MG Parsing as a Model of Gradient Acceptability in Syntactic Islands

MG Parsing as a Model of Gradient Acceptability in Syntactic Islands

Aniello De Santo
Department of Linguistics
Institute for Advanced Computational Science
Stony Brook University
aniello.desanto@stonybrook.edu

Abstract

It is well-known that the acceptability judgments at the core of current syntactic theories are continuous. However, an open debate is whether the source of such gradience is situated in the grammar itself, or can be derived from extra-grammatical factors. In this paper, we propose the use of a top-down parser for Minimalist grammars (Stabler, 2013; Kobele et al., 2013; Graf et al., 2017), as a formal model of how gradient acceptability can arise from categorical grammars. As a test case, we target the acceptability judgments for island effects collected by Sprouse et al. (2012a).

1 Introduction

The human judgments linguists use to evaluate the adequacy of syntactic theories fall in a wide, non-binary spectrum of acceptability — a fact well-known from the early days of generative grammar (Chomsky, 1956, 1965, a.o.). Nonetheless, mainstream syntax has long claimed that grammatical knowledge is, at its core, categorical, and that *gradience* in acceptability judgments comes from extra-grammatical factors (Sprouse, 2007, a.o.). However, the rise of experimental methods in theoretical syntax has renewed the question of whether gradience should be integrated in grammatical theories directly, for instance in the form of probabilistic models (Keller, 2000; Crocker and Keller, 2005; Sorace and Keller, 2005; Lau et al., 2014, 2015, 2017).

As the relation between grammaticality and acceptability is not transparent, constructing a well-specified theory of how gradient acceptability arises from grammatical knowledge is clearly valuable. From an empirical perspective, however, categorical approaches seem to be at a disadvantage when compared to gradient grammatical models rooted in quantitative, probabilistic frameworks.

There is an abundance of well-known proposals about the way syntactic structure and cognitive resources can be integrated to derive connections between acceptability and processing difficulty (e.g., Yngve, 1960; Wanner and Maratsos, 1978; Rizzi, 1990; Rambow and Joshi, 2015; Gibson, 2000; McElree et al., 2003; Lewis and Vasishth, 2005, a.o.). However, few models based on current grammatical formalisms have been implemented in precise computational frameworks (cf. Boston, 2010). In order to have a complete theory of how acceptability judgments correlate to categorical grammars, what seems to be necessary is a formal model of the syntactic structures licensed by said grammars, and a theory of how such structures interact with extra-grammatical factors to derive differences in acceptability. This would make it possible to test how assumptions about fine-grained syntactic details lead to quantifiable predictions for the gradient acceptability of individual sentences (Stabler, 2013; Sprouse et al., 2018).

Here, we suggest that a parser for Minimalist grammars (MGs; Stabler, 2013), coupled with complexity metrics measuring memory usage (Kobele et al., 2013; Graf et al., 2017, a.o.), is an effective model to address these issues. The MG parser has been used in the past to study which aspects of grammar drive processing cost for a vast set of off-line processing asymmetries cross-linguistically (Gerth, 2015; Graf et al., 2017; Zhang, 2017). Given the ability of MGs to encode rich syntactic analyses, the MG parser is especially sensitive to fine-grained grammatical information, and thus is able to generate quantitative predictions especially suited to our purposes.

In particular, we relate sentence acceptability to sentence structure by specifying: 1) a formalized theory of syntax in the form of MGs; 2) a parser as a model of how the structural representation of a

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 53-63.
New Orleans, Louisiana, January 2-5, 2020

sentence is built from its linear form; 3) a linking theory between structural complexity and acceptability in the form of metrics measuring memory usage. As a proof-of-concept for the validity of the linking theory, we model the acceptability judgments for three types of syntactic islands, using as a baseline the judgments reported in (Sprouse et al., 2012a).

Importantly, our main aim is not to settle the debate of whether gradience should be found in the grammar itself, or in the interaction between grammar and external factors (if such a debate could ever be settled). What we offer is a formalized, testable model of the latter hypothesis, in the hope of providing ground for a more principled investigation of categorical grammaticality and continuous acceptability.

2 MG Parsing

2.1 MGs

MGs (Stabler, 1997, 2011) are a lexicalized, mildly context-sensitive formalism incorporating the structurally rich analyses of Minimalist syntax — the most recent version of Chomsky's transformational grammar.

An MG grammar is a set of lexical items (LIs) consisting of a phonetic form and a finite, non-empty string of features. LIs are assembled via two feature checking operations: *Merge* and *Move*. Intuitively, Merge encodes subcategorization, while Move encodes long-distance movement dependencies. Here, we avoid most of the technical details of the formalism, and we limit our discussion to a general description of the data structures defined by these grammars.

MGs' *derivation trees* encode the sequence of Merge and Move operations required to build the phrase structure tree for a specific sentence (Michaelis, 1998; Harkema, 2001). In a traditional derivation tree, all leaf nodes are labeled by LIs, while unary and binary branching nodes are labeled as Move or Merge, respectively. However, as the details of the feature calculus are irrelevant to us, we adopt a simpler representation that discards the feature annotation of LIs, and labels internal nodes as standard in minimalist syntax. We also explicitly include dashed arrows indicating movement relations.[1]

[1]Note that, due to the fact that intermediate landing sites for moved phrases do not affect the traversal strategy, we do not explicitly highlight them with movement arrows.

