Understanding acceptability judgments: Additivity and working memory effects

Laura Staum Casasanto
Max Planck Institute for Psycholinguistics
Nijmegen, NL

Philip Hofmeister
Center for Research in Language
University of California - San Diego
La Jolla, CA

Ivan A. Sag
Department of Linguistics
Stanford University
Stanford, CA

Abstract

Linguists build theories of grammar based largely on acceptability contrasts. But these contrasts can reflect grammatical constraints and/or constraints on language processing. How can theorists determine the extent to which the acceptability of an utterance depends on functional constraints? In a series of acceptability experiments, we consider two factors that might indicate processing contributions to acceptability contrasts: (1) the way constraints combine (i.e., additively or super-additively), and (2) the way a comprehender's working memory resources influence acceptability judgments. Results suggest that multiple sources of processing difficulty combine to produce super-additive effects, but multiple grammatical violations do not. Furthermore, when acceptability judgments improve with higher working memory scores, this appears to be due to functional constraints. We conclude that tests of (super)-additivity and of differences in working memory can help to identify the effects of processing difficulty (due to functional constraints).

Introduction

Grammatical theories are designed to reflect, explain, and predict what is and is not possible to say in a language. Potential utterances are usually classified as "possible" or "impossible" on the basis of native speaker judgments of their acceptability. Whether they are the judgments of theorists themselves or of a sample of naive speakers, these judgments are not a perfect window into the speaker's grammatical competence: the judgments themselves are colored by performance factors. This problem has been discussed since Miller and Chomsky (1963) pointed out that some sentences that native speakers judge to be unacceptable, such as triple center embeddings (1), are better ruled out by their extreme difficulty than by grammatical constraints.

(1) The salmon that the man that the dog chased smoked fell.

Miller and Chomsky's assessment that functional constraints on the language processing system underlie the unacceptability of these examples is fairly uncontroversial. However, it is often difficult to determine what role functional constraints might play in other acceptability contrasts. In the domain of island violations, for example, both processing and grammatical constraints have been proposed to account for the unacceptability of island-violating sentences (Ross, 1967; Chomsky, 1973, 1986; Kluender, 1998, inter alia).

Assessing whether functional constraints underlie acceptability contrasts may be difficult, but it is critical in determining which acceptability contrasts should be taken as evidence for the existence of grammatical constraints. But what tools do we have for recognizing when acceptability contrasts are a consequence of functional constraints? This paper will explore two properties of processing constraints that could help theorists to recognize their effects on acceptability. First, individuals have a limited set of cognitive resources that they can use to understand language (Just & Carpenter, 1992; Kluender, 1998; Cowan, 2001). Extreme sentence processing difficulty can exhaust these resources, resulting in a strong perception of unacceptability, as in (1). Second, the extent of this limited pool of resources arguably varies from one individual to another, as suggested by Just and Carpenter (1992).

To explore the first property, we will consider what happens when multiple possible sources of unacceptability are combined. There are three logically possible outcomes of combining two manipulations that each individually cause acceptability decrements: a significantly smaller penalty than the sum of the two individual penalties (a result which we will refer to as *sub-additive*), a penalty that is statistically indistinguishable from the sum of the two individual penalties (which we will refer to as *additive*), or a penalty that is significantly larger than the sum of the two individual penalties (which we will call *super-additive*).

Super-additive effects may result from combining two manipulations that tax the same set of limited resources, if the manipulations are sufficiently strong to deplete the available resources. Thus, super-additivity could result when multiple sources of processing difficulty co-occur, depending on the degree of difficulty. On the other

hand, formal (grammatical) violations are not expected to combine super-additively. At least for theories that distinguish between formal and functional constraints, grammaticality violations do not cause decrements due to the taxing of a limited set of resources but to the violation of grammatical rules. They could combine additively – if each violation influences judgments independent of the other – or sub-additively, if one violation overwhelms the other or the overall acceptability of the sentence depends simply on the most egregious violation. When formal and functional sources of unacceptability appear in the same sentence, either additive or sub-additive decrements could be the result, for the same reasons, but again, this combination should not produce a super-additive decrement.

