The Role of Experimental Syntax in an Integrated Cognitive Science of Language

Jon Sprouse and Diogo Almeida

1. Introduction

Acceptability judgments form the primary empirical foundation for generative syntactic theories (Chomsky 1965, Schütze 1996). As such, the methodology of acceptability judgment collection has been a topic of research since the earliest days of generative syntax (e.g., Hill 1961, Spencer 1973). However, the past fifteen years have seen a dramatic increase in the number of articles devoted to the topic. It seems clear that the recent increase in interest in methodological issues is related to advances in technology that have made it easier than ever to construct, deploy, and analyze formal acceptability judgment experiments, which following Cowart (1997) have come to be called *experimental syntax* (a practice that we will follow in this chapter). The question at the center of this literature is deceptively simple: *How can formal acceptability judgment experiments help achieve the goals of generative syntax?* As we will see in this chapter, answering this question is surprisingly complex. A comprehensive answer to this question requires (at least) three components: (1) an explicit formulation of the goals of generative syntax, (2) an enumeration of the potential obstacles to those goals, and (3) an empirically-driven evaluation of the ability of formal experiments to eliminate those obstacles. In this chapter we will present a comprehensive review of the recent acceptability judgment literature with respect to these three components in an attempt to provide (our version of) an answer to the question of how formal judgment experiments can help generative syntactic theory.

2. The goals and obstacles of generative syntactic theory

Our starting assumption is that all cognitively-oriented language researchers share the goal of constructing a theory of language that integrates all three of Marr's famous levels of analysis: the computational level, the algorithmic level, and the implementational level (Marr 1982, see also discussion in Phillips 1996, Phillips and Lewis 2010, Kluender 1991, Frazier 1978, Embick and Poeppel 2005, and many others). Marr (1982) used cash registers as an illustrative example to define these three levels for information processing devices (such as the human brain). The computational level of the theory is a description of the properties of the problem that must be solved by the device, as well as the operations that the device must perform, that abstracts away

from the exigencies of actually solving the problem in practice. For a cash register, the computational level description is the theory of addition, with properties such as commutativity and associativity, abstracting away from the precise algorithms that are necessary to carry-out addition. For the sentence level phenomena of language, syntactic theories are computational level descriptions, as they describe the properties of the final syntactic structures that must be built, as well as the properties of the structure building operations that are required to build them, but abstract away from the requirements of real-time sentence processing. The algorithmic level of the theory is a description of the actual operations that must be deployed to solve the problem (i.e., an algorithm). For a cash register this could be the base-10 addition algorithm that we learned in school: start from the right, and "carry over the ones." For language, parsing theories are algorithmic level theories, as they describe the specific parsing operations that must be deployed during real-time sentence processing, including the strategies that dictate the deployment of those operations, and the ways in which parsing resources constrain the operation of the parser. Finally, the implementational level of the theory is a description of how the processes/strategies/resources are implemented in the hardware of the device. For a cash register, there are several hardware options that can influence this level (e.g., spinning drums versus electronic processors). However, for (a cognitive approach to) language, there is only one set of hardware, the human brain. Neurolinguistic theories, which seek to identify the cortical networks involved in various linguistic computations, are a first step toward implementational level descriptions (Embick and Poeppel 2005, Sprouse and Lau 2012).

There are at least two major obstacles to the construction of an integrated theory of language. The first is the black box problem: there is no method to directly measure cognitive mechanisms. What this means in practice is that researchers must (i) identify observable data types (behavior, electrophysiological responses, hemodynamic responses), and (ii) identify linking hypotheses that license (empirically valid) inferences from the observable data to the unobservable cognitive mechanisms. The black box problem affects all three levels of the theory; however, in this chapter we will focus on the data and linking hypotheses underlying syntactic theory (see Sprouse and Lau 2012 for a description of the data types and linking hypotheses at the algorithmic and implementational levels). Crucially, the black box problem presents a framework for investigating the empirical contribution of experimental syntax by focusing the discussion on the following questions: *To what extent is the data underlying current incarnations of syntactic theory sound?* And, *What types of inferences are licensed by the linking hypothesis between acceptability judgment data and syntactic theory?*

Whereas the first major obstacle to the construction of an integrated theory of language,

the black box problem, presents a framework for investigating the empirical contribution of experimental syntax, the second major obstacle presents a framework for understanding the historical and sociological context of recent investigations of experimental syntax. Even a cursory glance at the experimental syntax literature suggests that significantly more attention has been devoted to the question of the soundness of the data underlying syntactic theory than to the question of what inferences are licensed by the linking hypothesis between data and theory. As we will see in this chapter, we believe that this has been a distraction for the field, as there appears to be no evidence that the existing data is faulty, and growing evidence that traditional collection methods are appropriate for the majority of phenomena of interest to syntacticians. We believe that this distraction can be (at least partially) traced to the second major obstacle to the construction of an integrated theory of language: the difficulty in establishing linking hypotheses between the levels (computational, algorithmic, implementational) of the theory (see also Phillips 1996, Townsend and Bever 2001, Ferreira 2005).