The fundamental difference between a phrase structure tree and a derivation tree is that in the latter, moved phrases remain in their base position, and their landing site must be fully reconstructed via the feature calculus (cf. Fig. 1a and Fig. 1b). As a concequence, the final word order of a sentence is not directly reflected in the order of the leaf nodes in a derivation tree.

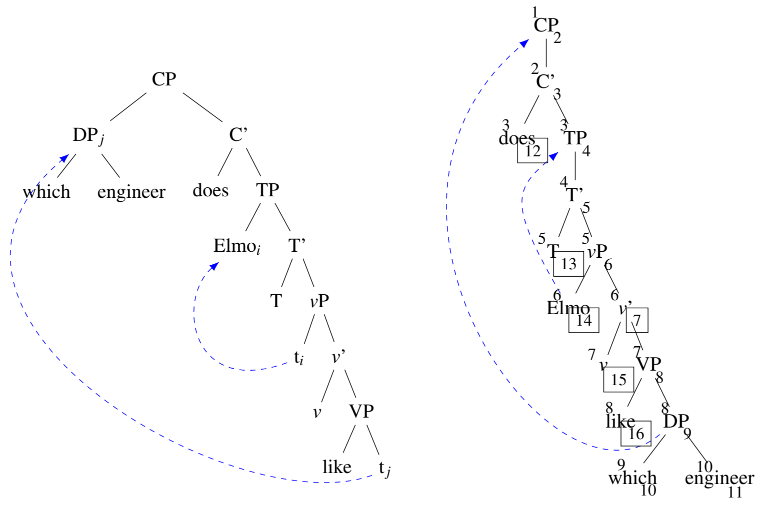
Importantly, MG derivation trees form a regular tree language, and thus can be regarded as a simple variant of context-free grammars (CFG), allowing us to exploit some of CFGs more established parsing algorithms.

2.2 Top-down MG Parsing

We follow recent sentence processing results, and adopt Stabler (2013)'s top-down parser for MGs. This parser is a variant of a standard depth-first, top-down parser for CFGs: it takes as input the string representation of a sentence, hypothesizes the structure top-down, verifies that the words in the structure match the input string, and outputs an encoding of the sentence structure in the form of a derivation tree. Importantly, the surface order of lexical items in the derivation tree is not the phrase structure tree's surface order. Thus, simple top-to-bottom and left-to-right scanning of the leaf nodes yields the wrong word order. While scanning the nodes then, the MG parser must also keep track of the derivational operations which affect the linear word order.

Memory plays a crucial role in this procedure: if a node is hypothesized at step i, but cannot be worked on until step j, it must be stored for $j-i$ steps in a priority queue. To make this traversal strategy transparent to the reader, we adopt Kobele et al. (2013)'s notation, in which each node in the tree is annotated with an *index* (superscript) and an *outdex* (subscript). Intuitively, the annotation indicates for each node in the tree when it is first conjectured by the parser (index) and placed in the memory queue, and at what point it is considered completed and flushed from memory (outdex). Consider the tree in Fig. 1b, explicitly annotated with the parsing steps. The node *does* is hypothesized at step 3. However, *which engineer* comes before it in the input, so *does* has to wait until step 12 to be flushed out of the queue.

Finally, note that Stabler's parser was originally given a search beam discarding the most unlikely predictions. Here though, we are not interested



(a) (b)

Figure 1: Phrase structure tree (a), and annotated MG derivation tree (b) for *Which engineer does Elmo like?*. Boxed nodes in (b) are those with tenure value greater than 2, following (Graf and Marcinek, 2014).

in the cost of choosing among alternative parsing choices, and want to focus on the specific contribution of the grammar to memory usage. Thus, we assume that the parser is equipped with a perfect oracle, which always makes the right choices when constructing a tree (Kobele et al., 2013). Essentially, the MG model employs a deterministic parsing strategy, where ambiguity has no role.

2.3 Measuring Memory Usage

Recently, Stabler (2013)'s MG parser has been used to investigate which aspect of grammatical structure affect off-line processing difficulty (Kobele et al., 2013; Graf and Marcinek, 2014; Gerth, 2015; Graf et al., 2017, a.o.).

In order to allow for psycholinguistic predictions, the behavior of the parser is related to processing difficulty via complexity metrics measuring how the structure of a tree affects memory. The MG model refers to three main notions of memory usage (Graf et al., 2017): (a) how long a node is kept in memory (*tenure*); (b) how many nodes must be kept in memory (*payload*); (c) how much information is stored in a node (*size*).

Tenure and payload for each node n in the tree can be easily computed via the node annotation scheme of Kobele *et al.*: a node's tenure is equal to the difference between its index and its outdex; the payload of a derivation tree is computed as the number of nodes with a tenure strictly greater than 2 (boxed nodes in our tree annotation scheme).[2] For instance, tenure for the node *does* in Fig. 1b is computed as $12-3=9$.