To explore the second property, we will compare the performance of people with different processing resources on acceptability judgment tasks. If an acceptability contrast reflects overtaxing the resources of comprehenders, then comprehenders with greater processing resources should experience less difficulty and the contrast should be reduced. However, when acceptability contrasts are due to grammaticality violations, comprehenders with greater processing resources should, if anything, show enhanced contrasts, because they are better able to parse the sentences and notice the rule violations.

In order to evaluate these predictions, we conducted three acceptability judgment studies, combining two processing manipulations (Experiment I), two grammaticality violations (Experiment II), and a processing manipulation with a grammaticality violation (Experiment III).

Experiment I: Processing Difficulty

To investigate the role of processing complexity in acceptability judgments, we manipulated the distance between two dependent arguments and their syntactic head.

Participants Stanford University students (n=32) participated in exchange for payment. All self-identified as native speakers of English.

Materials Twenty-four items were selected from Grodner and Gibson (2005). In these items, the hierarchical distance between a subject and object noun phrase and their subcategorizing verb was varied. This was achieved by varying (1) the presence/absence of a relative clause between the subject and verb [2a,2c vs. 2b,2d] and (2) positioning the object NP immediately after the verb or before the subject NP by relativizing it:

(2) a. The nurse from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room. [**short-short**]

b. The nurse who was from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room. [**long-short**]

c. The administrator who the nurse from the clinic supervised scolded the medic while a patient was brought into the emergency room. [**short-long**]

d. The administrator who the nurse who was from the clinic supervised scolded the medic while a patient was brought into the emergency room. [**long-long**]

These items were selected because reading time evidence from Grodner & Gibson (2005) show that increasing the hierarchical distance in examples like these leads to slower processing at the critical integration sites. The 24 experimental items appeared with 72 fillers (24 of which were the items from Experiment III).

Procedure To acquire acceptability ratings from participants, we used the thermometer judgment methodology described in (Featherston, 2008). This paradigm resembles the Magnitude Estimation (ME) technique of gathering judgments (Bard, Robertson, & Sorace, 1996; Sorace & Keller, 2005), where participants are asked to rate the magnitude of acceptability difference between test items and a reference sentence (e.g. twice as good, three times as good, half as good, etc.). In both ME and thermometer judgment experiments, participants are not limited to a particular set of values that they can assign to sentences - in principle, every sentence could receive a different judgment.

There are, however, several key differences between the ME and thermometer methods of judgment collection. In the latter paradigm, participants are not instructed to evaluate test items in terms of the magnitude of acceptability compared to the reference item, as evidence shows that participants ignore these instructions and rate sentences in terms of their linear distance from the reference. In addition, in thermometer judgment studies, participants judge items relative to two reference sentences. One of these references is quite good and the other quite bad, and we follow Featherston (2008) in assigning these sentences the arbitrary values 20 and 30. For all of our experiments, we used the same reference sentences.

(3) a. The way that the project was approaching to the deadline everyone wondered. = 20

b. The architect told his assistant to bring the new plans to the foreman's office. = 30

While participants could theoretically assign any real number value to the test items, including negative or decimal values, participants almost always assign positive integer values, typically between 10 and 40.

Sentences were presented to participants on a computer screen one word at a time for a fixed duration, via the DMDX software package (Forster & Forster, 2003). The duration that each word stayed on the screen varied with the number of characters in the word (250 ms + 33.34 * number of characters), so that longer words remained visible for longer periods. We chose word-by-word presentation over full sentence presentation to prevent participants from excessive introspection about the test sentences, and we used auto-paced presentation