Establishing a level-level linking hypothesis, for example between syntactic theories (computational) and parsing theories (algorithmic), requires the resolution of at least two complex theoretical issues. The first is to determine exactly how much of the sentence processing system should be captured by the syntactic theory; in other words, a line must be drawn that separates the aspects of the processing system that will be abstracted away from in building a syntactic theory, and the aspects of the system that will be directly captured by the syntactic theory. The second issue is to determine exactly what the linking hypothesis will be between the mechanisms in the syntactic theory and the mechanisms in the parsing theory. One early attempt at an integration of syntactic and parsing theories was the *Derivational Theory of Complexity* (DTC) (Miller 1962, McMahon 1963, Miller and McKean 1964, Gough 1965, 1966; for reviews see Fodor, Bever, and Garrett 1974, Berwick and Weinberg 1983, Pritchett and Whitman 1993, Phillips 1996, Townsend and Bever 2001). The DTC assumed an early version of transformational syntactic theory that contained structure building operations (e.g., transformations), but abstracted away from other aspects of sentence processing such as meaning, parsing strategies, probabilistic information, etc. The DTC also assumed an isomorphic linking hypothesis between structure building operations in the syntactic theory and parsing operations in the parsing theory. Under this view, for every transformation that was necessary for a given sentence in the syntactic theory, there was a complementary process in the parsing theory to 'un-do' the transformation during sentence comprehension. In this way, the DTC predicted that behavioral responses that tracked parsing difficulty (such as reaction times) would be directly affected by the number of transformations that were necessary to derive a

given sentence in the syntactic theory, as each transformation would trigger complementary processes during sentence comprehension. As is well known, this prediction did not hold for many types of complex sentences.

The failure of the DTC as a linking hypothesis between syntactic and parsing theories continues to shape the interaction of syntacticians and psycholinguists, as there is some truth to the observation that each side of the computational/algorithmic internalized a different lesson from the failure. Though it is clear that the failure of the DTC was likely due to problems with all three components (the syntactic theory, the parsing theory, and the isomorphic linking hypothesis between the two; see Phillips 1996, Townsend and Bever 2001, Phillips and Lewis 2010), syntacticians tend to be more suspicious of the veracity of parsing theories, and psycholinguists tend to be more suspicious of the veracity of syntactic theories. This latter suspicion, coupled with a long tradition of formal experimentation in psycholinguistics, may be the cause of the increased attention given to the soundness of acceptability judgment data, as unsound data would obviously lead to unsound theories (Edelman and Christiansen 2003, Ferreira 2005, Gibson and Fedorenko 2010a,b). However, as will become clear in the next section, the soundness of acceptability judgment data does not appear to be a true impediment to an integrated theory (see also Phillips and Lasnik 2003, Phillips 2009, and Culicover and Jackendoff 2010); instead, the real impediment seems to be the complexity of the problem, as the space of possible syntactic theories, the space of possible parsing theories, and the space of possible linking hypotheses that can account for the data that we do have are still all relatively large.

3. To what extent are the acceptability judgments underlying syntactic theory sound?

Perhaps the most obvious target of criticism for researchers who are skeptical of syntactic theories is whether the judgments reported in any given paper can be trusted to be a true reflection of the acceptability of the sentences in question. We will call this the *reliability* of judgment data. Establishing the reliability of judgment data is no easy task: the fundamental problem is that, unlike the properties of physical objects, there is no device that can objectively measure the properties of cognitive objects. Instead, cognitive scientists must rely on behavioral experiments to indirectly establish the quantity or quality of the cognitive objects in question. In the case of acceptability, the behavioral experiments in question actually ask the participants to report their judgment of acceptability; however, it should be clear that this report of acceptability is not necessarily the 'true' acceptability response generated by the cognitive

system of language. The process of establishing the reliability of judgment data is actually the process of establishing confidence that the reported values of acceptability accurately reflect the 'true' acceptability response (see also Featherston 2007, Myers 2009a, and Schütze and Sprouse 2011). The question then is how can experimental syntax help establish confidence in the acceptability judgments reported in the syntactic literature.

3.1 Criticisms of the reliability of syntactic data

To begin to see how experimental syntax can help to establish confidence in the data underlying syntactic theories, we can use recent criticisms of syntactic data as a roadmap. Perhaps the most well-known of recent criticisms is that of Gibson and Fedorenko (2010b). Gibson and Fedorenko argue that traditional data collection techniques have led to a preponderance of faulty data in the syntactic literature. As evidence for this, Gibson and Fedorenko discuss three phenomena that were originally reported using traditionally collected acceptability judgments. The first phenomenon is a preference for right-branching relative clauses over center-embedded relative clauses from Gibson (1991), as in (1):

(1) a. *The man that the woman that the dog bit likes eats fish.
b. ?I saw the man that the woman that the dog bit likes.

The second phenomenon is the triple-wh amelioration of the Superiority effect reported by Kayne (1983):

(2) a. *I'd like to know where who hid it.
b. ?I'd like to know where who hid what.

The third phenomenon is a comparison of two sentences involving the Superiority effect from Chomsky (1986a):

(3) a. What do you wonder who saw?
b. *I wonder what who saw.

Gibson and Fedorenko (2010b) re-tested each of these contrasts using formal experiments, and report that all three failed to replicate (i.e., the experiments detected no significant difference between the two conditions in each pair). From these results, Gibson and Fedorenko (2010b)

conclude that the syntactic literature is rife with unreliable data, and that formal experiments are required to correct the situation.

There are at least two fundamental problems with the Gibson and Fedorenko (2010b) studies that we can use to create a roadmap for applying experimental syntax to the question of reliability in syntactic data. The first problem is that we don't know how representative these three phenomena are of the data in the field as a whole. It could be the case that these are three examples from a large set of replication failures in the literature; or it could be that these are three examples from a small set of replication failures. The problem is that these three were chosen with bias (i.e., because they are replication failures). An unbiased test of the replication failure rate in syntax would either test the entire set of data points in the field (as in Sprouse and Almeida (2012), discussed in section 3.2), or test a truly random sample of data points from the entire set (as in Sprouse, Schütze, and Almeida (*submitted*), discussed in section 3.3). One could then compare the number of replication failures to the number of replications to derive a replication failure rate for the field as a whole, and ask whether that rate is substantially higher than the rate in other domains of experimental psychology. Without such comprehensive tests, the Gibson and Fedorenko (2010b) examples are not very informative.