Defining size in an informal way is slightly trickier, as it was originally based on how information about movers is stored by Stabler's top-down parser (for a technical discussion, see Graf et al., 2015). In practice, size measures the hierarchical length of a movement dependency, and is computed as the index of a mover minus the index of its target site. Considering again the tree in Fig. 1b, the size of *Elmo* is $6-3=3$.

In order to contrast derivations, past work has used these general concepts to define a vast set of complexity metrics measuring processing difficulty over a full tree (Kobele et al., 2013). For instance, tenure can be associated to metrics like MAXT $:= max(\{tenure\text{-}of(n)\})$ and SUMT $:= \sum_n tenure\text{-}of(n)$. MAXT measures the maximum amount of time any node stays in memory during processing, while SUMT measures the overall amount of memory usage for all nodes whose tenure is not trivial. It thus captures total memory usage over the course of a parse. As an illustrative example, consider one last time the tree in Fig. 1b. Tenure in this tree is mostly driven by the movement of the embedded object, thus MAXT is mea-

[2] We refer to tenure values ≤ 2 as *trivial*, since it arises naturally from the binary nature of derivation trees, and it's not due to extra waiting time in the priority queue (Graf and Marcinek, 2014).

sured at *does* and it is equal to $12-3=9$. Similar metrics can be defined for size. For instance, in Fig. 1b SUMS is given by the length of the object movement and the length of the subject movement: $(8-1)+(6-3)=10$.

These metrics have been surprisingly successful in accounting for a vast array of different processing phenomena, such as right embedding vs. center embedding, nested dependencies vs. crossing dependencies, as well as a set of contrasts involving relative clauses (Graf and Marcinek, 2014; Graf et al., 2015). However, Graf et al. (2015) argue that a better approach would make use of ranked metrics of the type $\langle M_1, M_2, \ldots, M_n \rangle$. Such rankings work in a way similar to constraint ranking in Optimality Theory (Prince and Smolensky, 2008): a lower ranked metric matters only if all higher ranked metric have failed to pick out a unique winner (e.g., if two constructions result in a *tie* over MAXT). Following this idea, Graf et al. (2017) show that when complexity metrics are allowed to be ranked in such a way the space of possible metrics quickly explodes (up to 1600 distinct metrics). Considering the total number of possible metrics, it is conceivable that some metric combination could explain any hypothetical processing asymmetry — thus reducing the explanatory power of the model. However, this does not seem to be the case. Graf et al. (2017) rule out the vast majority of these metrics, by showing their insufficiency in accounting for some crucial constructions across a variety of grammatical analyses.

Here then, we rely on previous work and focus on the predictions made by a ranked version of $\langle$MAXT, SUMS$\rangle$ in comparing memory burden for contrasting sentences (Zhang, 2017; Liu, 2018; Lee, 2018; De Santo, 2019; De Santo and Shafiei, 2019). In addition, our core linking hypothesis connects processing difficulty to acceptability by assuming that higher memory cost implies lower acceptability.

3 Gradient Acceptability in Syntactic Islands

Given the metrics' sensitivity to minor differences in syntactic structure, the MG parser's predictions are the most interpretable when used to compare the relative complexity of minimally different sentences. Careful comparisons across sentences as similar as possible in their underlying syntactic structure seem also to be desirable if we want to understand the source of gradient variation in acceptability judgments. For these reasons, we chose to model the data on the acceptability of syntactic islands collected by Sprouse et al. (2012a) (henceforth SWP), in a first investigation of the viability of the parser as a model of gradient acceptability.

Syntactic islands are well-known in linguistics (Chomsky, 1965; Ross, 1968) as a set of phenomena in which the acceptability of a sentence is degraded, in relation to the interaction of a long-distance dependency and its syntactic context. Consider the following sentences:

(1) a. What$_i$ did John say Bill saw t_i ?
   b. What$_i$ did John have dinner before Bill saw t_i?

In 1a, *what* is displaced from its lower position as the object of the verb *saw* to a sentence initial position. In 1b, this same displacement cannot take place, as *what* is inside an adjunct clause (headed by *because*). Thus, 1b is considered ill-formed by native speakers of standard American English. Since displacing an element from inside an adjunct leads to ungrammaticality, adjunct clauses are classic example of island structures.

SWP conducted an extensive investigation of the acceptability of island constructions, by collecting formal acceptability judgments for four island types using a magnitude estimation task. The acceptability contrasts in this study are optimal for our purposes for multiple reasons. First, while a categorical grammar would predict a binary split in sentence acceptability (violates an island/doesn't violate an island), the continuous scale the estimation task was based upon revealed a spectrum of gradient judgments. Second, the stimuli in SWP's design were based on a (2×2) factorial definition of island effects, and explicitly identify two structural factors that might affect acceptability: 1) the length of a movement dependency; 2) the presence of a so-called "island construction" (Kluender and Kutas, 1993). This careful dimensional decomposition of the test sentences, coupled with the continuous scale of the judgment task, resulted in a set of well-defined pairwise comparisons ideal for the MG parser's modeling approach.

