover, the same properties are observed for these islands with simple wh – lower probabilities of gap continuation than grammatical controls and higher than *because*-Adjunct islands. Additionally, conditions with simple wh show lower rates of gap-completion in all three island types, but the separation is larger in Whether and CNP islands than Adjunct islands.

SOSP-GD captures most of these patterns by making the following claims:

1. Whether islands undergo interpretable coercion – *wonder whether* is interpretably coerced into an approximation of *think that* (passing the slash feature but with the semantics of a wonder-complement). When the extracted wh is complex, coercion is predicted to happen even more often because of the support provided by the richer semantic content.
2. Complex NP islands also undergo interpretable coercion – *hear the rumor* is coerced into an approximation of *hear*. Again, complex CNP islands are predicted to undergo coercion more often that their simple counterparts.
3. *Because*-Adjunct islands do not undergo interpretable coercion either with complex or with simple wh (only uninterpretable coercion is available for *because*-Adjunct islands).

As such, SOSP-GD appears to provide a largely accurate, independently motivated account of the island structures under consideration

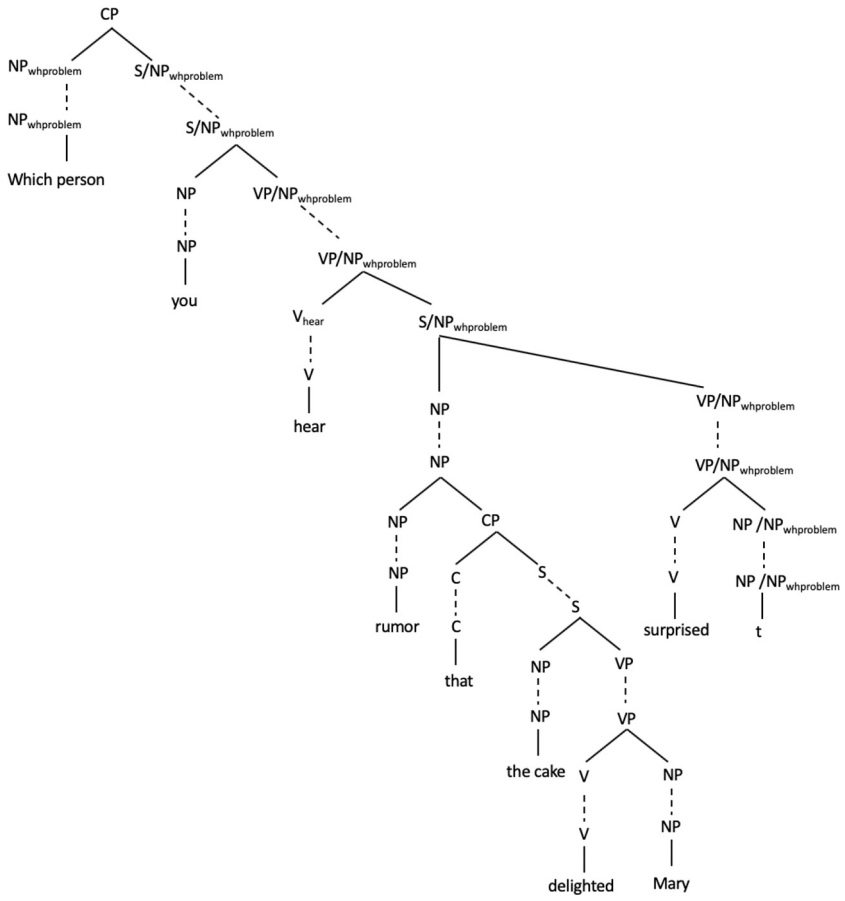

Fig. 10. Simplified tree illustrating a grammatical continuation for a sentence containing a Complex NP island (*Which person did you hear the rumor that the cake delighted Mary surprised?*).

here, including their gradient behavior. The SOSP mechanism for semi-grammatical extraction from weak islands is interpretable, but still mildly penalized, coercion.

4.3. Computational investigation of SOSP-GD coercion claims

[The code for the computational investigations reported here is available at https://github.com/WhitneyTabor].

4.3.1. A more detailed introduction to the formal model

Here, we work with a version of the SOSP model called "SOSP-GD". "GD" stands for "Grammar Derived", indicating that the model formulation is based on a (probabilistic) grammar model. (13) shows the procedure for deriving an SOSP-GD model from a probabilistic context free grammar (PCFG).

(13) Steps for deriving an SOSP-GD model from a PCFG:

PCFG ➔ Extended CFG ➔ Tree Harmony Values + Tree Locations ➔ Attractors ➔ Dynamical Processing Model (= SOSP-GD).

A PCFG, parameterized on the basis of a language sample, serves as the foundation of the SOSP-GD model. We summarize its formulation in *Model Formulation* and give an example of how it parses a sentence in *Example of a Parse* below.

4.3.2. Model formulation

In describing the formulation, we make reference to the simple-wh forms of the example shown in Fig. 7 (i.e. *What did you wonder whether the candidate solved?)*. The starting point is a PCFG, but we first focus on the CFG part (i.e. the grammar rules, ignoring their probabilities). As an illustration, consider the (P)CFG in Table 1.

From the CFG, we generate bonding-rules, which specify how a daughter of one treelet (i.e. rule) connects to the mother of another. Each bonding rule is associated with a local harmony value. The simplest and most harmonious form of bonding rule has the form X ➔ X, which bonds a daughter to a mother with the same label. Such rules are assigned the local harmony value, 1.[20] An example (14) of such a bonding rule that can be derived from the grammar G_1 in Table 1:

(14) Wh[bare] ~> Wh[bare] (Harmony = 1)

Less harmonious bonding rules involve a daughter above connecting to a mother below with a different label (15):

(15) V[sentcomp] ~> V[wonder] (Harmony = 0.8)

In this case, we have indicated a harmony of 0.8. This number reflects two properties of *wonder*: it takes a sentient subject and a sentence complement, so its argument geometry matches that of other sentence-complement verbs, but it is not a canonical sentence complement verb like *think* or *hear* – for example, its complement must be introduced by *whether* and cannot be introduced by *that*. Because it is a good fit, but not

[20] In the full model, these rules may receive harmony values that are slightly less than 1 – such values generate predictions about subtle variation in acceptability values that is attested even among grammatical sentences (e.g., Villata, Rizzi, & Franck 2016), but for present discussion, we will ignore this nuance.

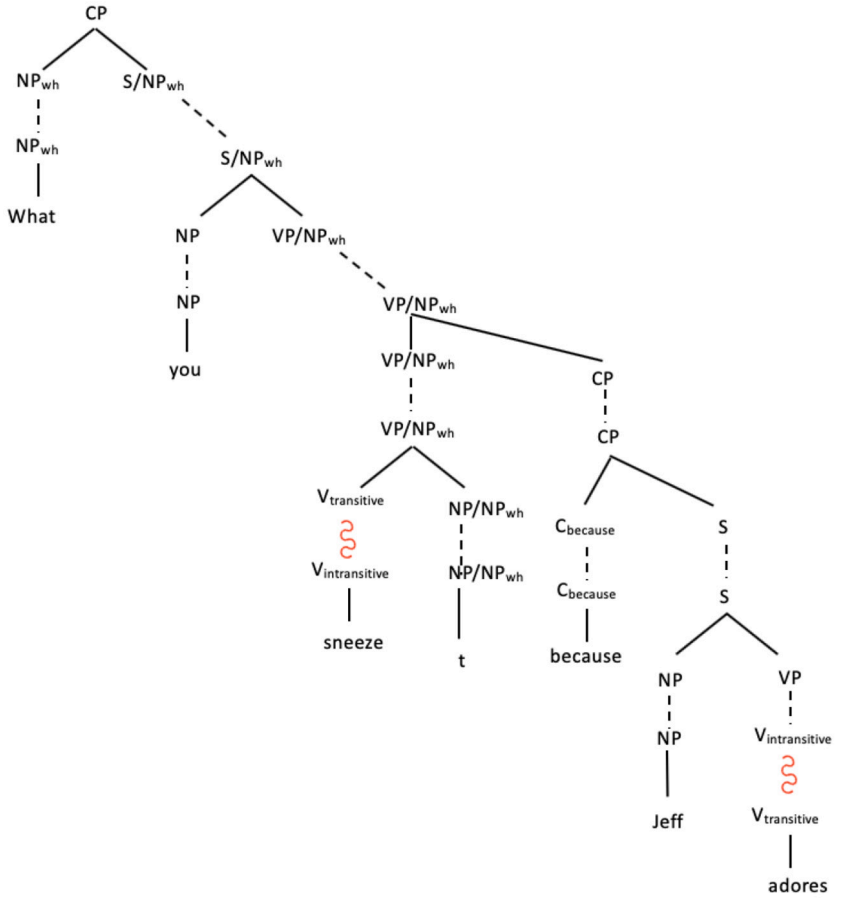

Fig. 11. Simplified tree for a *because*-Adjunct island (e.g. *What did you sneeze because Jeff adores?*) where intransitive *sneeze* is coerced to transitive and transitive *adores* is coerced to intransitive, illustrating a case of uninterpretable coercion.

Table 1
Grammar G_1, a slightly simplified fragment of the PCFG we used for modeling the cases in focus here (see Appendix I for the full grammar).