The second problem is that Gibson and Fedorenko (2010b) assume that when traditional methods and formal experiments yield conflicting results, as is the case with their three case studies, we should accept the formal experimental results as "true." This is in many respects begging the question: if the goal of the study is to determine which method is more reliable, then one can't simply assume that one method is a priori more reliable. To illustrate this problem, Sprouse and Almeida (*submitted b*) re-tested the Gibson and Fedorenko (2010b) phenomena using a more powerful judgment task (the Forced-Choice task, see section 3.4), and found that two of the three phenomena do show significant differences identical to those originally reported in the literature: 62 of 98 respondents favored the right-branching structure (1b) from Gibson (1991), p=.006 by sign test, and 58 of 98 respondents favored the triple-wh construction (2b) from Kayne (1983), p=.04 by sign test. In other words, two of the three replication failures reported in Gibson and Fedorenko (2010b) may not be replication failures at all, but instead may be examples of false negatives that arose due to insufficient statistical power in the reported experiments. This suggests that a systematic investigation of the relative statistical power of traditional methods and formal experiments is necessary to determine under what conditions each experiment type should be considered an appropriate tool for assessing acceptability (see section 3.4).

3.2 *The reliability of textbook data*

Given the problems raised by biased selection of phenomena in Gibson and Fedorenko (2010b), Sprouse and Almeida (2012) set out to provide a more accurate estimate of the reliability of data in syntax. They tested all 469 US-English data points from an introductory syntax textbook (Adger 2003) in formal experiments using 440 naïve participants, the magnitude estimation (Stevens 1957, Bard et al. 1996) and yes-no tasks, and three different types of statistical analyses (traditional frequentist tests, linear mixed effects models (Baayen et al. 2008), and Bayes factor analyses (Rouder et al. 2009)). The results of that study suggest that at least 98% of the data points in Adger (2003) replicate using formal experiments. Even following the assumption of Gibson and Fedorenko (2010b) that formal experiments provide the "true" results, this means that the maximum replication failure rate of the traditionally collected judgments from Adger (2003) is only 2%.

3.3 *The reliability of journal data*

Although the replication rate for judgments in Adger's (2003) introductory textbook is impressive (at least 98%), it is logically possible that the replication rate for judgments in journal articles could be substantially lower. To test this possibility, Sprouse, Schütze, and Almeida (*submitted*) identified every (acceptability-judgment-based) data point published in the journal *Linguistic Inquiry* from 2001 to 2010, for a total of 1743 data points. They then randomly sampled 292 data points (forming 146 pairwise phenomena), or about 17% of the full set of data points in *Linguistic Inquiry* 2001-2010. They then tested these 146 phenomena in formal experiments to estimate a replication rate for data points from *Linguistic Inquiry* 2001-2010. They found that 95% of the sampled phenomena replicated using formal experiments. Based on the size of the sample in relation to the full set of data points, this suggests that *Linguistic Inquiry* 2001-2010 has a minimum replication rate of 95% ±5. Taken together with the textbook replication from Sprouse and Almeida (2012), these results suggest that there is no evidence of a reliability problem in the syntax literature, and that the concerns raised by Gibson and Fedorenko (2010b) were empirically unfounded.

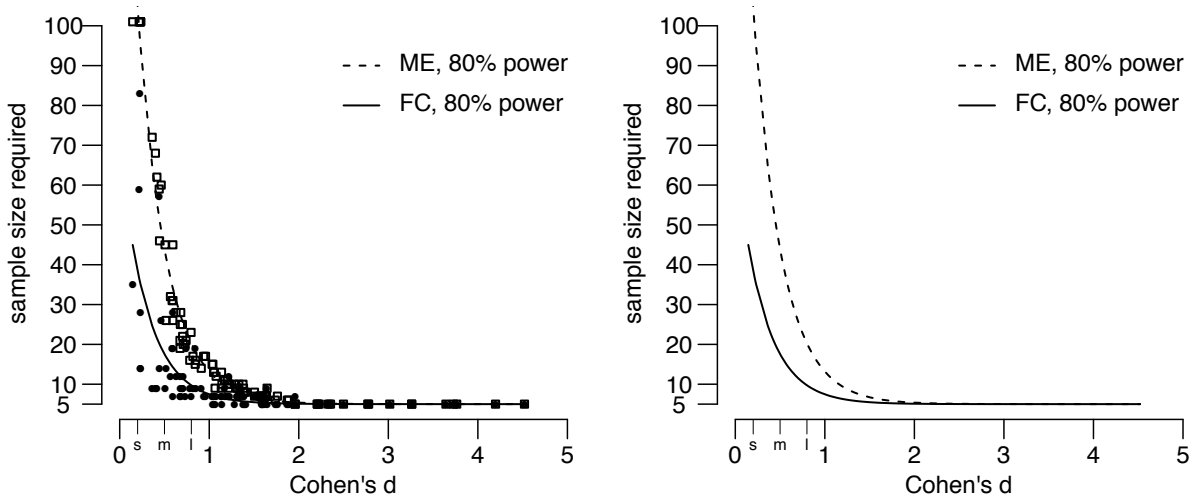
3.4 *A comparison of the statistical power of traditional methods and formal experiments*

Given the role that statistical power played in the interpretation of the Gibson and Fedorenko (2010b) results (two of the three case studies were false negatives due to low statistical power, see Sprouse and Almeida (*submitted a*)), the next logical step is to compare the statistical power of traditional methods and formal experiments. Intuitively speaking, statistical power is the

ability of an experiment to detect a difference between conditions when in fact there is a true difference. Statistical power is often expressed as a percentage: for example, 80% statistical power would indicate that an experiment would detect a true difference 80% of the time. Cohen (1962, 1988, 1992) has suggested that well-powered experiments in psychology should strive for 80%, although in practice most experiments in psychology are closer to 60% (Clark-Carter 1997). Many different factors influence the statistical power of an experiment, from the task chosen, to the size of the difference to be detected, to the number of participants in the experimental sample. As such any study interested in assessing the statistical power of acceptability judgment experiments must manipulate all of these factors to arrive at a comprehensive picture.