In what follows, we test whether the gradient of acceptability shown in SWP's data is predicted by a parser grounded in a rich categorical grammar. Before proceeding with our analysis though, it seems to be important to make an additional note about our aims. An expert reader might know that there is an ongoing debate in the literature about the nature of islands effects (see, for instance, Hofmeister et al., 2012a; Sprouse et al., 2012b; Hofmeister et al., 2012b, and references therein) — with classical syntactic accounts rooting them in grammatical constraints, while others arguing that such effects can be reduced to a conspiracy of processing factors.

Importantly, we are *not* attempting to reduce these effects to processing demands and, at least at this stage, it is not our purpose to directly engage with this debate. For the same reasons, we do not investigate the *super-additivity* found in SWP's paper, as we are *not* interested in modeling the grammaticality of an island violation per-se. Relatedly, we do not claim that the acceptability of island violations is *purely* syntactic in nature, as it has been shown to be sensitive to a variety of semantic factors (Truswell, 2011; Kush et al., 2018; Kohrt et al., 2018, a.o.). Crucially, we are "just" interested in exploring the idea that the *gradient* component of acceptability judgments arises due to processing factors. We focus on islands effects exclusively because of the optimal baseline offered by SWP's data.

We will return to the question of whether our model could give *any* insights into the question of separating processing and grammatical contributions to island effects in Sec. 5.

4 Modeling Results

SWP focused on English wh-movement dependencies to explore four types of islands constructions: Subject, Adjunct, Complex NP, and Whether islands. Since the MG parser is only sensitive to structural differences, in this paper we ignore the case of Whether islands and concentrate on the remaining three cases. Table 1 presents a summary of all modeling contrasts in the paper, compared with the experimental results of SWP.[3]

4.1 Subject Island: Case 1

First, we model Subject islands as in SWP's Experiment 1, comparing 4 sentence types across 2 conditions: subject/object extraction, and island/non-island. Note that here *island* does not imply a violation, but refers to the presence of an island structure (Kluender and Kutas, 1993).

[3] All scripts are available at https://github.com/CompLab-StonyBrook/mgproc.

Island Type	Sprouse et al. (2012)	MG Parser
Subject Island Case 1	2b > 2a	✓
	2b > 2d	✓
	2b > 2c	✓
	2a > 2c	✓
	2a > 2d	✓
	2c > 2d	2c < 2d
Subject Island Case 2	3a > 3b	✓
	3a > 3c	✓
	3a > 3d	✓
	3b > 3d	✓
	3c > 3b	✓
	3c > 3d	✓
Adjunct Island	4a > 4b	✓
	4a > 4c	✓
	4a > 4d	✓
	4b > 4d	✓
	4c > 4b	✓
	4c > 4d	✓
Complex NP Island	5a > 5b	✓
	5a = 5c	✓
	5a > 5d	✓
	5b > 5d	✓
	5c > 5b	✓
	5c > 5d	✓

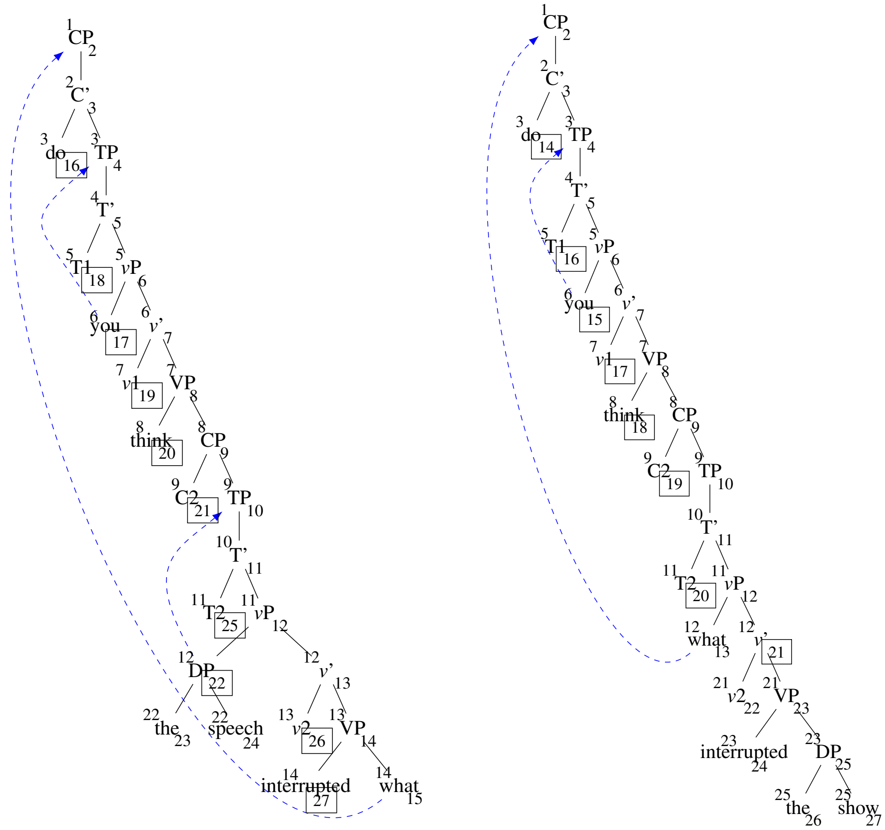
Table 1: Summary of results (as pairwise comparisons) from (Sprouse et al., 2012a), and corresponding parser's predictions (x > y: x more acceptable than y).