0.67 Root: S
0.33 Root: WhQ
1.00 WhQ ➔ Wh[bare] S/What
1.00 S: NP VP
0.67 S/What ➔ NP VP/What
0.33 S/What ➔ VP
0.48 VP ➔ V[intrans]
0.24 VP ➔ V[trans] NP
0.16 VP ➔ V[sentcomp] CP[that]
0.12 VP ➔ V[wonder] CP[whether]
0.67 VP[trans] ➔ V[trans] NP
0.33 VP/What ➔ V[trans]
…

a perfect fit, we have selected a harmony value less than 1, but relatively close to 1.[21]

Another relevant case is the bonding rule (16):

(16) Vtrans ~ > Vintr (Harmony = 0.2)

This bonding rule plays a role in the parse of the *because*-Adjunct example shown in Fig. 11. Here, the harmony value, 0.2, is greater than 0, reflecting the fact that all verbs share some structure (e.g. they all take a subject) but the argument geometries of the two types are topologically distinct (intransitives have one argument, transitives have two), so the harmony is far from 1. All bond harmonies lie in the interval, [0,1].

Adding such bonding rules to the original PCFG produces the "Extended CFG" in (13). We use the Extended CFG to generate trees and partial trees. The trees are parses of full sentences (some grammatical, some ungrammatical). A partial tree is any minimal contiguous tree structure by which the grammar generates an initial substring of a full sentence. We define the harmony of a tree or partial tree to be the product of the harmonies of its bonds. Thus, grammatical parses have a harmony of 1 and ungrammatical parses have harmony values strictly less than 1.

The next step is to use the tree structures to specify a dynamical system (for a vivid introduction to dynamical systems theory, see Abraham & Shaw, 2000). A dynamical system has a state space – the set of all possible states that the system can occupy. In the current model, the state space models the set of all mental states involved in parsing sentences. We assume that the state space is a continuum with a metric on it, so we can speak of a range of degrees of states being closer together or farther apart from one another. In many cases, dynamical systems have attractors in their state spaces. An attractor is a set in the space such that, when the system is placed close to the set (i.e. within some positive radius) and is allowed to evolve autonomously, its state converges to the set. A particularly simple kind of attractor is a fixed-point attractor – a single point to which the system converges from immediately surrounding points. In SOSP-GD, each possible tree structure described

[21] In the full formal model, these harmony values are not estimated by eyeballing, but are derived by a distribution comparison technique applied to treelets in a corpus generated by the original PCFG – this procedure is related to methods of using neural networks to model word prediction (e.g., Devlin, Chang, Lee, & Toutanova, 2019; Elman, 1990, 1991; Lau, Clark, & Lappin, 2017) and thus provides a theoretical bridge between such neural learning models and the classical theory of grammar.

above (including the partial tree structures that correspond to partial parses of sentences, ungrammatical and grammatical) is a fixed-point attractor. Since the cardinality of the tree structures just described is a countable infinity, different such attractors are separated in the continuum. The points in the continuum that lie between attractors can be thought of as corresponding to blends of multiple grammatical structures (see Cho, Goldrick, & Smolensky, 2017). These in-between points serve to model the transient states that the mind is hypothesized to be in when it is choosing which parse to adopt in response to a sequence of words presented to it (or choosing among various possible structures to employ when it is producing sentences).

Here, we will not precisely specify the geometry of the state space or the dynamical equations, because the focus of this paper is not on the details of moment-to-moment processing. Nevertheless, one can think of the model's trajectories in the state space as modeling the through-time interactions among treelets described in the preceding sections of the paper (see Smith & Tabor, 2018; Villata, Sprouse, & Tabor, 2019). A key part of this dynamic is a constellation of attractive fixed points corresponding to possible (partial and full) parses of a given sentence of interest. These fixed points are positioned so that, after the model has received a sequence of words, $w_1 \ldots w_n$, it is at a location intermediate between attractors corresponding to possible parses of $w_1 \ldots w_{n+1}$, where w_{n+1} can be any word in the vocabulary (i.e. the previously seen words plus one more word of unknown identity). Let [v] be one possible parse of $w_1 \ldots w_{n+1}$ (including a particular choice of vocabulary item for w_{n+1}) and let v be the attractor corresponding to [v]. At the point just before receiving the new word, the distance of the state from v is roughly inversely proportional to the probability, under the extended CFG, given the structural biases induced by words $w_1 \ldots w_n$, that the parse will be [v]. In other words, the state's position is biased to reflect the likelihood of observing each next structural state, given the current structural state under the PCFG. Given that it always has low-level noise in its activation values, if we run the model in generation mode, letting it gravitate to an attractor which indicates each next structure and the corresponding next word until it makes a complete sentence, the model will generate a distribution over structures that largely conforms to the distribution specified by the PCFG.[22]

4.3.3. Example of a Parse

The SOSP-GD model (last step of diagram in (13)) is simply a large ensemble of interconnected neuron-like units whose activations range from 0 to 1. At the beginning of comprehending any sentence that is experienced out of the blue, the model starts in an initial activation state that is the average position of all the attractors associated with all complete sentences that can be generated by the PCFG (each position is weighted in the average by the sentence probability). Since, in a natural language grammar, there are only a few ways to start sentences and many ways to finish them, the average tends to stack up activation on nodes encoding possibilities for the first word or the first few words. For comprehending a sentence like:

(17) What (do) you wonder whether the candidate solved?" (see Fig. 7)

the model first receives the word *What*. This results in the appearance in the dynamical system of an attractor corresponding to various possible parses of *what* as the first word of the sentence. What are these possible parses? In fact, under the Extended CFG, there is only one reasonable one (in examples (18)–(20) we show high harmony bonds in bold and low harmony bonds in bold and italic):

(18) [Root [**WhQ ~> WhQ** [WhQ [**Wh(bare) ~> Wh(bare)** [Wh (bare) what]] [Smax/What]]]]

(Harmony = 1)

Some other possible, but not very reasonable parses are (19, 20):

(19) [Root [***S* ~ > S** [S [NP [***NP* ~ > *Wh(bare)*** [Wh(bare) what]]] [VP]]]].

(Harmony = near zero)

(20) [Root [***S* ~ > *VP*** [VP [Vtrans [***Vtrans* ~ > *Wh(bare)*** [Wh(bare) what]]] [NP]]]].

(Harmony = very near zero)

In fact, under the Extended CFG, there are many other parses of sentence-initial *What* besides the coherent, grammatical parse, (18), but they all have very low harmony like (19, 20). In a dynamical system, the set of points around an attractor from which the system converges to the attractor is called the *basin* of the attractor. Basins can be wide or narrow. In SOSP-GD, the width of the attractor basin associated with a particular parse is proportional to the parse harmony. Recall that there is continuous small-magnitude noise in the activation values of SOSP-GD. Consequently, when the system starts at a point equidistant from multiple attractors, the probability of converging to a given attractor increases with its basin width. This means, when there are multiple attractors in its vicinity, SOSP-GD is more likely to converge to high harmony attractors than low harmony ones. In the current example, the system has a very high likelihood of parsing sentence-initial *What* with the structure (18) and a low likelihood of any other choice. In general, when one (or more) grammatical parses is accessible to it, SOSP-GD is very likely to converge to it (or one of them) – in other words, it expects grammatical continuations. However, there is a small likelihood that, even when processing a grammatical sequence, SOSP-GD will converge to an ungrammatical parse. Relatedly, if SOSP-GD is presented with an ungrammatical sentence (a sentence for which confusion is unavoidable), then the dominating basin at each word will be that of the highest harmony structure associated with the words seen thus far, but this harmony will be less than 1 – i.e. the model "optimizes" as discussed in the Introduction.

Returning to example (17), once the system has gravitated to whichever attractor it chooses after processing *What*, it receives the next word. In this case, for simplicity, Grammar G_1 ignores English do-support so the next word is *you*. As soon as *you* appears, the attractors corresponding to partial parses of *What* disappear and are replaced by attractors corresponding to parses of *What (do) you*. At this point, the state of the system is still the state it was in after converging to the attractor it chose after *What*. This state is at a positive distance from all the attractors associated with *What (do) you* because each previous parse state (for *What*) is an encoding of the average future after *What* (given the particular interpretation of each parse) and there are many futures. The model then gravitates (again with noise in the activations) to an attractor associated with *What (do) you*, reflecting its detection of the word *you* and its assignment of a particular interpretation to *you*. The process is repeated for each word perceived. In this fashion, the model proceeds through the sentence and arrives at a final attractor corresponding to its chosen interpretation of the whole sentence.[23]

To summarize key points of the account so far, it is through the mechanism of attractor convergence, including attractors for

[22] This is similar to the behavior of a standard, grammar-based surprisal model (e.g., Hale, 2001; Levy, 2008) except that this model does not place any probability on ungrammatical parses.

[23] The convergence times for each word generate predictions about processing times in tasks like self-paced reading (Just, Carpenter & Woolley, 1992) and eye-tracking during reading (Rayner, 1998).

ungrammatical as well as grammatical states that SOSP-GD has a way of interpreting both ungrammatical and grammatical input. Note that, when the full range of possible bonding rules is considered, SOSP-GD has at least one attractor for every finite string that can be formed from its vocabulary. In this sense, it is a general model of both grammatical and ungrammatical processing.