Sprouse and Almeida (*submitted a*) conducted just such a study in an effort to directly compare the statistical power of traditional methods and formal experiments. They tested 95 phenomena taken from Adger (2003) and *Linguistic Inquiry* 2001-2010. These 95 phenomena span the full range of effect sizes in syntactic data, allowing for a comparison of statistical power at every possible effect sizes, from very small differences to very large differences. In order to compare the power of traditional methods and formal experiments, the phenomena were tested using two different tasks: magnitude estimation, which is commonly used in formal experiments, and forced-choice, which is commonly used in traditional methods. Over 140 participants were tested using each task, and then resampling simulations were used to empirically estimate the statistical power for each phenomenon at every possible sample size between 5 and 100 participants. Figure 1 below presents the results of these resampling simulations by plotting the sample size required to reach 80% power (along the y-axis) for each effect size (the x-axis).

Figure 1: The sample size required (y-axis) to reach 80% power for the combined set of 95 effect sizes (x-axis) for both forced-choice (solid dots) and magnitude estimation (empty squares) experiments. The solid line is a non-linear trend line for the forced-choice results. The dashed line represents a non-linear trend line for the magnitude estimation results. The criteria for small, medium, and large effect sizes (following Cohen 1988, 1992) are indicated on the x-axis in a smaller font. The left panel includes both points and trend lines; the right panel presents the trend lines in isolation.



Contrary to the claims of critics, the results of the Sprouse and Almeida (*submitted a*) study suggest that traditional methods may be more powerful than formal experimental methods, at least with respect to detecting difference between conditions, as traditional methods require substantially fewer participants to reach 80% power (i.e., the suggested power level in experimental psychology; see Cohen 1988, 1992). Sprouse and Almeida (*submitted a*) also present a discussion of these results in relation to the distribution of effect sizes in syntactic theory (based on the *Linguistic Inquiry* sample from Sprouse, Schütze, and Almeida (*submitted*)), arguing that the results suggest that that the phenomena of interest to syntacticians tend to be substantially larger than those of interest to other areas of psychology (nearly 90% of phenomena are large enough to be visible to the "naked eye" according to the metrics of Cohen 1992), and that traditional methods will lead to at least 94.5% power for the mean effect size (as compared to 59% power for the mean effect size in other areas of psychology, Clark-Carter 1997). In other words, traditional methods should be seen as a valid, reliable, and well-powered set of methods for the investigation of the phenomena of interest to syntacticians.

3.5 *Cognitive bias*

Perhaps one of the most contentious aspects of traditional judgment collection is the use of professional linguists as participants. Several critics of traditional methods have suggested that this introduces the logical possibility of cognitive bias on the part of the participants: as professional linguists, the participants will likely be aware of the theoretical consequences of their judgments, and this awareness may impact the judgments that they ultimately report (Edelman and Christiansen 2003, Ferreira 2005, Wasow and Arnold 2005, Gibson and Fedorenko 2010a,b). Supporters of traditional methods counter this possibility with two logical

arguments. First, acceptability judgment experiments are easily replicable, as they require no special equipment. This means that any given data point can be replicated on the spot: audiences at conferences, reviewers of articles, and even the readership of journals can quickly and easily check the reported judgments for accuracy, and thus identify any influence of cognitive bias. Second, the theoretical awareness of professional linguists may provide a type of expert knowledge that increases the reliability, and possibly the sensitivity, of linguists' judgments over non-linguists' judgments (Newmeyer 1983, 2007, as well as Fanselow 2007, Grewendorf 2007, and Haider 2007 for possible examples in German, and Devitt 2006, 2010, Culbertson and Gross 2009, Gross and Culbertson 2011 for a discussion of what could be meant by 'expert knowledge'). The empirical question then is whether there is any evidence of cognitive bias in the judgments of professional linguists.

There are at least two methods for assessing the role of cognitive bias in syntactic theory. One method would be to compare the judgments of linguists with differing theoretical dispositions to see if theoretical knowledge is indeed affecting their judgments. To our knowledge, the only study to directly compare judgments from linguists with differing theoretical dispositions is Dabrowska (2010). Dabrowska compared the judgments of self-identified generative linguists with self-identified functional linguists in a rating study that included Complex NP islands (**What did John make the claim that Mary bought?*). One plausible prediction of the cognitive bias hypothesis is that generative linguists would rate the examples of island violations lower than the functional linguists because island constraints are a core part of the generative theory of syntax, whereas several functional linguists have argued that island effects are an epiphenomenon of language use (e.g., Kuno 1973, Deane 1991, Kluender and Kutas 1993, Goldberg 2007). In other words, generative linguists have a motivation to confirm the reliability of Complex NP islands, whereas functional linguists have a motivation to disconfirm the reliability of Complex NP islands. Dabrowska (2010) actually found that generative linguists' ratings of Complex NP islands were higher (i.e., more acceptable) than the functional linguists' ratings, which, contrary to the cognitive bias hypothesis, suggests that the generative linguists were actually biased *against* their own theoretical interests[i].