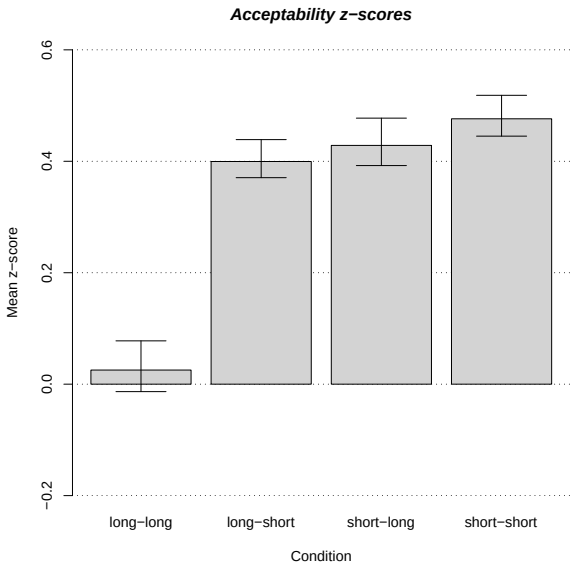

Figure 1: Acceptability z-scores for experiment I. Error bars show (+/−) one standard error.

rather than self-paced presentation to prevent differences in how long each participant studied a given stimulus.

Each participant also completed a reading span task during the same session, used to assess their working memory capacity (Daneman & Carpenter, 1980). For the analysis of reading span scores, we scored each test using using the partial credit method outlined in Conway et al. (2005): successful recall of a word in a study list counts toward the final reading span score, even if the entire item set was not recalled correctly.

Results

Prior to statistical analysis, we log-transformed judgment ratings to normalize the data and to reduce the effect of extreme data points. Subsequently, we computed z-scores for each subject on the basis of all data in the experimental data set (except practice items), including fillers. This reduces the impact of varying uses of the interval scale by subjects. Finally, we excluded data points with z-scores more than 2.5 standard deviations from the mean for each participant. For Experiment I, this outlier removal process affected 2.0% of the data. The resulting z-scores constitute the data on which we conducted statistical analyses.

For all experiments, we used linear mixed effects models to estimate the effects of the experimental manipulations. Such statistical analyses remove the need for prior averaging over subjects and items, are more robust in the presence of missing data, and do not require the assumptions of sphericity that are inherent to analyses such as repeated measures ANOVAs (Baayen, 2004, 2007). This method of statistical analysis also allows for the evaluation of additional factors such as list position alongside effects due to experimental manipulation. Prior to analysis, all predictors were centered—higher order variables (interactions) were also based on these centered predictors. Linear mixed effects models do not directly yield p-values (due to complications in estimating the degrees of freedom), but Monte Carlo Markov Chain (MCMC) sampling can be used to conservatively estimate p-values. For all p-values reported here, we utilized 25,000 MCMC samples to estimate the values.

The acceptability judgment results demonstrate main effects of both manipulations—-subject distance ($\beta = -.242$, t $= -6.412$, $p < .0001$) and object distance ($\beta = -.211$, t $= -5.584$, $p < .0001$). In addition, there was a highly significant interaction between these factors ($\beta = -.328$, t $= -4.343$, $p < .001$). This interaction reflects the result of combining multiple processing difficulties: the acceptability decrement produced by two processing manipulations was more than expected on the basis of the decrements produced by each manipulation in isolation.

Reading span score was also a highly significant predictor of acceptability scores ($\beta = .050$, t $= 3.685$, $p < .001$). In particular, higher reading span scores predicted higher judgments of acceptability. This effect appears to be largely driven by the conditions with multiple processing manipulations and a dislocated object phrase (the most difficult conditions according to Grodner and Gibson (2005)), which is reflected by the significant interaction of reading span score and the object distance manipulation ($\beta = .068$, t $= 3.858$, $p < .01$).

Discussion

According to the results, while these kinds of processing manipulations may have only minor effects on acceptability in isolation, they can have highly significant effects on judgments when combined. In this study, increasing the distance between a single dependent argument and its head only slightly lowered judgments. But when we increased the hierarchical distance of both dependents to their syntactic head, a sharp drop in acceptability judgments occurred. Consequently, these results indicate a super-additive effect on judgments resulting from the co-occurrence of multiple sources of difficulty in sentence processing.

In addition, estimates of working memory (operationalized as performance on the reading span test) indicate that better working memory predicts higher judgments of acceptability for items with processing challenges. This suggests that a positive linear relationship between reading span scores and acceptability scores indicates significant processing difficulty in the test items. The strength of these conclusions, however, depends on whether a similar relationship appears in sentences with grammatical violations.