4.3.4. Testing SOSP-GD on the strong/weak distinction

For the current project, as a first step toward investigating the plausibility of the SOSP-GD framework, we asked the following question: What are the harmonies of the attractors associated with each of the full island types of the experiments under discussion? If the model is to produce the main empirical results we mentioned corresponding to the strong-weak distinction, the model's best candidate for a *whether*-extraction should be the structure in Fig. 7, and for Complex NP extractions, the structure in Fig. 9. However, for *because*-Adjunct islands, the model should find only very low harmony structures. We focus, in this section, on modeling just the processing intermediacy properties – viz. intermediacy for Whether and CNP islands, non-intermediacy for *because*-Adjunct islands – that we described in Section 4.1 above. Although our model performs well in these cases, we are not as sure that our model can handle the D-linking pattern we observed. We make further comments on the D-linking case in the General Discussion.

To test this hypothesis, we ran a standard parser with three different grammars that progressively more accurately approximated the full SOSP-GD system. We examined grammatical control sentences and the three extraction types of interest here. First, we used just the Grammar G_1 rules. In this case, fully grammatical parses were found for grammatical control sentences and no parses were found for any of the critical extraction sentences. This makes sense because all three of the critical extraction sentences are ungrammatical. However, it fails to model the experimentally observed differences between the ungrammatical sentences.

Then we introduced three harmony-reducing bonding rules: V[sentcomp] ~ > V[wonder], C[that] ~ > C[whether], V[sentcomp] ~ > VP. These bonding rules implement the relatively high-harmony but nevertheless coerced semantic correspondences we argued for above. With only these bonding possibilities implemented, the parser found just the structure shown in Fig. 7 for the *whether*-extraction sentence, it found just the structure shown in Fig. 9 for the Complex NP extraction, and it found no structures for *because*-Adjunct islands. This is closer to what we observe in the data for these sentences, though the result implies that, if only these limited coercion possibilities were considered, SOSP-GD would reinterpret *because*-Adjuncts as grammatical or nearly grammatical sentences. This is because, in the absence of a proximal attractor for a low-harmony parse, the model will optimize by converging on a proximal higher harmony parse. In this case, we have only provided relatively high harmony bonding rules, with the result that the most proximal attractor for the *because*-Adjunct extraction corresponds to a misperception of the sentence as an entirely different and more grammatical sentence. This outcome is not consistent with the experimental results.

As a third test, we included more low-coherence bonding rules. This makes for an Extended PCFG inventory which much more closely approximates the set of tree structures underlying the full SOSP-GD dynamics. We introduced all coercions involving substitution of one verb for another (including, for example V[intrans] ~ > V[trans] and V[trans] ~ > V[intrans] – see Appendix I for the full list). Each of these cases generates an uninterpretable coercion because verb arguments are either lost, gained, or substituted. We also introduced the "slash-dropping rules", S[Max]/WhichN ~ > S[Max], S/WhichN ~ > S, CP/WhichN ~ > CP, and VP/WhichN ~ > VP. These implement events in which the processor forgets the properties of extracted element that was perceived at the beginning of a sentence while it is processing the rest of the sentence. With all these low-coherence bonding rules included, we are approximating more closely the full "Extended CFG" referred to in (13). Under these conditions, the parser found an additional 20 parses for Whether island extractions, an additional 25 parses for Complex NP island extractions, and it found 21 parses for *because*-Adjunct extractions. All 66 (= 20 + 25 + 21) of these additional parses were uninterpretable. Most of them involved dropping slash features or coercing transitive verbs into intransitives or vice versa. Of course, this more fully extended model also finds the grammatical parses for grammatical sentences and the two relatively high harmony parses for Whether and Complex NP islands mentioned above. Recall that each one of the parses produced here corresponds to an attractor in SOSP-GD. As we noted in Section 4.2.1, SOSP-GD tends to stabilize on the highest harmony attractors available for a given sentence. Thus, the findings just reported indicate that an SOSP-GD model with a full complement of bonding rules will tend to stabilize on high-harmony attractors for grammatical sentences, on relatively high harmony attractors for whether- and CNP-extractions, and on low harmony attractors for *because*-Adjunct extractions. The full model thus appears to align well with the parsing intermediacy results in the experimental data.

Two of the low-harmony parses found for the *because*-Adjunct extractions (Fig. 11 and example (21)) lead to processing predictions that can be tested.

(21) CP [NP Which problem] [Smax/Wh [S/Wh [NP you] [VP/Wh [**Vsc ~ > V_{intr}** sleep] [CP/Wh [**C_{that} ~ > $C_{because}$** because] [S [NP the candidate] [VP/Wh [V_{tran} solved]]]]]]]].

As a premise, we make the assumption that bond-coercion is detectable in online measurements – at the point of making the coercion, participants should slow down and possibly show other signs of processing difficulty (Smith & Tabor, 2018; Smith, Franck, & Tabor, 2021).

Regarding Fig. 11, we hypothesize that the coercion of *sleep* into a sentence-complement verb is quite unharmonious (because the meaning of *sleep* has no place for a propositional argument) but the coercion of *sleep* into a transitive verb, though generally ungrammatical (**sleep the house*, **sleep my arm*), is nevertheless on the border of grammatical for some combinations (*? sleep the baby*, *? sleep a project*). In this case, at the point of processing *because*, the model, detecting no better choice, will tend to coerce intransitive *sleep* into a transitive – see the first coerced (shown as red squiggle) bond in Fig. 11. Having dispensed with the slash feature, the model will proceed into the adjunct clause and eventually encounter the mandatorily transitive verb *solved* without an object. At this point, it has no choice but to coerce again (this time making *solved* function as an intransitive – see the second coerced bond in Fig. 11). Thus, the model eventually stabilizes on the parse shown in Fig. 11. In this case, then, the model predicts that an online experiment should reveal separate loci of difficulty at both *sleep* and *solved*.

In (21), by contrast, the *because*-Adjunct structure has been coerced to have a verb-complement syntax (like that of the grammatical control sentences and mildly ungrammatical Whether island extractions). Referring to (22) below, consider a situation in which it is known from the start that Jeff complains for precisely one of two reasons: either because Lorna plays Rachmaninoff or because Lorna plays Schönberg. In this case, (22) seems to be fairly interpretable, even though its quality is not great.

(22) Which composer did Jeff complain because Lorna played this afternoon?

Our suggestion is that this happens because the context makes the semantics of the because-clause more like that of a complement, and less

like that of an adjunct. Among other things, in the scenario described, the salience of causality is reduced, making the contribution of *because* to the sentence less contentful and more formal and therefore more like the role of *that* in marking a sentential complement.[24] In this circumstance, we expect the processing to favor structure (21) over the structure in Fig. 11. While (21) is expected to produce some strain around the region of the words *sleep because*, it is not expected to produce strain following the sentence-final verb. Overall, then, the SOSP-GD framework identifies a testable interaction between discourse interpretation bias and the pattern of local processing difficulty in adjunct-island extractions. We leave the testing of this prediction to future research.

5. General discussion

5.1. Summary of the case

We successfully parse ungrammatical sentences all the time. Yet, some ungrammatical sentences are easier to parse than others. Islands are possibly one of the most prototypical and challenging examples of gradient ungrammatical processing. Although the strong/weak island distinction has been traditionally accounted for in purely categorical terms, our understanding of this distinction has become more nuanced recently: several studies employing formal methodologies of acceptability judgment gathering have revealed intermediate acceptability patterns for weak islands with complex wh-phrases, rather than full acceptability (e.g. Atkinson, Aaron, Kyle, & Omaki, 2015; Sprouse, Caponigro, Greco, & Cecchetto, 2016; Villata, Rizzi, & Franck 2016; Villata, Tabor, & Sprouse, 2020a, 2020b). Moreover, a recent set of experiments employing the maze task indicates that participants are willing to establish a dependency inside of weak islands (and particularly so in weak islands with complex wh-phrases), while this tendency is virtually nonexistent for strong islands (Villata, Tabor, & Sprouse, 2020a, 2020b). These observations require a theory that addresses both the strong/weak island distinction and its pattern of gradience.

Traditionally, gradient effects (both in grammatical and in ungrammatical sentences) have been accounted for in terms of extra-grammatical mechanisms that intervene after the syntactic module generates its output. In the case of gradient effects in grammatical sentences, it is traditionally assumed that the grammar generates a fully well-formed sentence the acceptability of which is diminished by some extra-grammatical factor. The case of gradient effects in ungrammatical sentences which are interpreted, however, poses a challenge to such accounts: if the grammar is defined as a system of rules that generates all and only the grammatical sentences of a given language, ungrammatical sentences cannot be generated by the grammar. The generation of a meaningful ungrammatical sentence must thus be the result of some extra-grammatical factor that kicks in once the syntactic derivation of the sentence has failed. On one such approach, such a mechanism would have to cobble a structure together, which the semantic module could then extract meaning from. Alternatively, there could be a special semantic interpreter which is able to create meaning from word-sequences without making use of syntactic compositional structure. However, to our knowledge, no specific extra-grammatical mechanism along either of these lines has ever been fleshed out, hindering the explanatory power of this approach.

In this paper, we have presented a new theory that captures gradient effects in a grammar-internal fashion. The self-organized sentence processing grammar-derived (SOSP-GD) model we have described shares with traditional accounts the assumption that the grammar is a system of rules: treelets are pieces of syntactic information that encode phrase structure rules. However, although rule-based, this system is more flexible than traditional ones. In particular, sentential elements can be coerced under specific circumstances to play syntactic roles that do not fully fit them. For instance, a slash-feature blocking element like *whether* can be coerced to play the role of a slash-feature passing element like *that* under certain circumstances. The greater flexibility ensured by coercion is what enables structure formation even under suboptimal circumstances without the need to resort to extra-grammatical mechanisms. We distinguished two types of coercion: uninterpretable and interpretable coercion. The former is a "brutal" form of coercion: the system forces the structure to form, but the theta criterion is violated. Any string of words can be uninterpretably coerced. Interpretable coercion, though it also creates some tension in the system, is less jarring: the system forms a thematically coherent tree (i.e. the theta criterion is met), despite feature mismatch on some nodes.

