Jon Sprouse

Three Open Questions in Experimental Syntax

Abstract: This article presents a review of current research in experimental syntax, with a focus on three open questions and the (methodo)logical tools that have been developed to explore them. The three questions are: (1) Are the published data underlying syntactic theories valid?, (2) How can we determine the source of acceptability judgment differences?, and (3) What do gradient judgments tell us about the architecture of the grammar? The hope is that these three questions will illustrate the fundamental connection between experimental syntax and theoretical syntax, provide concrete demonstrations of the value of the tools of experimental syntax (e.g., random sampling, factorial logic, and gradient judgments), and point to new directions for future research.

Keywords: experimental syntax, acceptability judgments, validity, island effects, gradient grammars

DOI 10.1515/lingvan-2014-1012

1 Introduction

The use of formal experimental methods for the collection of acceptability judgments has grown in popularity over the past 15 to 20 years, thanks in no small part to Schütze's (1996) treatise on the empirical base of linguistics, and Cowart's (1997) textbook on acceptability judgment collection methods. Today the use of formal methods are so widespread that it is not uncommon to describe studies using such methods as part of a "field of experimental syntax", following the title of Cowart's textbook. In this article I would like to review three open questions in experimental syntax, with a particular focus on the logical and methodological tools that have proven useful in exploring these questions. Although this may seem like straightforward review material, there is one aspect of experimental syntax that makes selecting open questions a bit tricky: experimental syntax really isn't a distinct field. Experimental syntax is a set of methods for collecting data in service of *theoretical syntax*. In other words, to the extent that there is a field of experimental syntax, it simply inherits its questions directly from the field theoretical syntax. Complicating the picture even further, the field of theoretical syntax already has a set of experimental methods: the traditional, informal judgment collection methods that have been deployed in nearly every theoretical syntax investigation since (at least) the 1950s (for the logic of using judgments, see Chomsky 1965 and Newmeyer 1983). The informal methods of theoretical syntax and the formal methods of experimental syntax share many properties (such as the creation of tightly controlled experimental conditions, and the testing of the same behavioral response), and tend to differ as a matter of degree (such as the number of items tested, or the number of participants recruited), not category (the one categorical difference being the background knowledge of participants; see Section 2). So when one asks the question *What are the driving questions of experimental syntax?*, it seems to me that the meaning behind the question must be something closer to *What are the new (or important) theoretical syntax questions that become (more) tractable when the traditional informal judgment collection methods are formalized using experimental syntax techniques?* The latter is a mouthful, so the shorthand is understandable. But I believe it is important to keep the relationship between theoretical syntax and experimental syntax in mind while evaluating the role of experimental syntax in modern theoretical syntax, and its potential moving forward.

With all of that in mind, I have attempted to identify three open questions that highlight either the nature of syntactic data, or the relationship between syntactic data and syntactic theory:

Jon Sprouse, Department of Linguistics, University of Connecticut, Storrs, CT 06269, USA, E-mail: jon.sprouse@uconn.edu

1. Are the published data underlying syntactic theories valid?
2. How can we determine the source of acceptability judgment differences?
3. What do gradient judgments tell us about the architecture of the grammar?

As with any review, the choices here reflect a bit of an editorial bias on my part. For example, I have chosen to focus on acceptability judgment methods over other methods (such as reaction times and EEG). This is partly for practical reasons, as there isn't enough space to do justice to more than one method in a short review article. This is also partly scientific, as the (relatively more complex) linking hypothesis between these other data types and syntactic theories are only just beginning to be explored. I have also chosen to focus on three questions that have figured prominently in my own research. Editorial bias aside, I believe these three questions have properties that make them ideal for a review of this sort: each has important consequences for theories of syntax, each has brought several logical and methodological tools into focus that are likely to be of use as the fields of theoretical and experimental syntax move forward, and each is still currently open for debate and future research.

2 Are the published data underlying syntactic theories valid?

Given the central role that acceptability judgments play in the theoretical syntax literature, it is perhaps not surprising that the single most frequent question in the methodological literature is to what extent linguists can trust the acceptability judgments reported in the literature. This question has arisen in one form or another since the earliest days of generative grammar (e.g., Hill 1961; Spencer 1973), it has played a central role in the two books that ushered in the modern approach to experimental syntax (Schütze 1996; Cowart 1997), and it has led to several high-profile discussions in the past decade (see Ferreira 2005; Wasow and Arnold 2005; Featherston 2007; Gibson and Fedorenko 2010, 2013 for some criticisms of informal methods, and Marantz 2005 and Phillips 2009 for some rebuttals). The fundamental concern is that informal judgment collection methods may give rise to spurious results. If the results are spurious, it would follow that the syntactic theories themselves are also spurious (with domino effects for fields that build on syntactic theories, such as language acquisition and sentence processing). This is a serious concern, and perhaps fortuitously, also a concern that experimental syntax is well-positioned to address. As such, it is perhaps unsurprising that this has been a lively topic in the experimental syntax literature (see also Myers 2009 for a review of concerns about judgment collection methods, along with recommendations about how many concerns can be alleviated with relatively small experiments).