Experiment II: Grammatical Violations

Experiment II evaluates how multiple grammatical violations affect judgments when they co-occur in the same sentence. Since grammatical violations do not affect acceptability via overtaxing processing resources, combining them should not result in super-additive decre-

ments (unlike the processing manipulations in Experiment I). In addition, for the same reason, comprehenders with greater working memory capacities should if anything show a greater decrement for grammatical violations than low-capacity comprehenders (also unlike the results of Experiment I).

Participants Stanford University students (n = 28) who had not participated in Experiment I completed this experiment in exchange for payment.

Materials The 24 experimental items in Experiment II contained either 0, 1, or 2 grammaticality violations. We manipulated the grammaticality of two separate but nearby constituents to yield a 2 x 2 design. The first manipulation targeted the morphological form of a verb in a subject relative clause. Subjects either saw the correct form (4a,4b) or they saw a form that was missing the appropriate inflectional morphology (4c, 4d). Additionally, participants either read an object pronoun with the proper case-marking (4b,4d) or they read a pronoun with unlicensed nominative case-marking (4a,4c):

(4) a. The friend who **visited** Sue asked **she** whether the value of the house had dropped since the recession began. [**good-bad**]

b. The friend who **visited** Sue asked **her** whether the value of the house had dropped since the recession began. [**good-good**]

c. The friend who **visit** Sue asked **she** whether the value of the house had dropped since the recession began. [**bad-bad**]

d. The friend who **visit** Sue asked **her** whether the value of the house had dropped since the recession began. [**bad-good**]

72 filler items appeared along with the critical items.

Procedure Procedure was identical to Experiment I.

Results

Data were analyzed using the same methods as in Experiment 1. Outlier removal affected 1.2% of the data. The acceptability results indicate that the manipulations of both inflectional morphology ($\beta = -.415$, $t = -8.525$, $p < .0001$) and case ($\beta = -.624$, $t = -12.817$, $p < .0001$) had significant effects on acceptability judgments. There was also a statistically significant interaction ($\beta = .234$, $t = 2.402$, $p < .05$); however, the interaction differs from the interaction found in Experiment I. Here, it emerges because the case error produces lower judgments than the verbal inflection error. This interaction is *not* due to super-additivity, as it was in Experiment I; two errors yield acceptability decrements that are approximately the sum of decrements caused by sentences with each error in isolation.

In further contrast with the results from Experiment I, reading span scores do not show an overall significant linear relationship with acceptability z-scores. For the conditions judged the worst by participants, memory estimates actually exhibit a negative linear relationship with z-scores, i.e. individuals with higher reading span scores judged these conditions worse, compared to individuals with lower reading span scores. The difference between the conditions leads to a statistically reliable interaction of reading span score and the effect of the case manipulation ($\beta = -.098$, $t = -3.028$, $p < .01$).

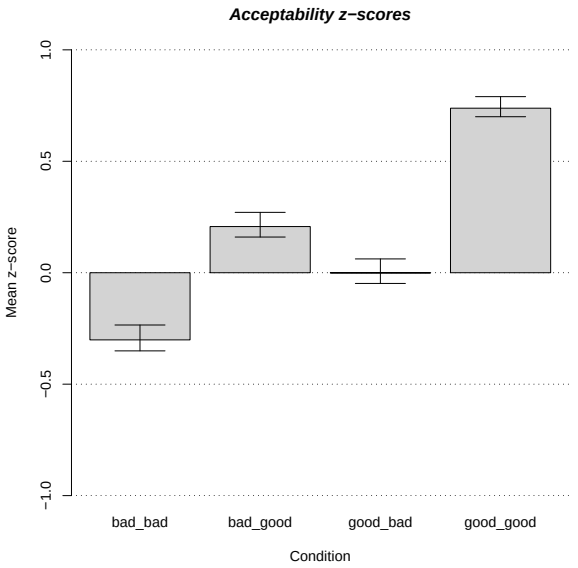

Figure 2: Acceptability z-scores for Experiment II. Error bars show (+/−) one standard error.

Discussion

Grammaticality violations appear to affect acceptability judgments in a qualitatively different way than processing manipulations. Most notably, grammaticality violations in this experiment combine additively—the effect of two co-occuring, proximal violations does not reduce judgments further than expected on the basis of each violation in isolation. These results align with independent evidence from Sorace and Keller (2005) that grammaticality violations combine additively.