(2) a. What do you think the speech interrupted *t*? **Obj/Non Island**
b. What do you think *t* interrupted the show? **Subj/Non Island**
c. What do you think the speech about global warming interrupted the show about *t*? **Obj/Island**
d. What do you think the speech about *t* interrupted the show about global warming? **Subj/Island**

Annotated MG derivation trees for these sentences are shown in Fig. 2 (object/subject with no island) and Fig. 3 (with island).[4] The parser's predictions (via MAXT) overall match the experimental results (see Table 1).[5]

[4] Due to space constraints, annotated derivations are provided just for the Subject island case, as an illustrative example. Derivations for all other island types can be easily reconstructed from standard minimalist analyses of the test sentences (e.g., Adger, 2003). Source files can also be found at https://github.com/aniellodesanto/mgproc/tree/master/islands.

[5] When a wh-element is displaced from an embedded position, we avoid intermediate landing sites due to successive cyclicity. As intermediate movement steps do not affect the



(a) (b)

Figure 2: Annotated derivation trees for (a) 2a (object, non-island) and (b) 2b (subject, non-island).

The factorial design of the original study helps us understand the model's predictions. The contrast between 2b and 2a,2d is correctly captured by MAXT. This is due to the wh-element spanning a longer, more complex structure comprising the whole embedded DP subject in the Island cases. Compare 2a and 2b, both with highest tenure on *do* (14 and 11, respectively — cf. Tbl. 2). In 2a, *do* is conjectured after *what* has been scanned from the input. But then it cannot be flushed out of memory until *what* is confirmed in its base position as the embedded complement. In 2b, *do* only has to wait until the embedded subject position is reached, and then it is discarded from memory.

Consider now 2c. Here the highest tenure is on the embedded *T* head, which has to wait for the wh-element in object position, and then for the whole complex DP in subject position, before it can finally be flushed out of the queue. The longer wh-dependency in the object case explains once again why 2b is preferred over 2c, and the additional complexity of the DP subject is crucial in driving the 2b > 2c contrast.

Clause Type	Ex. #	MaxT	SumS
Obj./Non Island	2a	14/*do*	19
Subj./Non Island	2b	11/*do*	14
Obj./Island	2c	23/*T2*	22
Subj./Island	2d	15/*do*	20
Short/Non Island	3a	5/*C*	9
Long/Non Island	3b	11/*do*	14
Short/ Island	3c	11/*T2*	9
Long/ Island	3d	17/*T2*	20

Table 2: Summary of MAXT (*value*/*node*) and SUMS by test sentence for Subject island in case 1 and 2 (*T2* marks the embedded *T* head.)

Finally, there is one case in which parser's predictions and experimental data disagree: the contrast between subject and object extraction in the island condition (2c vs 2d). The parser predicts that 2c should be more acceptable than 2d (*Subj/Island* > *Obj/Island*). This is not surprising,

traversal strategy, this choice does not significantly change our results (cf. Zhang, 2017).