When it comes to assessing experimentally collected data, we must first confront a fundamental fact about experimental science: we can never know whether our experiments are yielding measurements that are true reflections of the universe. To know that would require independent knowledge of the universe, thus circumventing the need for measurements in the first place. In lieu of certainty, we attempt to build confidence in our measurements. This usually revolves around two dimensions: the *validity* of the measurement, which is its ability to measure the property of interest, and the *reliability* of the measurement, which is its ability to consistently measure the property of interest.[1] To increase confidence in validity, we often look for consistency between two distinct methods intended to measure the same construct, look for consistency between our measurements and the predictions of uncontroversial theories, and look for consistency between the measurement methodology and the best practices agreed upon by the research community. To increase confidence in reliability, we often repeat measurements under unchanged conditions (*replication*). In the case of informal judgment collection, many of these markers of confidence are either covert or missing altogether. In addition, many (if not all) informally collected judgments come from professional linguists rather than linguistically-naïve participants. Professional linguists may be aware of the syntactic theories under

[1] For quantitative methods, once validity and reliability are established, we can also investigate the *accuracy* and *precision* of the measurements, but experimental syntax has not reached that point yet.

investigation, leading to a type of cognitive bias that may impact their judgments (an issue raised in nearly every previous discussion of syntactic methods). As such, it is not surprising that some researchers have expressed skepticism in the validity (and perhaps reliability) of informal methods.

There is a straightforward method for determining whether this lack of confidence in informal methods is justified: compare the results of informal methods with formal experimental methods. The results that converge between the two methods will benefit from the increase in confidence. The results that diverge can then be further investigated to determine which method is more likely giving the valid result (i.e., by manipulating the factors that give rise to concern in each method, such as the linguistic knowledge of the participants). The tools of experimental syntax make such studies eminently possible. However, there are at least two issues that require careful consideration. The first is how to select phenomena to be tested. It is not uncommon for discussions of this topic to present a few phenomena from the literature, along with formal experiments showing results that diverge from the informally collected results (e.g., Gibson and Fedorenko 2013). The concern with these sorts of studies is that the phenomena weren't selected randomly, or in statistical terms, were selected *with bias*. A biased sample of phenomena cannot be used to statistically generalize to a larger population of phenomena; instead, we either need to test the entire population of phenomena (exhaustive testing), which would reveal the exact number of divergent results, or randomly sample from the population, which would allow us to statistically estimate the number of divergent results within a margin of error.

The second issue is a practical one: at some point, a decision must be made as to whether the informal method is valid and reliable, or not. There are real costs to advocating a complete switch to formal methods (in terms of time and money). If informal methods are invalid and unreliable, the cost will be justified and the field will need to adjust accordingly. However, if informal methods are valid and reliable, then formal experiments can be reserved for the questions where they provide unique information (such as the questions in Sections 3 and 4). In null hypothesis significance testing (of the Neyman-Pearson variety), this question is often framed in terms of *Type I errors*, also known as *false positives* or *false rejections of the null hypothesis*, which are the errors that arise when a theorist acts as if there is a significant difference between one or more conditions when constructing a theory, but in fact no such difference is true of the world. It is common to talk about maximum Type I error rates, which is the rate of Type I errors (false positives) that would occur if a statistical test were repeated an infinite number of times. The same question arises in syntax: what is the maximum number of false positive results that we are willing to tolerate from informal methods? It is tempting to answer zero, but every experimental result carries a risk of error. For example, the conventionally agreed upon maximum Type I error rate in experimental psychology is 5% (which arises because the threshold for behaving as if a result is significant, called the alpha-criterion by Neyman-Pearson, has been conventionally set at $p < 0.05$). The question facing the field of syntax is whether we also want to adopt the same or a different criterion.

Crucially, once that criterion is decided, we can easily evaluate the convergence/divergence between informal and formal methods to determine if the extra cost of formal methods should be imposed on the field or not. Under the assumption that convergent results are very likely to be true positives (an assumption that can certainly be questioned), it follows that the divergence rate becomes a maximum Type I error rate for informal methods (the error rate would be zero if follow-up studies show that the informal method was always correct, and the error rate becomes the divergence rate if follow-up studies show that the formal method was always correct). There are currently two studies that take both of these issues (phenomena selection and divergence rate) into consideration. First, Sprouse and Almeida (2013) exhaustively tested every English phenomenon in a recent syntax textbook (Adger 2003) that could be tested in a simple acceptability judgment survey. Using the best practices of experimental syntax (multiple items, pseudorandomized surveys, large samples of naïve participants), and several statistical methods, they found a convergence rate of 98%, and therefore a divergence rate of 2%. Second, Sprouse et al. 2013 randomly sampled English phenomenon from 10 years of syntax articles published in *Linguistic Inquiry*. Again, using the best practices of experimental syntax, and based on the number of phenomena randomly sampled, they found a convergence estimate of 95% ± 5, and therefore a divergence estimate of 5% ± 5 (Figure 1).