The other important contrast between the first two experiments involves the relationship between reading span scores and acceptability scores. While we found a positive linear relationship between the two in Experiment I, in this experiment, reading span predicted *lower* judgments for the conditions judged worse (those with a case error). However, because the two types of manipulations were investigated in separate experiments, these high- and low-reading span participants were different individuals across experiments. In Experiment III, we directly compared the effects of grammaticality manipulations and processing manipulations in the same experiment and the same individuals.

Experiment III: Grammar and Processing

Participants This experiment was conducted in the same session as Experiment I, and involved the same 32

Stanford University students.

Materials Experiment III investigated how grammaticality violations and processing manipulations interact with one another. Experimental items appeared with either a correctly inflected verb (5a,5b) or incorrectly inflected verb (5c,5d). Dependency locality was utilized again to vary processing difficulty; the *wh*-dependencies in (5b) & (5d) are shorter than those in (5a) & (5c).

(5) a. They couldn't remember which lawyer that the reporter interviewed had defended the elderly man at the courthouse. **[hard-good]**

b. They couldn't remember which lawyer had defended the elderly man that the reporter interviewed at the courthouse. **[easy-good]**

c. They couldn't remember which lawyer that the reporter interviewed had defending the elderly man at the courthouse. **[hard-bad]**

d. They couldn't remember which lawyer had defending the elderly man that the reporter interviewed at the courthouse. **[easy-bad]**

The 24 experimental items were included alongside the materials from Experiment I and 48 additional fillers.

Procedure Procedure was identical to Experiments I and II.

Results

Data were analyzed using the same methods used in Experiments I and II. Removal of outliers affect 1.2% of the dataset. Results show that grammaticality significantly influences acceptability judgments ($\beta = .626$, $t = 15.583$, $p < .0001$). In contrast, the effect of processing difficulty on judgments is not statistically significant ($\beta = -.065$, $t = -1.628$, $p > .1$); however, there is a significant interaction between processing difficulty and grammaticality ($\beta = -.215$, $t = -2.680$, $p < .05$). As Figure 3 illustrates, this interaction arises because processing difficulty lowers judgments in sentences without grammatical violations, but it does not do so in sentences with grammatical violations.

While reading span does not emerge as a significant predictor for judgments across all condition types ($\beta = -.014$, $t = -.868$, $p > .1$), this seems to be because the grammatical and ungrammatical conditions pattern in different ways. The data reveal that individuals with higher reading span scores judge ungrammatical items worse, but in the grammatical conditions, better reading span performance predicts higher judgments of acceptability, leading to a significant interaction of reading span score and grammaticality ($\beta = .093$, $t = 4.986$, $p < .001$). In other words, estimates of memory capacity only show a positive linear relationship with judgments in the absence of grammar-based constraint violations.

Discussion

The results of Experiment III are consistent with our predictions and with the results of the first two experiments. Processing constraints and grammatical constraints combine sub-additively. Presumably, the grammaticality violations were so extreme as to "drown out" the effects of the processing manipulations in the ungrammatical conditions, i.e. a floor effect occurred. In general, if processing constraints and grammatical constraints reflect distinct and largely independent cognitive resources, super-additive combinations are unexpected. The present results support this hypothesis.

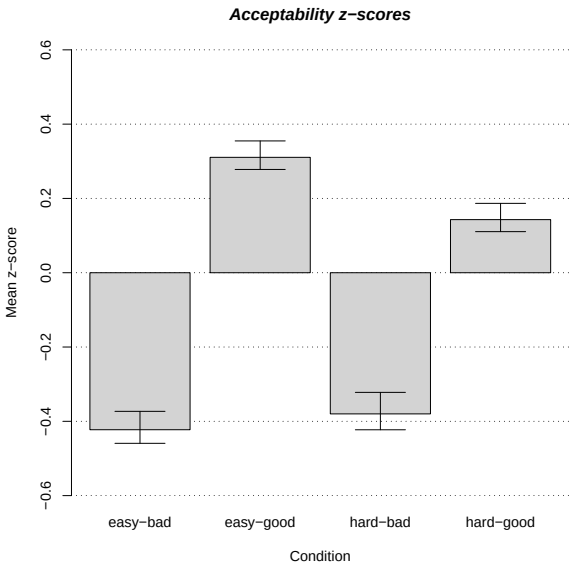

Figure 3: Acceptability z-scores for Experiment III. Error bars show (+/−) one standard error.