The second method is to compare linguists' judgments with the judgments of naïve participants. This method is necessarily more complicated, as a simple difference between two groups is never enough to establish which is correct, and as we've already seen, there are real statistical power differences between the two methods. One plausible prediction of the cognitive bias hypothesis would be that linguists' judgments would go in the opposite direction than naïve

participants' when it is theoretically advantageous to do so. If this prediction holds, then large-scale comparisons of the two groups, such as the studies by Sprouse and Almeida (2012) and Sprouse, Schütze, and Almeida (*submitted*), should reveal a large number of sign-reversals – that is, instances where the direction of the difference between two conditions reported by one group is the exact opposite of the direction of the difference reported by the other group. Out of the 250 pairwise phenomena investigated by these studies, only two sign-reversals were observed (less than 1% of cases), suggesting that cognitive bias has not had an influence on existing syntactic data.

3.6 The interpretation of variation across participants

Another issue that arises in the critical literature (e.g., Wasow and Arnold 2005, Gibson and Fedorenko 2010b), but often goes unexamined, is the question of how to interpret the variability in acceptability judgments across participants. To make this discussion concrete, imagine that a researcher is interested in the difference between two sentence types, as is typical in acceptability judgment experiments. In general statistics terminology, this difference is the *treatment effect*. When investigating the treatment effect, the researcher can ask if the sample as a whole shows a treatment effect by comparing the mean response of the sample for each condition. If the two means are different enough, traditional statistical significance testing (SST) will report that there is a significant treatment effect for the sample. However, finding a statistically significant treatment effect for the sample does not mean that every participant demonstrated the treatment effect. In practice, given sufficient statistical power, very few participants need to show the treatment effect in order for the sample to show a significant treatment effect. A recurring question in the experimental syntax literature is what to make of this variability. If 100% of the participants show the treatment effect, it is pretty clear that the effect is a robust fact for all of the members of the sample. However, what if 75% show the effect, and 25% do not? What if only 25% show the effect, and 75% do not?

There seem to be three different approaches to this question in the literature:

1. Since measurement involves noise, only the central tendency of the sample matters, and it is expected that not every participant or every item in the sample show the treatment effect.

2. If a large enough proportion of participants do not show the predicted treatment effect, this might be evidence for a different dialect.

3. Given a strong theory-data linking hypothesis that ungrammatical sentences should be overwhelmingly judged to be unacceptable, a large enough proportion of participants that fail to show the predicted treatment effect, or that judge supposedly ungrammatical sentences no worse than awkward will be taken as evidence that the theoretical prediction is disconfirmed.

The first approach assumes that the participants who do not show the treatment effect are simply being influenced by random noise. This is the default assumption in most domains of cognitive science, as it is assumed that all behavioral responses are the result of a combination of the experimentally manipulated behavior, and various sources of random noise (sometimes called unsystematic variation). Under this approach, it only matters whether the sample as a whole shows the treatment effect: if SST reveals a treatment effect in the sample, then there is a real treatment effect. This is by far the most common approach in the experimental syntax literature, as many of the best practices of experimental syntax, including the use of SST, have been directly adapted from experimental psychology. A second approach is to investigate whether participants who do not show the treatment effect are actually drawn from a different population than the participants who do show the effect. In most domains of cognitive science, the population of interest is all humans; in linguistics, the population of interest is all speakers of a given language. It is always a logical possibility that the participants who do not show an effect have a different grammar than the speakers who do show the effect (see also den Dikken et al. 2007). A third approach is to assume that only manipulations that yield a treatment effect in (almost) 100% of participants are real. While this is certainly a strong criterion to impose on experimental results, it is not without a certain logic. It is common in the syntactic literature to talk about *possible* sentences and *impossible* sentences. If one truly believes that a given sentence is impossible in a certain language, then one could also conclude that no amount of random noise should be enough to cause participants to rate that sentence as acceptable (see also Hoji 2010). This approach nonetheless makes several assumptions: (i) that acceptability judgments directly reflect the grammaticality of the sentences, without contamination from other cognitive systems, (ii) that fatigue and distraction do not affect judgments, (iii) that the crucial analysis is categorical (sentence A is acceptable or unacceptable), as opposed to relative differences (A is better or worse than B), and (iv) that the empirical domain of syntactic theory is only the difference between possible and impossible sentences.

Because of the domain-specific issues related to syntactic theory (e.g., language

variation, possible/impossible sentences), it is critical to keep these three approaches to variation in mind when interpreting the results of formal acceptability judgment experiments. Failure to do so can lead to substantially different interpretations of the data. For example, Langendoen et al. (1973) investigated the claim made by Fillmore (1965) and others that the first object of a double-object construction cannot be questioned:

(4) *Who did you buy a hat?
(cf. What did you buy Mary?)

Langendoen et al. performed an answer completion task to test this claim formally. They asked 109 students to answer questions like (7) with a complete sentence:

(5) Who did you show the woman?