The self-organizing treelet mechanism has the property that interpretable coercion is often associated with structural analogy: there is a grammatical form in which treelets come together in a particular constellation; the ungrammatical form uses many of the same treelets, but there are some treelets that have to be coerced (i.e. that experience feature clash at some bond sites). If this coercion can be achieved without violating thematic constraints – the interpretable coercion case – then speakers may notice an analogy between the coerced structure and the grammatical structure. On the other hand, if the self-organization produces a structure in which thematic constraints are violated, then speakers tend not to detect any analogy. Under the self-organization account, a natural explanation for this difference in perception is that an interpretably coerced structure is mentally proximal to its grammatical analog: the mind rests in a stable state corresponding to the abstract structure underlying both the coerced sentence and the analog, and can switch between the two by making small adjustments in a few pieces of the structure without disturbing its broad form. By contrast, in uninterpretable coercion, some of the bonds are so bad that the mental state is in effect, not a structure, but a collection of fragments. In this case, even if an analogy is explicitly suggested, it is very hard to switch between the two relevant mental states because the roles played by corresponding parts are not similar. This seems especially so in cases like the one we mentioned in the introduction, *the boy sleeps fell*. Although there is an analogy to *the boy sleeps peacefully*, it is not an easy analogy to grasp because *fell* in the first condition cannot effectively play the role that *peacefully* does in the second. For *because*-Adjunct islands like (4c), we suggested in Section 4.1 that it is possibly slightly easier, than in *the boy sleeps fell*, to detect an analogy provided lexical choice and contextual circumstances support it. Nevertheless, it is much harder to detect an analogy in *because*-Adjunct extractions than in Whether extractions and CNP extractions. We suggested that it is this difference which gives rise to the empirically observed strong/weak distinction.

5.2. Relation to other treatments of semi-grammaticality

Classical unification grammar builds trees by specifying constraints at nodes and allowing two nodes to unify (in SOSP-GD terms, to "bond") if the features of the daughter node above match the features of the mother node below. In relaxation approaches to ungrammaticality (e.g. Douglas & Dale, 1992; Fouvry, 2003; Vogel & Cooper, 1995), nodes can unify even if their features do not perfectly match. These methods massively generate ungrammatical structures by allowing the construction of structures in which features clash. Some of them offer accounts of how to triage among the many semi-grammatical options by defining degree of well-formedness based on the amount of clash and they posit that the system chooses greater well-formedness over lesser well-formedness whenever possible. Here, we have made a related claim, but with two important differences: (1) we let the dynamics of the interacting treelets perform the triage among the structures; (2) for a given language, the inventory of ungrammatical types that our system

[24] This explanation seems aligned with Goldberg (2006)'s backgrounding account, if we assume that the subordinator *because* itself can be foregrounded (21) or backgrounded (22), and that its complement has the opposite status in each case.

employs appears to be a proper superset of the inventory of types that relaxation methods allow.

Regarding (1), we note that the mathematics of unification is quite well-understood, and this offers a foundation for a nuanced theory of ungrammaticality. The work of Fouvry (2003) especially demonstrates this. Yet, if one starts with a discrete topology, such as the typed objects of unification theory, there are many ways to introduce a metric (for the triage), and it is not easy to guess, with only the discrete elements at hand, what the metric should be. Moreover, at this point, little seems to be known about how this system interacts with probabilistic properties of languages as observed in natural corpora. An advantage of the approach we offer here is that we start with a metrized continuum (hence a complete metric space), and we specify dynamics motivated by empirical data on sentence processing. The result is a system of fixed point attractor differential equations associated with an energy function, a framework for which the mathematics of probability has been very well developed (e.g. Oppenheim, Shuler, & Weiss, 1967).

Regarding point (2), relaxation allows cases where two feature structures with identical attribute geometry, but conflicting values, unify into a single feature structure with the conflicting values blended together into a possibly novel type (e.g. Fouvry, 2003; Vogel & Cooper, 1995). In our system, the effect of such relaxations can be achieved by employing bonding rules that alter the feature structures (e.g. $N_{sg} \sim > N_{pl}$, $N_{pl} \sim > N_{sg}$). However, in allowing rewrite rules like $V_{hear} \sim > VP$ (see Fig. 9), we seem to go beyond the range of realizable relaxation coercions. In Vogel & Cooper's model, unification is only possible if the graph structures of coerced elements are isomorphic (so Fig. 9 is not possible). Fouvry's system may be more permissive, but a radical coercion like Fig. 9 gives rise to a deep type clash so, at best, the result is treated as highly ungrammatical. Our model, relying on the semantic closeness of *hear* and *hear the report*, despite their contrasting type structure, produces a much milder clash. We are not claiming that this single example is strong evidence for our model's greater permissiveness but we offer, for empirical consideration, the thesis that there are many such semantically mild coercions that involve radical type incompatibility, so a model such as ours is needed. One might worry that such wide permissiveness would result in rampant generation of impossibly bad structures. However, the harmony values, working in concert with the dynamics, prevent this. Most of the parsing visits only the highest-harmony structures proximal to each point in the state space.

5.3. Relation to probabilistic grammar models of gradient grammaticality

Another, related approach to gradient grammaticality is to parameterize a probabilistic grammar on a natural corpus containing ungrammatical structures and thus to model the ungrammatical structures as well as the grammatical ones by essentially treating them as low-probability grammatical structures (see Foster, 2007; Hashemi & Hwa, 2016; Lau, Clark, & Lappin, 2017 for discussion). While this approach has practical usefulness for designing parsers that can handle ungrammaticality as it occurs in natural corpora, we find the approach theoretically dubious. Psycholinguistic investigations provide strong evidence that probability and grammaticality are distinct mental phenomena. For example, naïve participants have a very high probability of being garden pathed by certain sentences (famously, *The horse raced past the barn fell* – Bever, 1970) and they give these low acceptability ratings at first encounter. However, if they are informed about how to parse such sentences, they will typically adjust their acceptability rating radically upward. Unlike probabilistic grammar models of gradient grammaticality, our approach models grammaticality and probability as two distinct phenomena. Grammaticality is modeled as the harmony of the stabilized parse (computable as the product of its bond harmonies). Probability is determined by the dynamics of the self-organization. In the case of *The horse raced past the barn fell*, for example, the model exhibits a garden path effect because it has an inherent anti-complexity bias: the state of the dynamical system is closer to the main clause structure than to the reduced relative clause structure at the point of having processed *the horse* so the system is unlikely to discover the reduced relative reading (Hancock & Tabor, 2021) and will end up in a very low grammaticality state because it gets stuck in the garden path. However, if the model is given a prompt ahead of time that moves it closer to the reduced relative clause state, it will adopt that state as the parse and return a high grammaticality rating.

5.4. Shortcomings

The current SOSP-GD proposal has a number of shortcomings.

First among these is that, although it is a framework that points strongly to the need to better formalize linguistic and psycholinguistic theorizing, the description we give here is itself only partially formalized. We suggest that the computational analysis presented in Section 4.3 takes a helpful step in the direction of complete formalization but as we noted there, that effort does not specify an incremental processing model. In light of our observation that some island extraction environments afford grammatical continuation and others do not, and that this difference is not correlated with the strong/weak distinction, it is important to explain why the processor avails itself of grammatical continuation possibilities in some cases and not others. Also, to rigorously test the treelet self-organizing account outlined in the main sections of the paper, it will be necessary to formalize the claims about treelet interaction and show that effective parsing of many (normal) sentences occurs while, at the same time, the mixed patterns of the weak-island cases are observed. Going further into this point, we also note that the formal claims of SOSP-GD depend on the linguistic structural analyses (the treelet forms) it adopts, and which linguistic structural analyses are correct is not yet very solidly known. On this point, while recognizing the shortcoming, we see two reasons for a positive outlook: (1) the structural assumptions we have relied on in the present analyses are generally simple and not controversial; (2) one of the main suggestions of this paper is that having a more stable and accurate linguistic theory will be helped by having a good theory of graded acceptability. Our suggestion with regard to (2) contrasts with what many generative linguists have advocated since the field was established: rather than first solving the competence problem and then moving on to the performance problem, we suggest that it may be more effective to pursue a single architectural framework that addresses both of them. Another part of the theory that needs clear formalization is the identification of semantic proximities. We have made ad hoc arguments for the semantic proximity claims that underlie our analyses. It would be better to have a systematic working method and best to have a full-fledged formal theory to determine what makes two treelets semantically compatible. For all of these formal questions, we see some promise in the possibility of linking SOSP-GD to methods in machine learning which are currently making much headway in improving applied natural language processing (e.g. Devlin, Chang, Lee, & Toutanova, 2019; Levy & Goldberg, 2014). We are especially encouraged by efforts to establish an appropriately flexible semantic theory by combining traditional semantic tools with vector-space methods (Asher, Van de Cruys, Bride and Abrusán, 2016) and to derive the types of feature vectors that are used by current sentence processing models (including SOSP-GD) using machine learning tools (Smith & Vasishth, 2020).