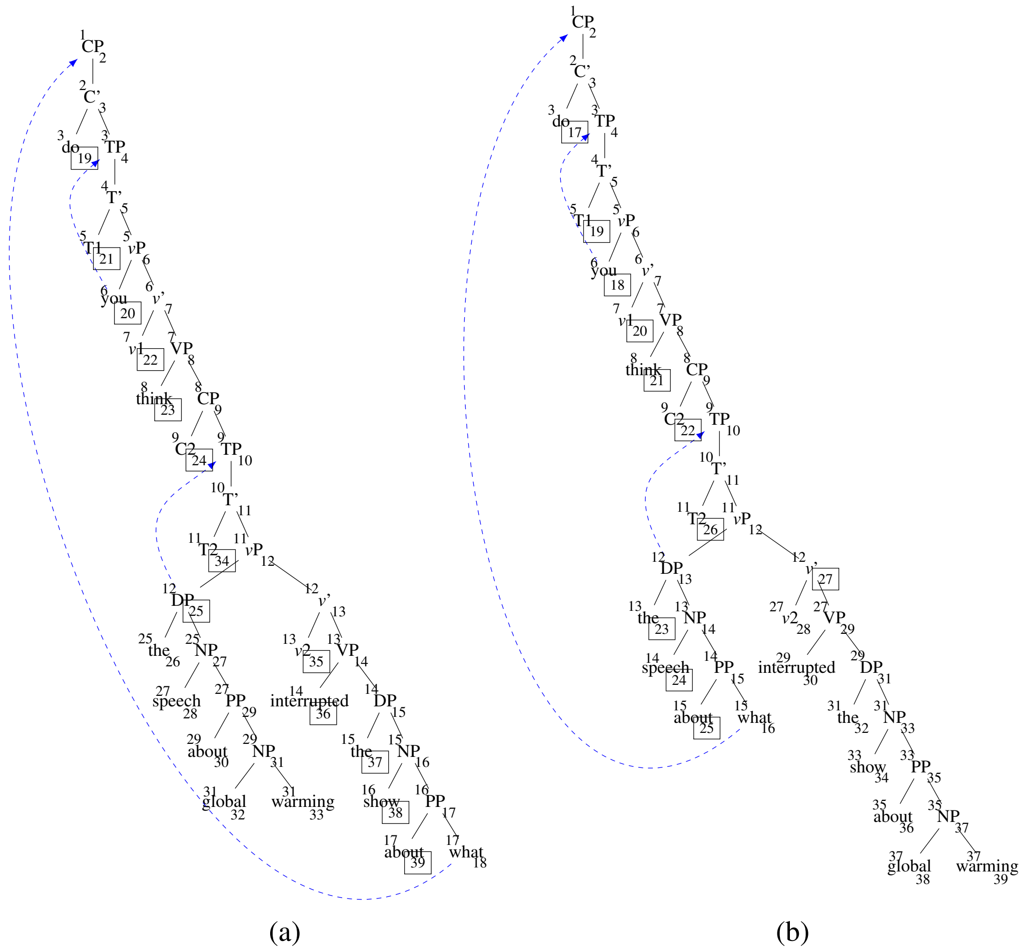


Figure 3: Annotated derivation trees for the test sentences in (a) 2c (object, island) and (b) 2d (subject, island).

as the memory metrics pick up on the additional length of the extraction in the object case, and thus obviously predict the preference for a subject gap. However, SWP show *Obj/Island* > *Subj/Island* — which is expected from a theoretical perspective since 2d is the ungrammatical condition (i.e., there is an extraction out of an island).

We will come back to the significance of this mismatch in Sec. 5. Crucially for our main claim though, the parser correctly predicts the gradient of acceptability for those conditions that, according to a categorical grammar, should all be equivalent (i.e., those containing no forbidden extraction).

4.2 Subject Island: Case 2

The previous section suggests that, when a grammatical violation coincides with processing factors (e.g., length of a dependency), parser and human judgments should match on all contrasts. Luckily, SWP offer us the chance to test such a prediction, with a second set of subject island sentences. SWP's Experiment 2 compares a *short* dependency and *long* dependency (*matrix* vs *embedded* extraction in the original paper), again in an island and non-island condition.

(3) a. Who *t* thinks the speech interrupted the primetime TV show? **Short/Non Island**

b. What do you think *t* interrupted the primetime TV show? **Long/Non Island**

c. Who *t* thinks the speech about global warming interrupted the primetime TV show? **Short/Island**

d. What do you think the speech about *t* interrupted the primetime TV show? **Long/Island**

As expected, parser's preferences and experimental data fully match in this case, as the ungrammatical condition (3d) is also the one in which the movement dependency is the longest. Here however, deriving the correct preferences requires the ranking of ⟨MAXT,SUMS⟩, instead of just MAXT alone (note also that SUMS by itself would not suffice, as it would not predict 3a > 3c, cf. Tbl. 2). Such a ranking also preserves the results in the previous section, which fully relied on MAXT. Interestingly, note how MAXT values for 3b (*Long/Non Island*) and 3c (*Short/ Island*) tie here, as the additional structural complexity of 3c does not interact with the main movement dependency (*who* raising from Spec,TP to Spec,CP). Moreover, the *Short/Non Island* (3a) and *Short/Island* (3c) conditions have very similar structures (with an extraction out of the main subject). Nonetheless, the memory metrics are able to capture subtle differences in the way the parser goes through the two sentences (arguably capturing the "island construction" cost of (Kluender and Kutas, 1993)).

4.3 Adjunct and Complex NP Islands

So far, we have been successful in replicating SWP's acceptability judgments via the MG parser. However, we might wonder whether this success is due to something peculiar in the way the Subject island test cases interact with the MG parsing strategy. Thus, we tested the MG parser on Adjunct and Complex NP islands, again using as a baseline the results in SWP's Experiment 1. The test sentences for the adjunct case were as follows:

(4) a. Who *t* thinks that John left his briefcase at the office? **Short/Non Island**
b. What do you think that John left *t* at the office? **Long/Non Island**
c. Who *t* laughs if John leaves his briefcase at the office? **Short/Island**
d. What do you laugh if John leaves *t* at the office? **Long/Island**

As for Subject islands in case 2, ⟨MAXT,SUMS⟩ correctly predicts the pattern of acceptability reported by SWP, matching the empirical results across all conditions (cf. Tbl. 1). Similar results are obtained for the Complex NP island, with test sentences as follows:

(5) a. Who *t* claimed that John bought a car? **Short/Non Island**
b. What did you claim that John bought *t*? **Long/Non Island**
c. Who *t* made the claim that John bought a car? **Short/Island**
d. What did you make the claim that John bought *t*? **Long/Island**

Clause Type	Ex. #	MaxT	SumS
Short/Non Island	4a	13/*PP*	10
Long/Non Island	4b	17/*PP*	18
Short/Island	4c	13/*PP*	11
Long/Island	4d	21/*PP*	28
Short/Non Island	5a	5/*C*	9
Long/Non Island	5b	13/*did*	19
Short/Island	5c	5/*C*	9
Long/Island	5d	15/*did*	21

Table 3: Adjunct Island and Complex NP Island: MAXT (*value/node*) and SUMS values by test sentence.