Their hypothesis was that if questions like (4) are indeed unacceptable, then the answers to (5) should *consistently* place the answer NP at the end of the sentence (*I showed the woman my daughter)*. Langendoen et al. reported two findings: that one-fifth of the participants responded with the NP in the first object position *(I showed my daughter the woman)* and these participants were all from the metropolitan New York City area. Langendoen et al. (1973) considered following approach two, i.e, concluding that there are two dialects at work in the sample: speakers from NYC, who can question first object, and everyone else, who cannot. Their favored conclusion, however, was more nuanced. Noticing the theoretical difficulty in incorporating the necessary restrictions in the grammar of English to explicitly rule out questions from the first object of a double objet construction (a point previously raised by Jackendoff and Culicover, 1971), Langendoen et al. (1973) proposed that these constructions are in fact licensed by the grammar of English. The difference between the population that finds them acceptable and the population that does not, they argue, is due to a different parsing strategy employed by the two groups. Taking a different perspective, Wasow and Arnold (2005) and Gibson and Fedorenko (2010a,b) have interpreted Langendoen et al (1973)'s result under approach three, and concluded that Fillmore's (1965) original claim is incorrect: it is, in fact, possible to question first objects. Of course, it is also possible to assume approach one: the one-fifth of participants who created first object answers did so because of random noise in the experiment. This means that 87/109 participants responded in accordance with Fillmore's (1965) claim. A one-tailed sign test yields $p = .0000000018$ – a significant result. What should

be obvious here is that the problem is not with the data itself, since no experimental result disputed the fact that, by and large, speakers of English found questions constructed from the first object of a double object construction to be unacceptable. The problem is with the interpretation of what these results might mean for the theory of grammar: Langendoen et al. (1973)'s favored interpretation was motivated first and foremost by theory-internal considerations, while Wasow and Arnold (2005)'s and Gibson and Fedorenko (2010a,b)'s conclusions were dictated by their data-theory linking hypothesis.

A similar situation arises with Wasow and Arnold's (2005) test of a claim from Chomsky (1955/1975) that the complexity of a noun phrase strongly determines the position of that noun phrase within a verb-particle construction. Chomsky's claim is twofold. First, he claims that the most natural place for multi-word NPs is after the particle, therefore both (a) and (b) below should be more acceptable than both (c) and (d). Second, he claims that complex NPs (relative clauses) are less acceptable than simple NPs when they occur between the verb and particle, therefore (d) should be less acceptable than (c).

(6) a. The children took in all our instructions. [3.4 out of 4]
b. The children took in everything we said. [3.3 out of 4]
c. The children took all our instructions in. [2.8 out of 4]
d. The children took everything we said in. [1.8 out of 4]

Wasow and Arnold (2005) ran a formal rating experiment, the results of which are in square brackets in (6). According to approach one, which assumes that only a difference in means is necessary to verify a claim, the formal results match Chomsky's informal results perfectly: there is a significant interaction between particle position and NP type ($p < .001$). However, Wasow and Arnold (2005, p. 1491) interpret the results as problematic because "17% of the responses to such sentences [d.] were scores of 3 or 4." It seems that Wasow and Arnold (2005) were assuming approach three, which requires that sentences be judged unacceptable close to 100% of the time if we are to accept their status as unacceptable.

It is crucial for the language community to be explicit about their assumptions regarding the variability of acceptability judgments moving forward. As we have already seen, several high profile criticisms of informal experiments rest upon the assumption that there should be little or no variability among participants (Wasow and Arnold 2005, Gibson and Fedorenko 2010b), but it is important to notice that when this very strong assumption between the relationship between the theory and the data is relaxed (for instance, to allow for things like

sampling error), the exact same set of results can be seen as providing strong evidence for the opposite conclusion (see also Labov (1996) for a similar discussion of the *wanna* contraction, and Raaijmakers (2003) for a similar discussion of the interpretation of variation across participants in the sentence processing literature).

3.8 The relative costs and benefits of traditional methods and formal experiments

In this section we have seen that experimental syntax techniques provide a useful toolkit for exploring precisely which properties of formal experiments should increase our confidence in the veracity of the results (i.e., whether there is indeed a true difference between conditions). Though it is relatively common to assume that formal experiments provide 'better' results than informal results, the current state of the field suggests that many of the perceived benefits of formal experiments ultimately disappear under closer empirical scrutiny. This raises the very real possibility that the problem facing acceptability judgment data is a sociological one, not an empirical one: researchers who are accustomed to formal experiments are disinclined to have confidence in the results of traditional methods, regardless of whether the informal experiments are empirically appropriate for the research questions that they are intended to address.

Choosing the appropriate methodology (in any field) requires the researcher to balance the costs and benefits of different methodologies relative to their specific research question. The benefits of the traditional methods over formal experiments are well known: (i) traditional methods are cheaper – formal experiments cost $2.20-$3.30 per participant on AMT; (ii) traditional methods are faster, at least with respect to participant recruitment – although AMT has diminished this advantage significantly (e.g., Sprouse 2011a reports a recruitment rate of 80 participants per hour on AMT); (iii) the tasks used in traditional methods, such as the forced-choice, appear to more powerful than the tasks used in formal experiments, such as magnitude estimation; and (iv) this increased statistical power often makes traditional experiments the only option for languages with few speakers (Culicover and Jackendoff 2010) or for studies of variation between individuals (den Dikken et al. 2007). On the other hand, the benefits of formal experiments typically revolve around the types of information that are necessary to answer the research question of interest (see also Section 4). For example, the numerical rating tasks typically used in formal experiments provide more information than the forced-choice and yes-no tasks used in traditional methods, such as the size of the difference between conditions (see also Schütze and Sprouse 2011, though as Myers 2009b points out, non-numerical tasks can be used to approximate size measurements if necessary). Furthermore, if one wishes to construct a complete theory of the gradient nature of acceptability judgments, an enterprise which has

gained in popularity over the past decade (e.g., Keller 2000, Featherston 2005b) then one will clearly need numerical ratings of acceptability. The bottom line is that there is no single correct answer when it comes to choosing a methodology. Syntacticians (and indeed all researchers) must be aware of the relative costs and benefits of each methodology with respect to their research questions, and be allowed to make the decision for themselves. Science cannot be reduced to a simple recipe.