Another area that needs development is the treatment of D-linking effects. In SOSP-GD, if we assume that simple-wh are encoded as an average of the many complex-wh phrases, then the slash-encoding for simple-wh is close to a zero vector. This makes fixed points for the encoding of simple-wh slash bundles close to fixed points associated with no slash bundle. If two attractors in a noisy dynamical system are close together, then there is a relatively high likelihood that the system will jump from one to the other one. Therefore, with simple-wh, there is a higher probability of dropping the slash bundle than with complex-wh (recall the harmony-reducing possibility of dropping the slash-bundle introduced in Section 4.3). This predicts generally greater

ungrammaticality with simple than with complex-wh extractions, a claim that is roughly consistent with the data. However, this account predicts that Complex NP islands should show a D-linking effect. As we noted above, although the empirical evidence is somewhat mixed, our recent findings make us skeptical that there are D-linking effects in Complex NP islands. We suspect that the model's treatment of D-linking needs to be enhanced to encode semantic distinctions that are relevant for Whether islands and not for Complex NP islands (see Szabolcsi, 2002).

One may also wonder about the computational tractability of exploring general flexible parsing. Here, we explored trees up to a fixed depth, limiting the set of coercions entertained. To build a full SOSP-GD model, we would need to consider all possible coercions. Keeping depth fixed at d and considering all ways of coercing one rule symbol to another, if the number of rules is r, then a lower bound on the number of attractors is $n(r,d) = r^{2^d-1}$. This is quite intractable for exhaustive exploration of even a grammar of very modest size (e.g. n(8 rules, depth 8) $\approx 1.9 \times 10^{230}$). However, most of the coercions produce very bad structures so they exert very little influence on the dynamics and can be ignored. Additionally, at any juncture between words, the dynamics are dominated by the influence of a few proximal attractors corresponding to ways of parsing the next word – a situation only slightly more expensive to compute than standard parsing on account of the presence of some reasonably viable coercions. Thus, although more investigation needs to be done, we think the tractability challenge is approachable.

Finally, we recognize that the current results are based on a very small set of cases: only three island types. The theory needs to be explored across a wider range of structural types before its viability can be solidly assessed.

5.5. *The role of similarity in linguistic and psycholinguistic theorizing*

The notion of analogy/similarity, which is key to interpretable coercion, has generally not played a central role in linguistic theorizing. Relativized Minimality, which we identified in Introduction, is an exception to this claim. It is a theory of locality constraints in linguistic information structure that makes predictions about a number of island phenomena. As we noted, this theory attributes the ungrammaticality of wh-islands (and possibly other types of weak islands) to the syntactic featural similarity between the extracted element and a structurally intervening one (see, among others, Atkinson, Aaron, Kyle, & Omaki, 2015; Friedmann, Belletti, & Rizzi, 2009; Starke, 2001; Rizzi, 1990, 2001, 2004; Villata, Rizzi, & Franck 2016). The theory has several commonalities with the similarity-based interference account that has been developed in the psycholinguistics literature (see Villata, 2017 for discussion). This account aims to capture the increased processing difficulty (and the consequent decrease in acceptability) of otherwise fully grammatical sentences, such as object relative clauses as compared to subject relative clauses (see, among many others, McElree, 2000; Gordon, Hendrick, & Levine, 2002; McElree, Foraker, & Dyer, 2003; Lewis & Vasishth, 2005; Van Dyke & McElree, 2006; McElree, 2006; Van Dyke, 2007; Friedmann, Belletti, & Rizzi, 2009; Van Dyke & McElree, 2011). In particular, the account claims that object relative clauses are harder to parse than subject relative clauses because of the intervention of the subject between the object and its trace. It has been shown that the more similar the two elements are (e.g. both are full NPs vs. one is a full NP and the other is a pronoun or a proper name), the harder the sentence becomes (e.g. Gordon, Hendrick, & Johnson, 2001). The hypothesis is that similarity decreases the distinctiveness of the elements in memory, which hinders their retrieval from memory. In subject relative clauses, by contrast, no interference arises because all elements are in their source position, which renders processing straightforward.

Although both coercion and similarity-based interference theories grant a key role to similarity, similarity turns out to play opposite roles in the two (facilitatory in the former, detrimental in the latter). This is because their explanatory goals are different. (Interpretable) coercion aims to account for the parsability of ungrammatical sentences, i.e. why sentences that are claimed to violate a rule of the grammar turn out to be parsable, understandable, and relatively acceptable. Theories of interference, on the contrary, aim to explain either why ungrammatical sentences are ungrammatical (as is the case in Relativized Minimality, which basically aims at explaining why wh-islands are islands), or why some grammatical sentences are harder to parse than others. Hence, while our focus on coercion here is on what makes an ungrammatical sentence "better", the focus of theories of interference is on what makes a grammatical sentence "worse" or an ungrammatical sentence ungrammatical. Relatedly, the elements that matter in the calculation of similarity differ in the two cases. In the case of coercion, what matters is the similarity between sentential elements of the target sentence (e.g. *wonder whether* in a Whether island) and sentential elements of a nearby grammatical sentence (e.g. *think that* in an extraction from a declarative). In the case of interference, what counts instead is the similarity between two elements of the target sentence itself, namely a long-distant element and an intervening one (e.g. the object and the subject in a relative clause; the extracted wh-element and the intervening one in a wh-island). In fact, SOSP-GD generates similarity effects of both types (see Smith, Franck, & Tabor, 2018, 2021).

Another theory of syntactic processing that grants a key role to similarity is the "creative analogy" approach of Bever & colleagues (e.g. Bever, 1974; Bever, Carroll, & Hurtig, 1976; Carroll, 1980). This theory claims that some ungrammatical strings are understandable because they are analogous to fully grammatical existing forms. However, the presence of a fully grammatical analogy also conspires to keep these sentences unacceptable (since the linguistic system already has a grammatical way to express that content). Although the concept of coercion as adopted here is reminiscent of the concept of creative analogy, it also differs from it in some relevant respects. First, in coercion, the analogy is not a synonymous alternative to the ungrammatical form, as is the case for creative analogy. The to-be-coerced structure must be abstractly close in meaning to the grammatical form, but there is no requirement of identity of meaning. Hence, the conditions of applicability of coercion are looser than those of creative analogy. Second, as discussed in Sections 4.1 and 5.1 above, the possibility of interpretable coercion in SOSP-GD does not depend on the existence of a sentential analog per se – instead, it depends on what the treelets activated by the input sequence are able to do. In principle, SOSP-GD can support interpretable coercions for which no grammatical analog exists.

Given that we are arguing for an expanded role of coercion in the theory of grammar, there is much more to be said on the topic of coercion. Although it is beyond the scope of this paper to go deeply into the issues in this area, we note that our framework has a desirable property with regard to one of the core questions connected with coercion: how to constrain its use (Lauwers & Willems, 2011). Generative linguistics usually takes its *metier* to be the specification of the set of grammatical constructions that a language employs. Under this approach, if coercion is a possibility, then there has to be a good way to keep it from overgenerating. So far, it has proved difficult to discover a simple mechanism that can draw the right boundary. Our approach offers a fresh angle on this problem: do not draw a boundary at all; instead, accept complex gradation into failure of sense and model the whole gamut. Even though this is also a tall order, we suspect the search for a boundary is misplaced: the edge of grammaticality may be inherently ragged (Tabor, Cho, & Szkudlarek, 2013).

6. Conclusion

We have presented a self-organizing theory of the strong/weak island distinction including gradient aspects of this phenomenon. The theory takes seriously the claim that syntax is a rule-based system, but, unlike traditional theories of grammar, it allows for greater flexibility in the syntactic system. In a self-organizing parser, structure is formed even under sub-optimal circumstances through coercion, a mechanism that

allows elements to play roles that do not fully fit them but may be the best way for them to satisfy their combination requirements at a particular moment in processing. Coercion comes in two forms: interpretable coercion, where the theta-criterion is fully satisfied, and uninterpretable coercion, where some thematic constraints are violated. We noted that interpretable coercion often involves a mildly ungrammatical form that is structurally analogous to a grammatical form. We argued that weak islands are weak because the structures can be interpretably coerced, allowing the parser to establish a dependency between an extracted element and a gap inside a putative island domain. Correspondingly, what renders strong islands strong is the fact that these structures cannot undergo interpretable coercion; in this case, self-organization produces a vacuous quantification structure. Gradience reflects the ease with which the system can undergo coercion: interpretable coercions involve a small number of feature clashes in a thematically coherent structure; in uninterpretable coercion, entire arguments are either present and unlicensed or missing, but required, so the featural clash is extensive. This gives rise to a both fine gradations of acceptability values and a binary contrast between coherent and incoherent outcomes.

All in all, we argue that SOSP-GD offers a valuable new way of approaching the strong/weak islands distinction, and the relationship between grammar and processing more generally. It is based on generative linguistic theory. Nevertheless, it differs in non-trivial ways from traditional assumptions, notably continuity in the grammar, a more flexible rule-based system, and a central role for processing in grammatical explanation. We hope our results will spur new discussion on these topics.

Credit author statement

Sandra Villata and Whitney Tabor were both involved in the development of the theory. Whitney Tabor implemented the computational model. Sandra Villata and Whitney Tabor jointly wrote the manuscript.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We are extremely grateful to Jon Sprouse for his collaboration in designing and analyzing the acceptability judgment and maze task experiments we discuss in this paper (and that will be reported in greater detail in a separate paper), as well as for countless discussions on gradient effects in islands and several theoretical aspects of this paper. The second author extends his thanks to Jessica Brown for many engaging discussions about related island constructions, the theory of syntax, and its relation to semantics. This work has also benefited from presentation and discussion at a variety of venues, including CUNY 2019, CUNY 2020, and AMLaP 2020. The research reported in this article was supported by a postdoctoral grant from the Marica De Vincenzi Foundation, as well as by research grants from the Institute for Brain and Cognitive Science (IBACS) and the Neurobiology of Language Program (NBL) at the University of Connecticut.