Once more, the parser matches the acceptability preferences reported in SPW correctly in all conditions. Particularly interesting is the absence of a contrast between 4a and 4c. This is again due to the absence of a real interaction between the additional structural complexity of the island and the main movement dependency. The fact that this results in a tie stresses how movement dependencies and structural complexity conspire with the top-down strategy of the MG parser in non-trivial ways to drive memory cost.

5 Discussion

This paper argues for an MG parser as a good, non probabilistic formal model of how gradient acceptability can be derived from categorical grammars. In doing so, we provide one of the first quantitative models of how processing factors and fine-grained, minimalist-like grammatical information can conspire to modulate acceptability. As a proof-of-concept, we replicated the gradient acceptability scores for the island effects in (Sprouse et al., 2012a). These results are certainly preliminary, but the success of the parser on this baseline is encouraging.

As mentioned in the Introduction, many hypotheses have been formulated in the past about the way memory and grammatical factors conspire to produce processing differences across sentences. Thus, it is reasonable to wonder what are the benefits of the particular linking hypothesis implemented here. As we pointed out before, one of the main advantages of our model is the tight connection between the parser behavior and the rich grammatical information encoded in the MG derivation trees. This allows for rigorous evaluations of the cognitive claims made by modern syntactic theories.

In line with recent work using the MG parser as a model of processing difficulty, Section 4 focused on the predictions made by MAXT and SUMS, Clearly, one could easily conceive of metrics that take different syntactic information into account (for example, by counting the amount of bounding nodes or phases). However, tenure and size arguably rely on the simplest possible connection between memory, structure, and parsing behavior — as they exclusively refer to the geometry of a derivation tree, without additional assumptions about the nature of its nodes.

Of course, a question remains about the cognitive plausibility of such metrics. While this model is certainly not the first to formalize memory cost as associated to the length of movement dependencies, the previous discussion highlighted how size-centered metrics do not simply depend on the length of a movement steps. Instead, they pick up on the non-trivial changes in the behavior of the parser, based on how long-distance dependencies interact with local structural configurations. Thus, they cannot trivially be identified with other length-based measures (cf. Gibson, 1998; Rambow and Joshi, 2015, a.o.). As previous work points out, in the future it will be important to explore the relation between these complexity metrics, and psychological insights about the nature of human memory mechanisms (De Santo, 2019).

Similarly, as one reviewer suggests, it would be interesting to see whether SPW's results can be derived from different cognitive hypotheses; for instance by implementing in the MG model the variety of constraints explored by Boston (2012) for a dependency parser. Moreover, in this study we employ a deterministic parser to exclusively focus on the relation between structural complexity and memory usage. However, it is known that structural and lexical frequency influence islands' acceptability (Chaves and Dery, 2019, a.o.). Thus, informative insights would come from implementing information-theoretical complexity metrics over the MG parser (Hale, 2016), and explore the predictions of expectation-based approaches.

Obviously, the target judgments modeled here are part of a restricted set. Future studies in this sense will benefit from wider comparisons among minimally different variants of acceptable and unacceptable sentences (cf. Sprouse et al., 2013, 2016). As mentioned, the nature of the model makes comparisons beyond pairs of minimal sentences hard to interpret. However, in future it might be possible to define normalization measures for memory metrics computed over sentences with widely different underlying structures.

Finally, in Section 3 we avoided discussing the nature of island effects, as we do not mean for the MG model to address the debate of whether island violations are reducible to processing factors, or are instead tied to core grammatical constraints. Importantly, while this approach might superficially be construed as a reductionist theory, it is not: for instance, the MG parser by itself is not able to explain the difference between sentences that are simply hard to process, and sentences considered unacceptable/ungrammatical. Thus, the model is theoretically neutral with respect to grammatical or reductionist frameworks.

However, consider the first case of Subject islands we analyzed in Sec. 4. The parser produced the right predictions for all test sentences except when, in the presence of an island construction, the longest movement dependency and the island violation did not coincide (2c and 2d). This mismatch is not only explained, but it is actually expected, if we embrace a grammatical theory of island constraints. Under such theory, 2d is preferable from a processing perspective (as it involves shorter dependencies), but its acceptability is lowered by the fact that it violates a grammatical constraint, while 2c does not.

While we have to be careful in formulating hypotheses based on a single data point, this contrast suggests that the MG model could help us investigate those aspects of acceptability that are fundamentally tied to grammatical constraints.

Acknowledgements

Sincere thanks to Thomas Graf, Jon Sprouse, and three anonymous reviewers for insightful feedback on this work.

References

David Adger. 2003. *Core syntax: A minimalist approach*, volume 33. Oxford: Oxford University Press.

Marisa Ferrara Boston. 2010. The role of memory in superiority violation gradience. In *Proceedings* *of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–44. Association for Computational Linguistics.

Marisa Ferrara Boston. 2012. *A computational model of cognitive constraints in syntactic locality*. Ph.D. thesis, Cornell University.

Rui P Chaves and Jeruen E Dery. 2019. Frequency effects in subject islands. *Journal of Linguistics*, 55(3):475–521.

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2:113–124.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, volume 11. MIT Press.

Matthew W Crocker and Frank Keller. 2005. Probabilistic grammars as models of gradience in language processing. In *Gradience in grammar: Generative perspectives*. Citeseer.