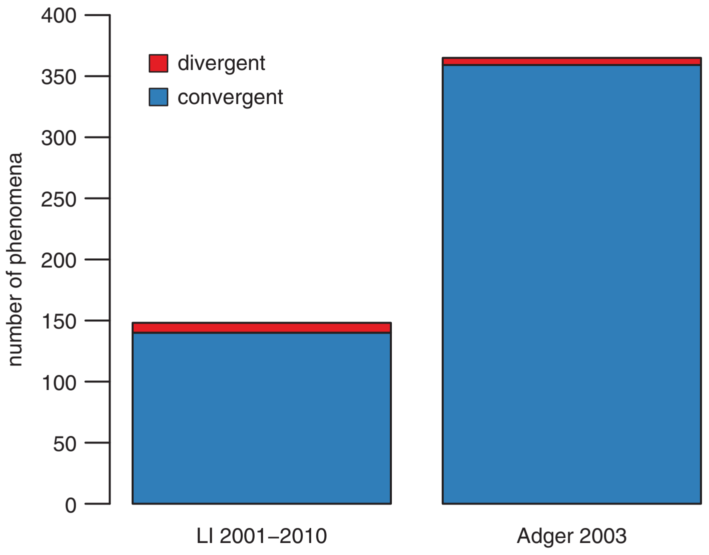

Figure 1 Convergent and divergent phenomena in *Linguistic Inquiry 2001–2010* and Adger (2003)

Both of these rates are at or below the criterion set by experimental psychology, so depending on what the field decides should be the conventional criterion in syntax, it may be the case that informal methods will ultimately be considered a valid and reliable method for phenomena that do not require the extra information provided by formal methods. However, a note is in order about the scope of these results. In the case of the Adger textbook study, the selection method was exhaustive, therefore the results are a great estimate of the convergence for the data in Adger (2003), but we don't know how well they extend to other populations of phenomena. For the *LI* study, the selection method was random, which allows for generalization to the full population (10 years of *LI* articles), but we again restricted that selection process to English phenomena that could be tested in a standard acceptability judgment survey (Figure 2).

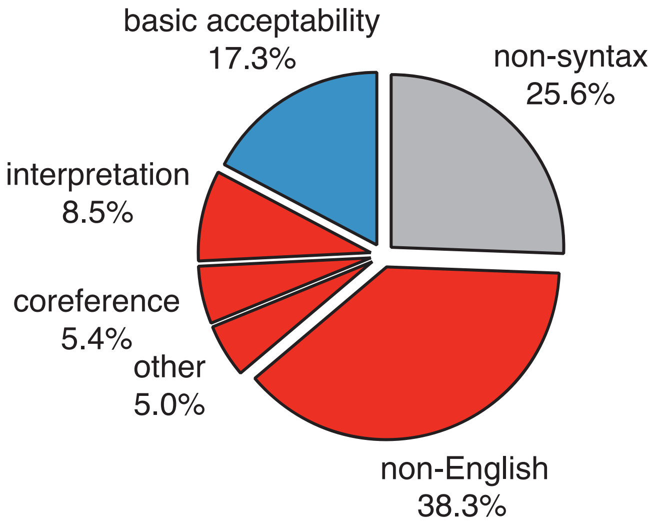

Figure 2 The distribution of types of data in *Linguistic Inquiry 2001–2010*

This means that in both cases, there is much work to be done examining divergence rates for other types of phenomena, and for languages other than English. Experimental syntax is well-positioned to provide these future studies, but it will require participation from a diverse array of linguists. It may be the case that other data types and other languages will reveal substantially different divergence rates, in which case the validity and reliability of informal methods will be a more nuanced issue.

3 How can we determine the source of acceptability judgment differences?

The second big question in experimental syntax arises because of the fundamental problem of cognitive science: we cannot directly observe cognitive operations. To get around this, we use data types that we can observe to make inferences about the underlying cognitive operations. In the syntax literature, acceptability judgments are used to identify differences in acceptability between two (or more) minimally differing sentences types (in experimental syntax we can call each sentence type a *condition* of the experiment). When a difference in acceptability is detected between conditions, syntacticians attempt to explain that difference as a consequence of the functioning of the grammar (e.g., as a violation of a syntactic constraint). However, acceptability judgments are a behavioral response that occurs as a by-product of sentence processing. This means that, in principle, every cognitive system that contributes to sentence processing contributes to acceptability judgments. This in turn means that any difference in acceptability that is detected between conditions could be explained as a consequence of these other cognitive systems. Given the sheer number of cognitive systems at work during sentence processing (syntax, semantics, pragmatics, ambiguity resolution, working memory, etc.), it is often nearly impossible to create conditions that manipulate syntactic properties without also manipulating properties of these other cognitive systems. From the point of view of syntactic theory these differences from other systems would be considered confounds, but from the point of view isolating the source of acceptability judgment effects, these differences in other cognitive systems are potential non-syntactic explanations for the observed effects. The question then is whether we can use experimental syntax to help tease apart explanations that require syntactic constraints (*syntactic explanations*) from explanations that reduce the effect to consequences of other cognitive systems (*reductionist explanations*).