Appendix A. Grammar used for the computational investigation in Section 4.3

Abbreviations: WhQ = Wh question phrase, Wh[bare] = phrasal node that generates *what*, Wh[complex] = phrasal node that generates *which N* (since there is only one noun in this grammar, only one form of this phrase occurs), S = sentence, VP = verb phrase, CP = complementizer phrase, NP = noun phrase, C = complementizer, N = noun, [max] = maximal phrasal node of the category indicated, [intrans] = intransitive (verb), [trans] = transitive (verb), [sentcomp] = sentence complement (verb)

Notes (see also the discussion in Section 4.3): (1) The node "Root" is the starting point of all derivations; (2) Although the rules of the grammar are divided into Grammatical and Lexical Rules to make it easy to study its structure, the parsing algorithm makes no distinction between the two types; (3) the grammar only generates wh-questions, since all the cases under consideration are wh-questions; (4) for simplicity, the grammar does not implement the rules of English do-support or English subject-verb agreement (including these would not change the results); (5) the notation provided here employs features on node names (like [bare], [trans], /What), thus highlighting the feature-passing structure of slash propagation and type-inheritance. However, the implementation performs no explicit feature passing – under the theory of metarules (Gazdar, 1982), unambiguous labeling suffices to implement feature-passing logic in a CFG.

Grammatical Rules:

Root ➔ WhQ Q
WhQ ➔ Wh[bare] S[max]/What
WhQ ➔ Wh[complex] Smax/WhichN
S[max] ➔ S
S[max] ➔ S CP[because]
S[max]/What ➔ S/What
S[max]/What ➔ S/What CP[because]
S[max]/WhichN ➔ S/WhichN
S[max]/WhichN ➔ S/WhichN CP[because]
S ➔ NP VP
S/What ➔ NP VP/What
S/WhichN ➔ NP VP/WhichN
VP ➔ V[intrans]
VP ➔ V[trans] NP
VP ➔ V[sentcomp] CP[that]
VP ➔ V[wonder] CP[whether]
VP/What ➔ V[trans]
VP/What ➔ V[sentcomp] CP[that]/What
VP/WhichN ➔ V[trans]
VP/WhichN ➔ V[sentcomp] CP[that]/WhichN
NP ➔ N

NP ➔ N CP[that]
CP[that] ➔ C[that] S
CP[that] ➔ S
CP[that]/What ➔ C[that] S/What
CP[that]/What ➔ S/What
CP[that]/WhichN ➔ C[that] S/WhichN
CP[that]/WhichN ➔ S/WhichN
CP[because] ➔ C[because] S
CP[whether] ➔ C[whether] S
Lexical Rules:
Q ➔?
Wh[bare] ➔ what
Wh[complex] ➔ whichN
NP ➔ report
V[intrans] ➔ sleep
V[trans] ➔ adore
V[trans] ➔ hear
V[sentcomp] ➔ hear
V[wonder] ➔ wonder
C[because] ➔ because
C[that] ➔ that
C[whether] ➔ whether

A.1. Coercions considered in the computational analysis

Table A1 shows the set of coercions we explored. To track adherence to thematic requirements, we have classified each coercion with a 1 if it maintains thematic coherence and a 0 if it violates thematic coherence. All grammatical bonds (these are not listed in Table A1) have a thematic coherence of 1. As noted in the text, each coercion can be implemented by adding a rule to the grammar of the form Daughter-above ➔ Mother-below. To find all coerced analyses of a sentence, we parsed the sentence using the augmented grammar, prohibiting iterated coercions. For the coercion V [hearsent] ➔ VP we took the thematic status to be 1 if the VP spanned exactly *hear report* and 0 otherwise.

Table A1
Coercions considered in the computational analysis. C = complementizer, V = verb, VP = verb phrase.

Coercion source (daughter node above)	Coercion outcome (mother node below)	Type	Thematic Status
C[that]	C[whether]	C-to-C	1
C[that]	C[because]	C-to-C	1
C[whether]	C[that]	C-to-C	1
C[whether]	C[because]	C-to-C	1
C[because]	C[that]	C-to-C	1
C[because]	C[whether]	C-to-C	1
V[intrans]	V[trans]	V-to-V	0
V[intrans]	V[sentcomp]	V-to-V	0
V[intrans]	V[wonder]	V-to-V	0
V[trans]	V[intrans]	V-to-V	0
V[trans]	V[sentcomp]	V-to-V	0
V[trans]	V[wonder]	V-to-V	0
V[sentcomp]	V[intrans]	V-to-V	0
V[sentcomp]	V[trans]	V-to-V	0
V[sentcomp]	V[wonder]	V-to-V	1
V[wonder]	V[intrans]	V-to-V	0
V[wonder]	V[trans]	V-to-V	0
V[wonder]	V[sentcomp]	V-to-V	1
V[sentcomp]	VP	V-to-VP	1/0
S[max]/WhichN	S[max]	Slash drop	0
S/WhichN	S	Slash drop	0
VP/WhichN	VP	Slash drop	0
CP[that]/WhichN	CP[that]	Slash drop	0

References

Abeillé, A., Hemforth, B., Winckel, E., & Gibson, E. (2020). Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition, 204*, 104293.

Abraham, R. H., & Shaw, C. D. (2000). *Dynamics: The geometry of behavior*. Aerial Press, Inc.

Abrusán, M. (2011). Wh-islands in degree questions: A semantic approach. *Semantics and Pragmatics, 4*, 5–1.

Abrusán, M. (2014). *Weak island semantics (Vol. 3)*. OUP Oxford.

Almeida, D. (2014). Subliminal wh-islands in Brazilian Portuguese and the consequences for syntactic theory. *Revista da ABRALIN, 13*, 55–93. https://doi.org/10.5380/rabl.v13i2, 39611.

Ambridge, B. (2015). Island constraints and overgeneralization in language acquisition. *Cognitive Linguistics, 26*(2), 361–370.

Ambridge, B., & Goldberg, A. E. (2008). The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics, 19*(3), 357–389.

Asher, N., Hunter, J., Morey, M., Benamara, F., & Afantenos, S. (2016, May). *Discourse structure and dialogue acts in multiparty dialogue: The STAC corpus*.

Asher, N., Van de Cruys, T., Bride, A., & Abrusán, M. (2016). Integrating type theory and distributional semantics: A case study on adjective–noun compositions. *Computational Linguistics, 42*(4), 703–725.

Atkinson, E., Aaron, A., Kyle, R., & Omaki, A. (2015). Similarity of wh-phrases and acceptability variation in wh-islands. *Frontiers in Psychology, 6*, 2048. https://doi.org/10.3389/fpsyg.2015.02048

Bailey, K. G., & Ferreira, F. (2003). Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language, 49*(2), 183–200.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*(Feb), 1137–1155.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: John Wiley and Sons.

Bever, T. G. (1974). The ascent of the specious, or there's a lot we don't know about mirrors. In D. Cohen (Ed.), *Explaining linguistic phenomena* (pp. 173–200). Washington, DC: Hemisphere.

Bever, T. G., Carroll, J. M., & Hurtig, R. (1976). Analogy or ungrammatical sequences that are utterable and comprehensible are the origins of new grammars in language acquisition and linguistic evolution. In T. G. Bever, J. J. Katz, & D. T. Langendoen (Eds.), *An integrated theory of linguistic ability* (pp. 149–182). New York: T. Y. Crowell press.

Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology, 43*(2), 83–128.

Boeckx, C. (2012). *Syntactic islands*. Cambridge University Press.

Braze, D., Shankweiler, D., Ni, W., & Palumbo, L. C. (2002). Readers' eye movements distinguish anomalies of form and content. *Journal of Psycholinguistic Research, 31*(1), 25–44.

Brown, J. (2015). Blackholes and subextraction. *CLS, 51*, 67–81.

Brown, J. (2017). *Heads and Adjuncts. PhD. Dissert., U. Cambridge*.

Carroll, J. M. (1980). Creative analogy and language evolution. *Journal of Psycholinguistic Research, 9*(6), 595–617.

Cho, P. W., Goldrick, M., Lewis, R. L., & Smolensky, P. (2018, January). Dynamic encoding of structural uncertainty in gradient symbols. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)* (pp. 19–28).

Cho, P. W., Goldrick, M., & Smolensky, P. (2017). Incremental parsing in a continuous dynamical system: Sentence processing in gradient symbolic computation. *Linguistics Vanguard, 3*(1), 1–10.

Cho, P. W., Goldrick, M., & Smolensky, P. (2020). Parallel parsing in a gradient symbolic computation parser. Doi:10.31234/osf.io/utcgv.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1986). *Barriers*. Cambridge: MIT Press.

Chomsky, N. (1995). Categories and transformations. In N. Chomsky (Ed.), *The Minimalist Program* (pp. 219–394). Cambridge, Mass: MIT Press.

Chomsky, N. (2000). Minimalist inquiries: The framework. In R. Martin, D. Michaels, & J. Uriagereka (Eds.), *Step by Step* (pp. 89–155). Cambridge, Mass: MIT press.

Chomsky, N. (2001). Derivation by phase. In M. Kenstowicz (Ed.), *Ken Hale: A Life in Language* (pp. 1–50). Cambridge, Mass.: MIT press.