4. What types of inferences are licensed by the linking hypothesis between acceptability judgments and syntactic theory?

The majority of experimental syntax studies have focused on the reliability of the data underlying syntactic theory. As the previous section made clear, we believe that this has been a (necessary) distraction: there appears to be no evidence that the existing data is faulty, and growing evidence that the informal methods are appropriate for the majority of phenomena of interest to syntacticians. However, there is reason to believe that experimental syntax techniques also provide new tools to investigate the inferences licensed by the linking hypothesis between acceptability judgments and syntactic theory. In this section we will review two ways in which experimental syntax has added to our understanding of the nature of syntactic theory: (i) testing reductionist claims about the correct locus of acceptability judgment effects (so-called "processing" explanations), and (ii) examining the complex theoretical issues surrounding the interpretation of continuous acceptability judgments.

4.1 Reductionist approaches to acceptability judgment effects

The fundamental component of the linking hypothesis between judgment data and syntactic theories is the assumption that manipulations of the structural properties of a sentence will lead to modulations of acceptability. Regular readers of the syntactic literature are aware that it is relatively common for syntacticians to establish the structural nature of acceptability differences; by holding non-syntactic factors constant (semantics/plausibility, phonetics/phonology, morphology/lexical properties), syntacticians can be relatively certain that it is the structural manipulation that is driving the effect. However, because acceptability judgments are the result of successful sentence processing, and because the operation of the parser is (by definition) intricately tied to structural properties of sentences, there is always a possibility that acceptability effects may be driven by properties of the parsing system rather than grammaticality per se (cf. the conclusion reached by Langendoen et al. 1973 for the

questions based on the double dative construction mentioned in the previous section). The question then is how experimental syntax techniques can help tease apart acceptability differences due to grammaticality effects, and acceptability differences due to properties of the parsing system.

The first step in teasing apart this ambiguity is to be clear about what is meant when one suggests that acceptability differences are driven by properties of the parsing system. These types of accounts are sometimes called "processing explanations" to contrast with "syntactic explanations," but as several researchers have remarked, this label is less than ideal (Phillips 2011, Sprouse et al. 2012). The problem with this label is that, by definition, structural manipulations will result in different behavior by the parser. This is precisely what we want: the theory of syntax (a computational theory of structural properties of sentences) should be closely related (by a level-level linking hypothesis) to the theory of parsing (an algorithmic theory of syntactic structure building). Viewed from this perspective, every "syntactic explanation" is a "processing explanation", as the syntactic theory is a form of abstraction or idealization of the structure-building component of the parser. Because of this tight relationship between the syntactic theory and the structure-building component of the parser, the syntactic properties of a sentence will necessarily affect the behavior of the syntactic structure-building component of the parser. This means that in order for the so-called "processing explanations" to be distinct from "syntactic explanations", the "processing explanation" must not be related to the syntactic structure-building component of the parser.

To clarify the content of these types of questions, Phillips (2011) and Sprouse et al. (2012) suggest the term *reductionist* instead of "processing explanation". They argue that the logic of these types of explanations is clearly reductionist, in that the acceptability effect is argued to be reducible to non-structure-building components of the parser, such as parsing strategies or parsing resource capacity. Under a reductionist approach, the relationship between the acceptability effect and the structural manipulation that is normally assumed to be driven by the syntactic system is actually epiphenomenal; the true causal nexus lies between the extra-syntactic factors, such as parsing strategies or parsing resources, and the acceptability effect. The second order correlation between acceptability and the structural manipulation arises because structural manipulations necessarily affect parsing strategies or parsing resource allocations. In this way, the complexity of the syntactic system is reduced in favor of extra-syntactic components. To the extent that these extra-syntactic components are independently necessary. *reductionist explanations* may be preferred to syntactic explanations according to theory building metrics such as Occam's razor.

As a concrete example, take the two island effects that we have discussed previously: Whether islands and Complex NP islands (see also Alexopoulou and Keller 2007, Sprouse 2008, and Sprouse et al. 2011 for other examples of the parsing system affecting acceptability judgments). The standard analysis within the syntactic literature is that these island effects are rated unacceptable by native speakers because there is a syntactic constraint, such as the Subjacency Condition, that rules these structures out as ungrammatical. However, several researchers have proposed alternative explanations that do not involve syntactic constraints at all, but rather potentially independently motivated properties of the parsing system such as working memory (Kluender and Kutas 1993, Kluender 1998, 2004, Hofmeister and Sag 2010), attention (Deane 1991), and focus (Erteshik-Shir 1973, Goldberg 2007). For example, Kluender and Kutas (1993) argue that island violations such as (2b) and (3b) are in fact grammatical structures, but that the unacceptability reported by speakers is in fact the result of a combination of two relatively resource-intensive processes that are necessary to successfully parse the sentences. These two processes require more resources than are available to the parsing system, and therefore cause the parsing system to fail to successfully parse the sentences, resulting in the perception of unacceptability.