In principle, the method for assessing the likelihood of a reductionist explanation is straightforward: manipulate the properties of the critical non-syntactic system(s) while holding the syntactic properties constant, and look for changes in the observable acceptability difference. However, in practice, it is often difficult to disentangle reductionist explanations from syntactic explanations. We can use island effects as a concrete example. Island effects, as a phenomenon, can be defined as the decrease in acceptability that occurs when a long-distance dependency originates within certain structures, which we can call island structures (in the examples below, the island structure is indicated with square brackets, the head of the dependency by italics, and the tail of the dependency by an underscore):

(1) a. *What* do you wonder [whether John bought __]?
b. *What* did you make [the claim that John bought __]?
c. *What* do you think [the rumor about __] prompted the congressional hearing?
d. *What* do you worry [if John forgets __ at the office]?

In the syntax literature, island effects tend to be analyzed as the consequence of one or more syntactic constraints, often descriptively called island constraints, that make this specific configuration ungrammatical (e.g., Ross 1967; Chomsky 1986). However, Kluender and Kutas (1993) observe that sentences containing island effects always contain two properties that could potentially decrease acceptability due to processing difficulty without any contribution from a syntactic constraint: (i) the presence of a long distance dependency, and (ii) the presence of complex structure (i.e., the island structure). These two properties are always part of the structural description of an island effect, so both syntactic and reductionist explanations can explain the acceptability decrease in (1).

In principle, experimental syntax provides a number of methods for attempting to tease apart syntactic and reductionist explanations. One possibility is to look for secondary differences in the judgments to different types of unacceptable sentences, such as the fact that some unacceptable sentences appear to increase in acceptability after repeated exposure. This effect is often called satiation in the syntactic literature, and has been studied extensively for island effects (as well as other phenomena). In principle, the idea is that if two phenomena differ in the source of the effect (e.g., syntax vs reductionism), their satiation properties (either the presence of satiation, or the rate of satiation) may also differ. In practice, however, satiation studies have yielded conflicting results (e.g., Snyder 2000; Hiramatsu 2000; Sprouse 2009; Francom 2009). Another possibility is to use the real time behavior of the parser to test predictions of reductionist theories. For example, Phillips (2006) argues that some reductionist theories might predict that the parser cannot actively complete dependencies when the gap is inside an island. He then shows reaction time evidence that the parser actively attempts to complete dependencies when the gap is inside certain subject islands, despite the fact that those sentences are judged to be unacceptable in offline studies. This suggests that the source of the unacceptability cannot be due to reductionist theories that predict that dependency completion is impossible. In a similar vein, Yoshida et al. (2013) argue that some reductionist theories might predict that all dependencies that involve working memory costs should respect island constraints. They then show that the parser does not respect island constraints for "backward" binding dependencies in which a pronoun appears at the beginning of a sentence and the antecedent appears inside of an island later in the sentence (but crucially the parser does respect island effects for wh-dependencies). This selectivity of island effects runs counter to the prediction of some reductionist theories.

Because satiation studies are currently inconclusive, and reaction time studies depend upon questions about how the parser works in real time, in the rest of this subsection I would like to review two additional tools that can be used to help tease apart syntactic and reductionist theories in more traditional offline acceptability judgment experiments. The first is the concept of (fully-crossed) factorial designs. In experimental design, a *factor* is a property that can be manipulated, such as the length of a dependency, or the presence or absence of an island structure. Each value that a factor can take is called a *level*. By choosing factors and levels that instantiate the components of a reductionist explanation, it is possible to isolate the contributions of each component. Again using island effects as a concrete example, we can quantify the effect of a long distance dependency with a factor called LENGTH, and two levels, SHORT and LONG, as in (2a) and (2b), such that the subtraction [(2a)–(2b)] yields a measure of the effect of dependency length. We can quantify the effect of island structures with a factor called STRUCTURE, and two levels, NON-ISLAND and ISLAND, as in (2a) and (2c), such that the subtraction [(2a)–(2c)] yields a measure of the effect of island structures.

(2) a. *Who* __ thinks that John bought a car? SHORT | NON-ISLAND
b. *What* do you think that John bought __? LONG | NON-ISLAND
c. *Who* __ wonders [whether John bought a car]? SHORT | ISLAND
d. *What* do you wonder [whether John bought __]? LONG | ISLAND

To make this a fully crossed design, we add (2d) which is a sentence that combines both the LONG level of LENGTH, and the ISLAND level of STRUCTURE, and is also the critical island-violating sentence. Because we've isolated the effects of length and structure in the first three sentences, we can make a prediction for the fourth: if the acceptability of island effects is completely explainable by the effects of long distance dependencies and island structures, then the acceptability of (2d) will be the sum of those two effects. In mathematical terms: [(2a)–(2d)] = [(2a)–(2b)] + [(2a)–(2c)]. In graphical terms, if the island effect is the sum of those two effects, a graph of the four conditions will yield two parallel lines as in the left panel of Figure 3. On the other hand, if there is more to island effects than just the sum of the length effect and the structure effect, then the acceptability of (2d) will be *lower* than predicted by the other three conditions. This will yield non-parallel lines in a graph as in the right panel of Figure 3. In mathematical terms: [(2a)–(2d)] > [(2a)–(2b)] + [(2a)–(2c)]. This is also called a superadditive effect, or superadditive interaction, in experimental studies.