Chomsky, N. (2008). On phases. In R. Freidin, C. Otero, & M.-L. Zubizarreta (Eds.), *Foundational Issues in Linguistic Theory* (pp. 133–166). Cambridge, Mass.: MIT Press.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson, & P. Kiparsky (Eds.), *A festschrift for Morris Halle. New York: Holt, Rinehart & Winston* (pp. 232–286).

Christensen, K. R., Kizach, J., & Nyvad, A. M. (2013). Escape from the island: Grammaticality and (reduced) acceptability of wh-island violations in Danish. *Journal of Psycholinguistic Research, 42*, 51–70.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology, 42*(4), 368–407.

Cinque, G. (1990). *Types of Ā-dependencies*. Cambridge: MIT.

Clifton, C., Jr., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Lang & Ling Compass, 2*(2), 234–250.

Cottrell, G. W., & Small, S. L. (1983). *A connectionist scheme for modelling word sense disambiguation* (Cognition & Brain Theory).

Davies, W. D., & Dubinsky, S. (2003). On extraction from NPs. *Natural Language & Linguistic Theory, 21*(1), 1–37.

De Vincenzi, M., Job, R., Di Matteo, R., Angrilli, A., Penolazzi, B., Ciccarelli, L., & Vespignani, F. (2003). Differences in the perception and time course of syntactic and semantic violations. *Brain and Language, 85*(2), 280–296.

Deane, P. (1991). Limits to attention: A cognitive theory of island phenomena. *Cognitive Linguistics, 2*, 1–63.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for language understanding. *Proceedings of NAACL-HLT, 1*, 4171–4186.

Dillon, B., Andrews, C., Rotello, C. M., & Wagers, M. (2019). A new argument for co-active parses during language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(7), 1271.

Dillon, B., Staub, A., Levy, J., & Clifton, C., Jr. (2017). Which noun phrases is the verb supposed to agree with?: Object agreement in American English. *Language, 93*(1), 65–96.

Dillon, B., & Wagers, M. (2019, November 19). Approaching gradience in acceptability with the tools of signal detection theory. Doi:10.31219/osf.io/apxru.

Douglas, S., & Dale, R. (1992). Towards robust PATR. In *COLING 1992 volume 2: The 14th international conference on computational linguistics*.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review, 112*(3), 531.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*(2), 195–225.

Erteschik-Shir, N. (1973). *On the nature of island constraints*. MIT: PhD dissertation.

Featherston, S. (2005). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua., 115*, 1525–1550.

Ferreira, F., & Bailey, K. G. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences, 8*(5), 231–237.

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language, 30*(6), 725–745.

Fodor, J. D., & Inoue, A. (1998). Attach anyway. In J. D. Fodor, & F. Ferreira (Eds.), *21. Reanalysis in Sentence Processing. Studies in Theoretical Psycholinguistics* (pp. 101–141). Springer Netherlands. https://doi.org/10.1007/978-94-015-9070-9_4.

Fodor, J. D., Ni, W., Crain, S., & Shankweiler, D. (1996). Tasks and timing in the perception of linguistic anomaly. *Journal of Psycholinguistic Research, 25*, 25–57.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods, 41*(1), 163–171.

Foster, J. (2007). Treebanks gone bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal on Document Analysis and Recognition, 10*, 129–145.

Fouvry, F. (2003). *Robust processing for constraint-based grammar formalisms*. University of Essex: Doctoral dissertation.

Franck, J., & Wagers, M. (2020). Hierarchical structure and memory retrieval mechanisms in agreement attraction. *Plos One, 15*(5), e0232163. https://doi.org/10.1371/journal.pone.0232163

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research, 1*(3), 183–192.

Friedmann, N., Belletti, A., & Rizzi, L. (2009). Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua, 119*(1), 67–88.

Gazdar, G. (1982). Phrase structure grammar. In P. Jakobson, & G. K. Pullum (Eds.), *The nature of syntactic representation* (pp. 131–186). D. Reidel: Dordrecht, Holland.

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language & Cognitive Processes, 28*(3), 229–240.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Goodall, G. (2015). The D-linking effect on extraction from islands and non-islands. *Frontiers in Psychology, 5*, 1493.

Gordon, D. M. (2010). *Ant encounters: Interaction networks and colony behavior*. Princeton, NJ: Princeton University Press.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1411.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language, 51*(1), 97–114.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science, 13*(5), 425–430.

Haken, H. (2004). *Synergetics. An Introduction* (3rd). Berlin: Springer-Verlag.

Haken, H. (2008). Self-organization. *Scholarpedia, 3*(8), 1401.

Hale, J. (2001). *A probabilistic Earley parser as a psycholinguistic model* (In Second meeting of the north American chapter of the association for computational linguistics).

Hancock, R., & Tabor, W. (2021, March). *Interaction between local coherence and garden path effects supports a nonlinear dynamical model of sentence processing.Poster presented at the 34th annual CUNY Conference on Human Sentence Processing*. Philadelphia, PA, USA: University of Pennsylvania.

Hashemi, H. B., & Hwa, R. (2016). An evaluation of parser robustness for ungrammatical sentences. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1765-1774)*. Association for Computational Linguistics.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language, 86*(2), 366.

Hofmeister, P., & Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology, 5*, 1237.

Huang, C. T. J. (1982). *Logical relations in Chinese and the theory of grammar. Ph.D. thesis, MIT*.

Jackendoff, R. (2002). *Foundations of language: Brain*. Meaning: Grammar, Evolution, Oxford University Press.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111*(2), 228–238.

Kalyan, S. (2012). Similarity in linguistic categorization: The importance of necessary properties. *Cognitive Linguistics, 23*(3), 539–554.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. USA: Oxford University Press.

Keller, E. F., & Segel, L. A. (1970). Conflict between positive and negative feedback as an explanation for the initiation of aggregation in slime mould amoebae. *Nature, 227*, 1365–1366. https://doi.org/10.1038/2271365a0

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality (doctoral dissertation)*.

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*(3), 273–290.

Keshev, M., & Meltzer-Asscher, A. (2019). A processing-based account of subliminal wh-island effects. *Natural Language & Linguistic Theory, 37*(2), 621–657.

Kim, B., & Goodall, G. (2016). Islands and non-islands in native and heritage Korean. *Frontiers in Psychology, 7*, 134.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30*(5), 580–602.

Kluender, R. (1998). On the distinction between strong and weak islands: A processing perspective. In P. Culicover, & L. McNally (Eds.), *Syntax and semantics 29: The limits of syntax* (pp. 241–279). San Diego, CA: Academic Press.

Kluender, R. (2004). Are subject islands subject to a processing account. In *Proceedings of WCCFL (Vol. 23, pp. 475-499). Somerville, MA*. Cascadilla Press.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language & Cognitive Processes, 8*, 573–633.

Kluender, R. E. (1991). *Cognitive constraints on variables in syntax*. University of California, San Diego: Doctoral dissertation.

Konieczny, L. (2005, August). The psychological reality of local coherences in sentence processing. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1178–1183). Stresa, Italy: Cognitive Science Society.

Koschmieder, E. L. (1993). *Bénard cells and Taylor vortices*. Cambridge: Cambridge University Press.

Kuno, S. (1976). Subject, theme, and speaker's empathy: A reexamination of relativization phenomena. In C. Li (Ed.), *Subject and topic* (pp. 417–444). New York: Academic Press.

Kush, D., Lohndal, T., & Sprouse, J. (2019). On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language, 95*(3), 393–420.

Kush, D., Lohndal, T., & Sprouse, J. (2018). Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory, 36*, 743–779.

Lakoff, G. (1973). Fuzzy grammar and the performance/competence terminology game. In C. Claudia, T. Cedric Smith-Stark, & A. Weiser (Eds.), *Papers from the 9th meeting of the Chicago linguistic society, 271–291*. Chicago.

Lasnik, H., & Saito, M. (1984). On the nature of proper government. *Linguistic Inquiry, 15*, 235–289.

Lasnik, H., & Saito, M. (1992). *Move alpha: Conditions on its application and output*. Cambridge, MA: MIT Press.

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science, 41*(5), 1202–1241.

Lauwers, P., & Willems, D. (2011). Coercion: Definition and challenges, current approaches, and new trends. *Linguistics, 49*(6), 1219–1235.

Legendre, G., Wilson, C., Smolensky, P., Homer, K., & Raymond, W. (1995). Optimality and Wh-extraction. In papers in optimality theory (J. Beckman, S. Urbanczyck, and L. Walsh, eds.). *UMOP, 18*, 607–636.

Levy, O., & Goldberg, Y. (2014, June). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 302–308).

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

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research, 25*(1), 93–115.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*(3), 375–419.

Lu, J., Thompson, C. K., & Yoshida, M. (2020). Chinese wh-in-situ and islands: A formal judgment study. *Linguistic Inquiry, 51*(3), 611–623.

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, Mass: MIT Press.

Marée, A. F. M., & Hogeweg, P. (2001). How amoeboids self-organize into a fruiting body: Multicellular coordination in Dictyostelium discoideum. In , *98. Proceedings of the National Academy of Sciences of the United States of America* (pp. 3879–3883). https://doi.org/10.1073/pnas.061535198

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah: Lawrence Erlbaum Associates.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology, 4*, 503.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science, 38*(6), 1139–1189.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological Review, 88*(5), 375.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research, 29*(2), 111–123.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language, 48*(1), 67–91.