Aniello De Santo. 2019. Testing a Minimalist grammar parser on Italian relative clause asymmetries. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL) 2019*, June 6 2019, Minneapolis, Minnesota.

Aniello De Santo and Nazila Shafiei. 2019. On the structure of relative clauses in Persian: Evidence from computational modeling and processing effects. Talk at the 2nd North American Conference in Iranian Linguistics (NACIL2), April 19-21 2019, University of Arizona.

Sabrina Gerth. 2015. *Memory Limitations in Sentence Comprehension: A Structural-based Complexity Metric of Processing Difficulty*, volume 6. Universitätsverlag Potsdam.

E Gibson. 2000. The dependency locality theory: a distance-based theory of linguistic complexity. In *2000, Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT press.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Thomas Graf, Brigitta Fodor, James Monette, Gianpaul Rachiele, Aunika Warren, and Chong Zhang. 2015. A refined notion of memory usage for minimalist parsing. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 1–14, Chicago, USA. Association for Computational Linguistics.

Thomas Graf and Bradley Marcinek. 2014. Evaluating evaluation metrics for minimalist parsing. In *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 28–36.

Thomas Graf, James Monette, and Chong Zhang. 2017. Relative clauses as a benchmark for Minimalist parsing. *Journal of Language Modelling*, 5:57–106.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Henk Harkema. 2001. A characterization of minimalist languages. In *International Conference on Logical Aspects of Computational Linguistics*, pages 193–211. Springer.

Philip Hofmeister, Laura Staum Casasanto, and Ivan A Sag. 2012a. How do individual cognitive differences relate to acceptability judgments? A reply to Sprouse, Wagers, and Phillips. *Language*, pages 390–400.

Philip Hofmeister, Laura Staum Casasanto, and Ivan A Sag. 2012b. Misapplying working-memory tests: A reductio ad absurdum. *Language*, 88(2):408–409.

Frank Keller. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Ph.D. thesis.

Robert Kluender and Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and cognitive processes*, 8(4):573–633.

Gregory M Kobele, Sabrina Gerth, and John Hale. 2013. Memory resource allocation in top-down minimalist parsing. In *Formal Grammar*, pages 32–51. Springer.

Annika Kohrt, Trey Sorensen, and Dustin A Chacón. 2018. The real-time status of semantic exceptions to the adjunct island constraint. In *Proceedings of WECOL 2018: Western Conference on Linguistics*.

Dave Kush, Terje Lohndal, and Jon Sprouse. 2018. Investigating variation in island effects. *Natural language & linguistic theory*, 36(3):743–779.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1618–1628.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

So Young Lee. 2018. A minimalist parsing account of attachment ambiguity in English and Korean. *Journal of Cognitive Science*, 19(3):291–329.

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3):375–419.

Lei Liu. 2018. Minimalist Parsing of Heavy NP Shift. In *Proceedings of PACLIC 32 The 32nd Pacific Asia Conference on Language, Information and Computation*, The Hong Kong Polytechnic University, Hong Kong SAR.

Brian McElree, Stephani Foraker, and Lisbeth Dyer. 2003. Memory structures that subserve sentence comprehension. *Journal of memory and language*, 48(1):67–91.

Jens Michaelis. 1998. Derivational minimalism is mildly context–sensitive. In *International Conference on Logical Aspects of Computational Linguistics*, pages 179–198. Springer.

Alan Prince and Paul Smolensky. 2008. *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Owen Rambow and Aravind K Joshi. 2015. A processing model for free word-order languages. *Perspectives on sentence processing*.

Luigi Rizzi. 1990. *Relativized minimality.* The MIT Press.

J. R. Ross. 1968. Constraints on variables in syntax. *Ph.D. dissertation, MIT*.

Antonella Sorace and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115(11):1497–1524.

Jon Sprouse. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, 1:123–134.

Jon Sprouse, Ivano Caponigro, Ciro Greco, and Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in english and italian. *Natural Language & Linguistic Theory*, 34(1):307–344.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Jon Sprouse, Matt Wagers, and Colin Phillips. 2012a. A test of the relation between working-memory capacity and syntactic island effects. *Language*, 88(1):82–123.

Jon Sprouse, Matt Wagers, and Colin Phillips. 2012b. Working-memory capacity and island effects: A reminder of the issues and the facts. *Language*, 88(2):401–407.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C Berwick. 2018. Colorless green ideas do sleep furiously: Gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Edward P Stabler. 1997. Derivational minimalism. In *International Conference on Logical Aspects of Computational Linguistics*, pages 68–95. Springer.

Edward P Stabler. 2011. Computational perspectives on minimalism. In *The Oxford Handbook of Linguistic Minimalism*.

Edward P Stabler. 2013. Two models of minimalist, incremental syntactic analysis. *Topics in cognitive science*, 5(3):611–633.

Robert Truswell. 2011. *Events, phrases, and questions*. 33. Oxford University Press.

Heinz Wanner and Michael P. Maratsos. 1978. An ATN approach to comprehension. In *Linguistic theory and psychological reality*. MIT Press.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Chong Zhang. 2017. *Stacked Relatives: Their Structure, Processing and Computation*. Ph.D. thesis, State University of New York at Stony Brook.