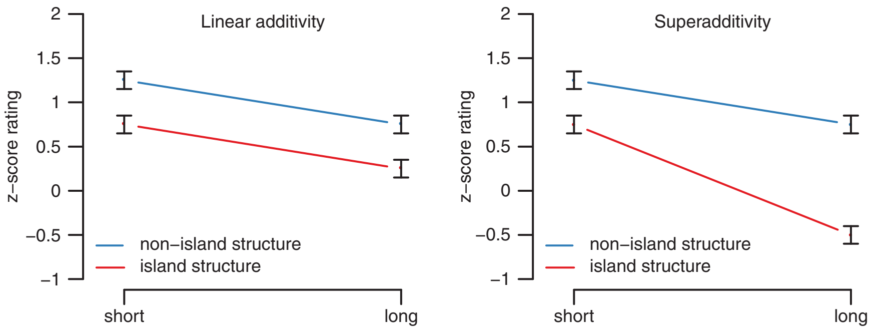

Figure 3 Examples of linear additivity (left) and superadditivity (right)

Factorial designs like the one above provide a first test for reductionist theories: if the critical effect can be captured by the sum of reductionist components, then the reductionist theory is likely true. However, if a superadditive effect is observed, the results are ambiguous. It could either be that there is a syntactic constraint causing the superadditive effect, or it could be that the two reductionist components interact in a complex way to yield a superadditve effect. In other words, superadditivity is a necessary condition for a syntactic constraint, but not a sufficient one. Once superadditivity is identified, it becomes the effect in need of an explanation: syntactic theories explain it with a syntactic constraint, while reductionist theories likely explain it with a complex interaction of non-syntactic effects. To resolve this ambiguity, we must use the second tool that experimental syntax makes available: correlating diverse data types. Although syntactic theories primarily use acceptability judgments as evidence, reductionist theories are predicated upon non-syntactic cognitive systems, which are likely to be amenable to investigation using other types of behavioral responses. If one can identify the cognitive system that is thought to give rise to the acceptability effect in question, and then identify a behavioral response that is also affected by that specific cognitive system, it may be possible to use the formal results of experimental syntax studies to look for statistical correlations between the superadditive effect observed in acceptability judgments and the other behavioral response in order to assess the likelihood of the reductionist theory.

As a concrete example, Sprouse et al. (2012) found that four island types in English all show superadditive patterns using a factorial design as in Figure 4.

This suggests either a syntactic explanation, or a complex reductionist explanation. Kluender and Kutas (1993) provide one such complex reductionist explanation. They suggest that both long distance dependencies and island structures might draw on the same set of working memory resources in order to be successfully parsed. If true, this would predict that attempting to parse both, as in island effects, might result in a larger-than-expected effect (a superadditive interaction). Sprouse et al. argue that one plausible prediction of this theory is that individual differences in working memory capacity will lead to differences in the size of the superadditive effect that individuals report using acceptability judgments. To test this, they asked a large number of participants to complete both a series of working memory tests and an acceptability judgment experiment, and looked for correlations between the results of the two experiments.

Although Sprouse et al. (2012) observe no significant correlations, casting some doubt on the complex reductionist explanation they tested, the more interesting result is the proof of concept that two of the tools made available by experimental syntax, factorial design and data correlation, can be used to explore the source of acceptability judgment effects. These tools have only just begun to be explored in the experimental syntax literature, but they promise to be useful for any number of phenomena, such as other dependency constraints, that are potentially amenable to reductionist theories (see also Sprouse et al. 2011 for island effects in Japanese, Hofmeister, Culicover, and Winkler *in press* for freezing effects in English, and Sprouse *et al. in press* for island effects with relative-clause formation in English and Italian).