McElree, B., & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1), 134.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation* (p. 46). San Diego: Academic Press, Vol.

Michaelis, L. (2004). Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive Linguistics, 15*, 1–67.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119). Curran Associates.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *vol. II. Handbook of mathematical psychology* (pp. 269–321). New York, NY: Wiley.

Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics, 14*, 15–28.

Müller, G. (2019). The Third Construction and Strength of C: A Gradient Harmonic Grammar Approach. In Ken Ramshøj Christensen, Henrik Jørgensen, & Johanna L. Wood (Eds.), *The Sign of the V – Papers in Honour of Sten Vikner* (pp. 419–448). Aarhus: Dept. of English, School of Communication and Culture, Aarhus University.

Ni, W., Fodor, J. D., Crain, S., & Shankweiler, D. (1998). Anomaly detection: Eye movement patterns. *Journal of Psycholinguistic Research, 27*(5), 515–539.

Nicolis, G., & Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems. From dissipative structures to order through fluctuations*. New York, London, Sydney, Toronto: J. Wiley & Sons.

Omaki, A., Fukuda, S., Nakao, C., & Polinsky, M. (2019). Subextraction in Japanese and subject-object symmetry. *Natural Language and Linguistic Theory*. https://doi.org/10.1007/s11049-019-09449-8

Oppenheim, I., Shuler, K. E., & Weiss, G. H. (1967). On the decay of initial correlations in stochastic processes. *The Journal of Chemical Physics, 46*(10), 4100–4111.

Paape, D., & Vasishth, S. (2016). Local coherence and preemptive digging-in effects in German. *Language and Speech, 59*(3), 387–403.

Pañeda, C., Lago, S., Vares, E., Veríssimo, J., & Felser, C. (2020). Island effects in Spanish comprehension. *Glossa: a journal of general linguistics, 5*(1).

Pesetsky, D. (1987). Wh-in-situ: Movement and unselective binding. In E. Reuland, & A. Ter Meulen (Eds.), *The representation of (in)definiteness* (pp. 98–129). Cambridge, MA: MIT Press.

Philip, W., & de Villiers, J. (1992). Monotonicity and the acquisition of weak wh-islands. In *Proceedings of the 24th Annual Child Language Research Forum* (pp. 99–111). Cambridge University Press.

Phillips, C. (2006). The real-time status of island phenomena. *Language, 82*(4), 795–823.

Piñango, María M., Zurif, Edgar, & Jackendoff, Ray (1999). Real-time processing implications of enriched composition at the syntax–semantics interface. *Journal of Psycholinguistic Research, 28*, 395–414.

Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago and Stanford: University of Chicago Press and CSLI Publications.

Prince, A., & Smolensky, P. (2008). *Optimality theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372.

Rizzi, L. (1982). Violations of the wh-island constraint and the subjacency condition. In L. Rizzi (Ed.), *Issues in Italian syntax* (pp. 49–76). Dordrecht: Foris.

Rizzi, L. (1990). *Relativized minimality*. The MIT Press.

Rizzi, L. (2001). On the position "Int(errogative)" in the left periphery of the clause. In G. Cinque, & G. Salvi (Eds.), *Current studies in Italian syntax, essays offered to Lorenzo Renzi* (pp. 287–296). Amsterdam: Elsevier.

Rizzi, L. (2004). Locality and the left periphery. In A. Belletti (Ed.), *Structures and beyond: The cartography of syntactic structures vol. 3* (pp. 223–251). Oxford: Oxford University Press.

Ross, J. R. (1967). *Constraints on variables in syntax*. MIT, Cambridge, Mass: Doctoral Dissertations.

Ross, J. R. (1972). The category squish: Endstation Hauptwort. In *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society* (pp. 316–338).

Sag, I., & Pollard, C. (1991). An integrated theory of complement control. *Language, 67*(1), 63–113.

Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject–verb number agreement. *Cognitive Science, 42*, 1043–1074.

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology, 124*, 101356.

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 (pp. 138-143)*.

Smith, G., & Vasishth, S. (2020). A principled approach to feature selection in models of sentence processing. *Cognitive Science, 44*(12), Article e12918.

Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory*. Colorado Univ at Boulder Dept of Computer Science.

Smolensky, P., & Goldrick, M. (2016). Gradient symbolic representations inGrammar: e Case of French Liaison. *Rutgers Optimality Archive, 1286*.

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science, 38*(6), 1102–1138.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua, 115*(11), 1497–1524.

Sprouse, J. (2007). *A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge (doctoral dissertation)*.

Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory, 34*(1), 307–344.

Sprouse, J., & Messick, T. (2015). How gradient are island effects. *In 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.

Starke, M. (2001). *Move dissolves into merge: A theory of locality*. Doctoral Dissertation: University of Geneva.

Steedman, M. (2000). *The syntactic process*. The MIT Press.

Stepanov, A., Mušič, M., & Stateva, P. (2018). Two (non-) islands in Slovenian: A study in experimental syntax. *Linguistics, 56*(3), 435–476.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research, 23*(4), 295–322.

Szabolcsi, A., & Lohndal, T. (2017). Strong vs. weak islands. In *The Wiley Blackwell companion to Syntax* (2nd ed., pp. 1–51).

Szabolcsi, A., & Zwarts, F. (1990). Semantic properties of composed functions and the distribution of whphrases. In M. Stokhof, & L. Torenvliet (Eds.), *Proceedings of the Seventh Amsterdam Colloquium* (pp. 529–555). Amsterdam: ILLI.

Szabolcsi, A., & Zwarts, F. (1993). Weak islands and an algebraic semantics for scope taking. *Natural Language Semantics, 1*(3), 235–284.
Szabolcsi, A. (2002). In I. Kenesei, & P. Siptár (Eds.), *'Hungarian disjunctions and positive Polarity* (pp. 217–241).
Szabolcsi, A., & Zwarts, F. (1997). Weak islands and an algebraic semantics for taking scope. In *Ways of Scope Taking* (pp. 217–262). Dordrecht: Springer.
Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. *Expert Systems, 17*(1), 41–56.
Tabor, W. (2009). A dynamical systems perspective on the relationship between symbolic and non-symbolic computation. *Cognitive Neurodynamics, 3*(4), 415–427.
Tabor, W. (2011). Recursion and recursion-like structure in ensembles of neural elements. In *unifying themes in complex systems*. In *Proceedings of the viii international conference on complex systems (pp. 1494-1508). Cambridge, MA: New England*. Complex Systems Institute.
Tabor, W. (2021). On the relationship between syntactic and semantic encoding in metric space language models. *Journal of Cognitive Science, 22*(2), 135–155.
Tabor, W., Cho, P. W., & Szkudlarek, E. (2013). Fractal analysis illuminates the form of connectionist structural gradualness. *Topics in Cognitive Science, 5*, 634–667.
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.
Torrego, E. (1984). On inversion in Spanish and some of its effects. *Linguistic Inquiry, 15*, 103–129.
Traxler, M. J., Pickering, M. J., & McElree, B. (2002). Coercion in Sentence Processing: Evidence from Eye-Movements and Self-Paced Reading. *Journal of Memory and Language, 47*(1), 530–547.
Truswell, R. (2017). Extraction from adjuncts and the structure of events. *Lingua, 117*, 1355–1377.
Van Dyke, J. A. (2002). *Retrieval effects in sentence parsing and interpretation*. Doctoral dissertation: University of Pittsburgh.
Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 407.
Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language, 49*(3), 285–316.
Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language, 55*(2), 157–166.
Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language, 65*(3), 247–263.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., … Polosukhin, I. (2017). *31st conference on neural information processing systems (NIPS 2017)*. CA, US: Long Beach.
van der Velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences, 29*, 37–108.
Villata, S. (2017). *Intervention effects in sentence processing*. University of Geneva: Unpublished doctoral dissertation.
Villata, S., & Franck, J. (2019). *Similarity-based interference in agreement comprehension and production: Evidence from object agreement*. Journal of Experimental Psychology: Learning, Memory, and Cognition.
Villata, S., & Lorusso, P. (2020). When initial thematic roles attribution lingers: Evidence for digging-in effects in Italian relative clauses. In V. Torrens (Ed.), *Typical and impaired processing in Morphosyntax*. Amsterdam: John Benjamins.
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. (2019). Modeling ungrammaticality: A self-organizing model of islands. CogSci. In A. K. Goel, C. M. Seifert, & C. Freska (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 1178–1184). Montreal, QB: Cognitive Science Society.
Villata, S., Tabor, W., & Sprouse, J. (2020a). Gap-filling in English syntactic islands: Evidence from forced choice and maze tasks. In *Poster presented at the 26TH Architectures and Mechanisms for Language Processing Conference. September 3*.
Villata, S., Tabor, W., & Sprouse, J. (2020b). Gap-filling in syntactic islands: Evidence for island penetrability from the maze tasks. In *The 33st Annual CUNY Conference on Human Sentence Processing, Amherst (US)* (pp. 19–21). March.
Vogel, C., & Cooper, R. (1995). Robust chart parsing with mildly inconsistent feature structures. *Nonclassical feature systems, 10*, 197–216.
Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and a lexicalist grammar. *Cognition, 75*, 105–143.
Vosse, T., & Kempen, G. (2009). The unification space implemented as a localist neural net: Predictions and error-tolerance in a constraint-based parser. *Cognitive Neurodynamics, 3*(4), 331–346.
Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language, 61*(2), 206–237.
Zhabotinsky, A. M. (1991). A history of chemical oscillations and waves. *Chaos, 1*, 379–386.