In addition, the relationships between reading span scores and judgments are as expected based on Experiments I and II: comprehenders with higher working memory scores find ungrammatical sentences worse, but difficult sentences better, compared to their low working memory counterparts. Experiment III shows that these contrasts can be observed even with the same set of subjects and minimally different items.

General Discussion and Conclusions

Three word-by-word acceptability judgment studies showed that (1) grammaticality violations combine additively, (2) differences that stem from functional constraints can combine super-additively with one another, and (3) grammaticality violations and processing manipulations can combine sub-additively with one another. These patterns suggest that when two constraints combine super-additively in acceptability decrements, it is likely that they are both functional constraints. Furthermore, participants' reading span scores predict sentence judgments differently for different types of manipulations. Participants with higher reading spans tend to judge ungrammatical sentences as being worse than their low-span counterparts do, yet they tend to judge difficult sentences as being better than participants with lower reading spans.

It might be tempting to extend the findings of these experiments to the inverses of the relationships we have reported here; that is, if super-additivity indicates a processing contribution to an acceptability decrement, then does the absence of super-additivity rule out processing contributions? While it would be helpful for interpreting acceptability judgments if this were true, neither the present results nor general principles of language processing license this inference. It two sources of processing difficulty are sufficiently weak, they will not overtax the available resources, and should combine additively. Likewise, if two sources of processing difficulty were sufficiently extreme, the presence of just one might so overwhelm processing resources that the presence of the other was undetectable, resulting in a sub-additive combination.

It is also not possible to infer the inverse of the relationship we reported between reading span scores and processing difficulty. The lack of a positive linear relationship between reading span scores and acceptability judgments does not entail that the sentences do not cause processing difficulty. Further experiments not reported here involving center-embeddings show that reading span scores do not exhibit a positive linear relationship with acceptability judgments in the presence of massive processing difficulty. This could occur if language comprehension is not likely at a certain level of difficulty, and thus having greater language comprehension abilities might not produce better judgments. In other words, some stimuli may be so hard to process that virtually no one will have sufficient cognitive resources to understand the stimuli.

Given that these inverse inferences are not supported, tests of the functional origins of acceptability contrasts that seek to take advantage of the relationships of super-additivity and working memory capacity we demonstrate here must be designed accordingly. When super-additivity and/or positive linear relationships between acceptability and working memory measures are observed, however, these relationships will support conclusions that grammatical constraints are not necessary to account for the observed acceptability contrasts. This paper is a first step in what we hope will be a continuing process of developing criteria for establishing the role of functional constraints in acceptability contrasts – a necessary part of collecting and assessing evidence for grammatical theory-building.

Acknowledgments We gratefully acknowledge discussions and input from Daniel Casasanto and the assistance of David Kettler in conducting the experiments.

References

Baayen, R. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers*, *1*, 1–45.

Baayen, R. (2007). *Analyzing Linguistic Data: A Practical Introduction to Statistics.* Cambridge, UK: Cambridge University Press.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude Estimation of Linguistic Acceptability. *Language*, *72*(1), 32–68.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (p. 232-86). New York: Holt, Reinhart & Winston.

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

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*(12), 769–786.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–66.

Featherston, S. (2008). Thermometer judgments as linguistic evidence. In M. Claudia & A. Rothe (Eds.), *Was ist linguistische Evidenz?* Aachen: Shaker Verlag.

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods*, *35*(1), 116–124.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, *29*, 261–290.

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*(98), 122–149.

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* (p. 241-279). San Diego, CA: Academic Press.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, p. 419-492). New York: Wiley.

Ross, J. R. (1967). *Constraints on Variables in Syntax.* PhD Thesis. MIT.

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