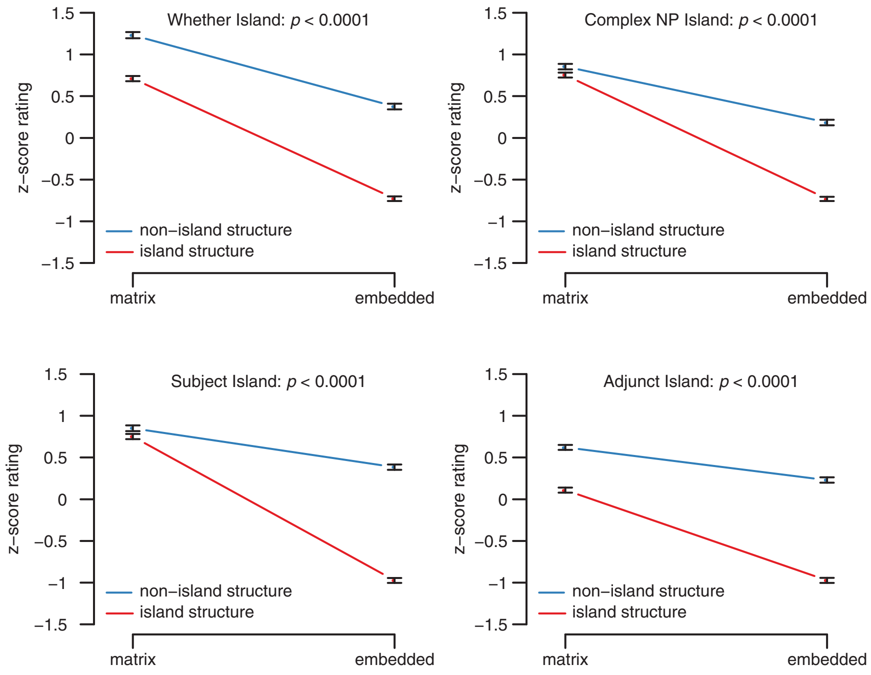

Figure 4 Island effects in English using the factorial design (Sprouse et al. 2012). The *p*-values are reported for the interaction term in a two-way linear mixed-effects model

4 What do gradient judgments tell us about the architecture of the grammar?

The third question concerns a fact about acceptability judgments that is simultaneously the most obvious to observe and perhaps the most difficult to explore: acceptability judgments appear to be continuous. Figure 5 plots the acceptability of the 300 sentence types tested by Sprouse et al. (2013) in order of

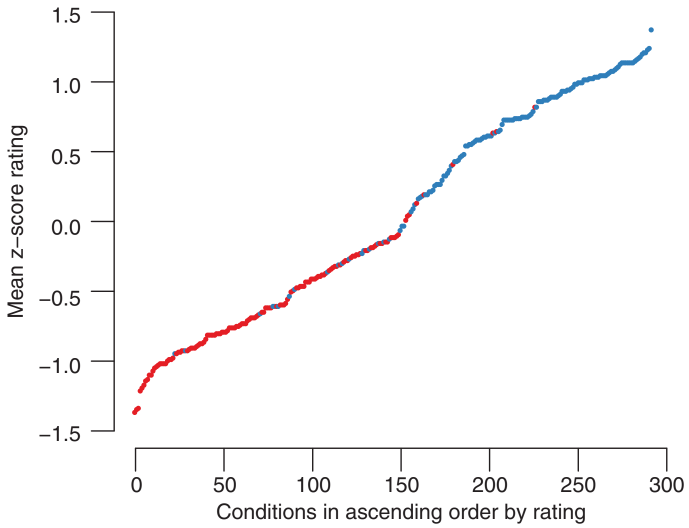

Figure 5 Acceptability for 300 sentence types randomly sampled from *Linguistic Inquiry 2001–2010* plotted in ascending order (Sprouse et al. 2013). Red dots indicate sentence types that were given a diacritic (*,?, or a combination) in LI. Blue dots indicate sentence types that were given no diacritic

increasing acceptability (the y-axis is mean z-score transformed ratings, the x-axis is simply ascending order of acceptability).

As Figure 5 demonstrates, with a large enough random sample of sentence types, the acceptability of the sentences is fairly evenly distributed across the (non-countably infinite) range of possible acceptability levels. This continuity, sometimes called gradience, appears to be a fact of acceptability judgments. This fact has long been acknowledged in the syntax literature (e.g., Chomsky 1964), and has even been leveraged as evidence to distinguish different types of syntactic constraints (e.g., Subjacency violations vs. Empty Category Principle violations in Chomsky 1986). What has changed with experimental syntax is that formal judgment experiments, especially those that use a continuous response scale (e.g., magnitude estimation), or approximate a continuous scale (e.g., z-score-transformed Likert scales), bring the gradience of acceptability judgments into sharp focus. The question, then, is what does this gradience tell us about the architecture of the grammar?

In principle syntactic architectures can be divided into two broad classes: binary-categorical theories and weighted-constraint theories. Binary-categorical theories are the most common, and likely the most familiar. In binary-categorical theories the grammar either generates a sentence or does not generate a sentence. Since the syntax only yields two values, this means that the gradience of judgments must derive from non-syntactic cognitive systems (pragmatics, real-world plausibility, parsing difficulty, etc.). In contrast, in weighted-constraint theories the syntax plays a larger role in accounting for gradient judgments. In weighted-constraint theories each constraint in the syntax is associated with a value, such that combining these values leads to a large range of possible levels of grammaticality (2^N levels, where N is the number of constraints). Weighted-constraint theories still assume that non-syntactic cognitive systems contribute to acceptability, but the relative contribution of syntax is higher. There are at least three prominent weighted-constraint theories in the experimental syntax literature: Keller's (2000) Linear Optimality Theory (see also Sorace and Keller 2005), Bresnan's (2007) Stochastic Optimality Theory, and Faetherston's (2005a) Decathlon Model (see also Featherston (2005b) for an example of a previously undetected Superiority effect in German that is potentially due to a syntactic constraint). There are also several instances in the literature where weighted-constraints were inserted in otherwise binary-categorical theories, such as the distinction between Subjacency violations and ECP violations (e.g., Chomsky 1986), or the distinction between strong and weak island effects (e.g., Szabolcsi 2006). For space reasons I won't review the details of specific theories here, but instead focus on the empirical facts that have been revealed by experimental syntax, and that must be addressed by both classes of theories.

The first fact is variation in effect sizes across phenomena. Syntacticians define an effect as a difference in acceptability. Experimental syntax has brought into focus the fact that different phenomena lead to different sizes of acceptability differences. This can again be illustrated with the large, random sample of phenomena tested from *Linguistic Inquiry* in Sprouse et al. (2013). Figure 6 plots the size of the acceptability effect for each of the 150 phenomena investigated.

Binary-categorical syntactic theories must account for these differences in effect size across phenomena based on non-syntactic factors impacting the acceptability of each of the sentences. The simplest theory would posit a single effect size for ungrammaticality (e.g., 0.5 z-score units), and then explain the variation through the linear addition of other factors, such as difficulty parsing, or even the ease of parsing in the case of effects that are smaller than the ungrammaticality effect. A more complex theory could postulate interactions (super- or sub-additive effects) among the non-syntactic factors. In either case, the explanatory burden is to find a set of factors that can capture the effect size variability within a binary-categorical grammar, and that will make predictions about the acceptability of future sentence types. Weighted-constraint theories can in principle capture the different effect sizes by postulating a set of constraints and values that give rise to the different effect sizes. In this case, the explanatory burden is to provide an account that goes deeper than just capturing the acceptability effects. This can be accomplished by tying the weights to an independent property (e.g., probability of occurrence), or by associating this property with units that are smaller than the sentence, which can then make predictions for future sentence types that can

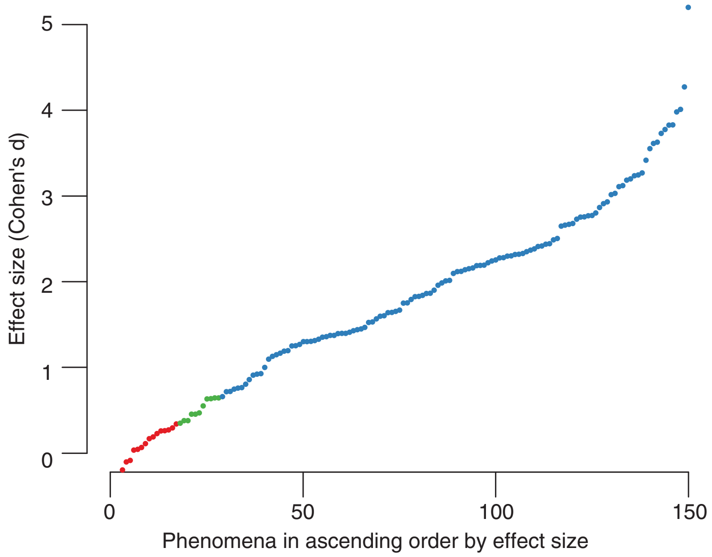

Figure 6 Effect sizes in ascending order for 150 (two-condition) phenomena randomly sampled from *Linguistic Inquiry 2001–2010*. Color indicates the category of effect size: red indicates "small" effect sizes, green indicates "medium" effect sizes, and blue indicates "large" effect sizes

be empirically tested. For both grammatical architectures, once the predictions are worked out, experimental syntax methods can be used to assess the success of the predictions.

The second fact is variation in effect sizes between phenomena that appear to involve the same, or at least closely related, constraints. For example, if one takes the superadditive component of the factorial design for island effects as a measure of effect size (see Section 3), then we can compare the size of island effects across island types, across dependency types, and even across languages. Any variation in effect sizes must be explained. As a concrete example we can compare the sizes of island effects for whether, complex NP, subject, and adjunct islands in English with bare wh-word dependencies, and complex *which*-phrase dependencies (see Sprouse et al. *in press*).

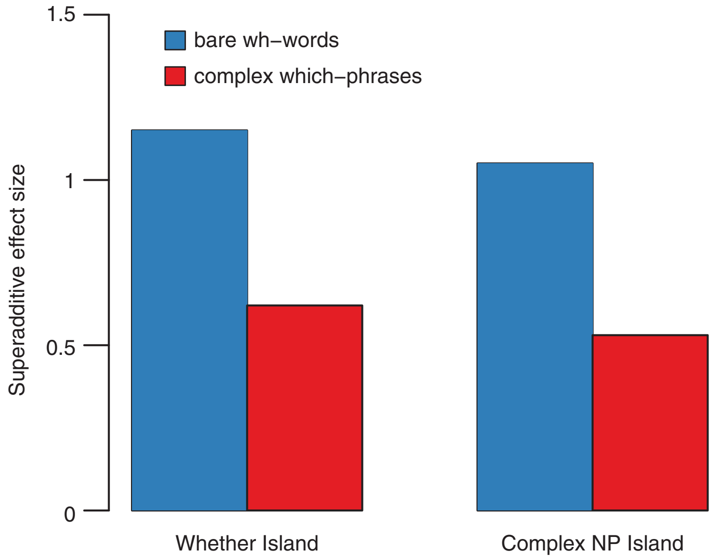

Figure 7 A comparison of the size of *whether* and complex NP island effects (in terms of the superadditive effect size) for bare wh-words and complex *which*-phrases

The results suggest that both whether and complex NP island effects are substantially smaller with complex *which*-phrases (Figure 7), corroborating claims in the literature that complex *which*-phrases tend to ameliorate certain island effects. For binary-categorical theories, this fact must be explained with non-syntactic factors that happen to differ across islands, dependencies, or languages. For weighted-constraint theories, this fact can be explained either as two distinct constraints with distinct weights (one for each dependency type), or different weights for the same constraint in the two dependency environments.

The field is still in the first stages of collecting facts about gradience, with very few definitive analyses. Progress will require both a concerted effort to collect facts across constructions, constraints, and languages, and a strong push to fully elaborate both classes of theories. Experimental syntax provides tools for at least two stages of this exploration: quantifying the gradience facts, and exploring the predictions of novel theories. However, the theorizing will require good old fashioned logical (and creative) thinking.

5 Conclusion

Experimental syntax provides a set of formal data collection methods to further explore the central questions of theoretical syntax. Current research in experimental syntax is focused on exploring questions that would not otherwise be answerable with traditional informal data collection methods, such as investigating the validity of traditional methods themselves, investigating the source of acceptability judgment effects, and investigating the source of gradient acceptability judgment effect sizes. Although there have been several tantalizing initial results in these explorations, we have only begun to scratch the surface of the potential of experimental syntax to shed light on these questions. A concerted effort to apply experimental syntax methods to a larger selection of constructions and languages will likely lead to a rapid expansion in the number of mysteries in need of explanation, and presumably, new avenues for exploring the architecture of the grammar.

References

Adger, David. 2003. *Core syntax*. Oxford: Oxford University Press.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 75–96. Berlin: Mouton de Gruyter

Chomsky, Noam. 1964. Degrees of grammaticalness. In J. A. Fodor & J. J. Katz (eds.), *The structure of language*, 384–389. Englewood, NJ: Prentice-Hall..

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1986. *Barriers*. Cambridge, MA: MIT Press.

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

Featherston, Sam. 2005a. The Decathlon Model of empirical syntax. In M. Reis & S. Kepser (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 187–208. Berlin: Mouton de Gruyter.

Featherston, Sam. 2005b. Magnitude estimation and what it can do for your syntax: Some WH-constraints in German. *Lingua* 115. 1525–1550.

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33. 269–318.

Ferreira, Fernanda. 2005. Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review* 22. 365–380.

Francom, Jared. 2009. *Experimental syntax: Exploring the effect of repeated exposure to anomalous syntactic structure—Evidence from rating and reading tasks*. PhD diss., University of Arizona.

Gibson, Edward & Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14. 233–234.

Gibson, Edward & Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28. 88–124.

Hill, Archibald A. 1961. Grammaticality. *Word* 17. 1–10.

Hiramatsu, Kazuko. 2000. *Accessing linguistic competence: Evidence from children's and adults' acceptability judgments*. PhD diss., University of Connecticut.

Hofmeister, Philip, Peter W. Culicover & Susanne Winkler. in press. Effects of processing on the acceptability of 'frozen' extraposed constituents. *Syntax*.

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD diss., University of Edinburgh.

Kluender, Robert & Marta Kutas. 1993. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8. 573–633.

Marantz, A. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22. 429–445.

Myers, James. 2009. Syntactic judgment experiments. *Language and Linguistics Compass* 3. 406–423.

Newmeyer, Frederick. 1983. *Grammatical theory: Its limits and its possibilities*. Chicago: University of Chicago Press.

Phillips, Colin. 2006. The real-time status of island constraints. *Language* 82. 795–823.

Phillips, Colin. 2009. Should we impeach armchair linguists? In S. Iwasaki, H. Hoji, P. Clancy & S. -O. Sohn (eds.), *Proceedings of the 17th Conference on Japanese/Korean Linguistics*, 49–64. Stanford, CA: CSLI.

Ross, John Robert. 1967. *Constraints on variables in syntax*. Cambridge, MA: MIT dissertation.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Snyder, W. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31. 575–582.

Sorace, Antonella & Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115. 1497–1524.

Spencer, N. J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2. 83–98.

Sprouse, Jon. 2009. Revisiting satiation. *Linguistic Inquiry* 40. 329–341.

Sprouse, Jon & Diogo Almeida. 2013. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48. 609–652.

Sprouse, Jon, Ivano Caponigro, Ciro Greco & Carlo Cecchetto. in press. Experimental syntax and the cross-linguistic variation of island effects in English and Italian. *Natural Language and Linguistic Theory*.

Sprouse, Jon, Shin Fukuda, Hajime Ono & Robert Kluender. 2011. Reverse island effects and the backward search for a licensor in multiple *WH*-questions. *Syntax* 14. 179–203

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

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

Szabolcsi, Anna. 2006. Strong vs. weak islands. In M Everaet, H. van Riemsdijk, R. Goedemans & B. Hollebrandse (ed.), *The Blackwell companion to syntax volume 4*, 479–531. Oxford: Blackwell.

Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115. 1481–1496.

Yoshida, Masaya, Nina Kazanina, Leticia Pablos & Patrick Sturt. 2013. On the origin of islands. *Language and Cognitive Processes* 29. 761–770.