Kluender and Kutas (1993, see also Kluender 1998, 2004) are very explicit about the two processes that they believe are the cause of the unacceptability, and about the resources in question. They argue that the first process is the maintenance of a displaced wh-word in working memory during the processing of the sentence between the wh-word and the downstream gap site. The second process is the construction of the island structure itself, which as can be seen in (2b) and (3b) involves a CP clause that is in some ways more complex than CPs headed by *that*. Kluender and Kutas (1993) argue that each of these processes requires a certain amount of working memory resources to be deployed. Although each of these processes can be deployed in isolation, when deployed simultaneously, the combined resource requirements are greater than the pool of available resources. At this point it should be clear that this reductionist theory defines island effects as a *psychological interaction* of two (sets of) parsing processes that occurs because the processes rely upon a single pool of resources. Sprouse et al. (2012) suggest that this psychological interaction can be translated into a *statistical interaction* between two factors, each with two levels: LENGTH (short, long) and STRUCTURE (non-island, island) (see also Myers 2009b for a discussion of the use of factorial designs in syntax). The factor LENGTH manipulates the length of the wh-dependency at two levels: within a bi-clausal constituent question, a short dependency is created by extraction of the matrix subject; a long dependency is created by extraction of the object argument in the embedded clause. The factor STRUCTURE

refers to the STRUCTURE of the embedded clause.

LENGTH	STRUCTURE	Example
short	non-island	Who __ thinks that John bought a car?
long	non-island	What do you think that John bought __?
short	island	Who __ wonders whether John bought a car?
long	island	What do you wonder whether John bought __?

Table 1: Independent manipulation of dependency length and island structures

Defining island effects in this way has several advantages. First, it allows us to isolate the effect of each of the individual factors on continuous acceptability ratings. For example, the effect of processing long-distance wh-dependencies can be seen by comparing the short, non-island condition to the long, non-island condition, and the effect of processing island structures can be seen by comparing the short, non-island condition to the short, island condition. Second, it allows us to quantify the statistical interaction of the two factors. If there were no statistical interaction between the two factors (i.e., if the two sets of processes impact acceptability ratings independently), we would expect a graph like that in Figure 2a. Figure 3a is an example of simple linear additivity between each factor in which the cost of each process leads to a decrement in acceptability ratings, and in which each cost sums linearly with respect to the short/non-island condition. This linear additivity in decrements leads to two parallel lines. However, if there were an interaction between the two factors, we would expect a graph like that in Figure 2b: super-additivity when the *long* and *island* levels of the two factors are combined, leading to non-parallel lines. (The hypothetical ratings in Figure 2 are displayed in terms of standardized z-scores, which can be derived from any approximately continuous rating measure, such as Likert scales or magnitude estimation.)

a. b.

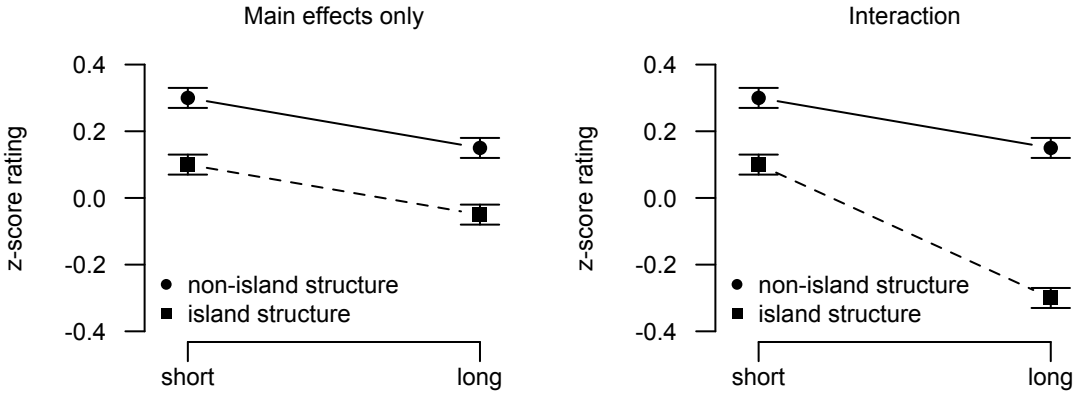


Figure 2: Example results for main effects and interaction

Figure 2b is in fact the pattern that is consistently observed when the factors LENGTH and STRUCTURE are independently manipulated in acceptability experiments, although there is variation in the size of the effect of island structures alone (Kluender and Kutas 1993, Sprouse 2007a, Sprouse et al. 2012). In this way, island effects can be defined as a statistical interaction between two structural factors, exemplified by a super-additive decrease in acceptability in the long, island condition.

Sprouse et al. (2012) used this interaction-based definition to test the role of working memory resources in the acceptability of island effects. They argued that the Kluender and Kutas (1993) analysis would predict an inverse relationship between working memory capacity in individuals and the strength of the super-additive interaction because the super-additive interaction (as opposed to linear additivity) arises due to insufficient working memory capacity. Therefore one would expect participants with higher working memory capacity to have smaller interactions, and participants with lower working memory capacity to have larger interactions. Sprouse et al. (2012) tested over 300 participants with two different working memory tasks and two different acceptability judgment tasks (Likert scales and magnitude estimation), and found no relationship between working memory capacity and the size of the super-additive interaction. From these results, they conclude that it is unlikely that working memory capacity is driving the unacceptability of island effects. In this way, a combination of factorial designs, numerical acceptability judgment tasks, and parsing resource tests (working memory tests, etc.) can be used to investigate the probability of reductionist versus grammatical explanations for acceptability effects, and help establish the data-theory linking hypothesis between acceptability judgments and syntactic theory.

Though this is just one example of using experimental syntax to investigate a

reductionist explanation for acceptability effects, it does suggest a general methodology for future studies. The first step is to identify a set of (non-structure-building) parsing-related factors that are hypothesized to drive the effect. The simplest possible reductionist explanation is one in which the acceptability effect of each of the factors will sum (linearly) to the full effect (e.g., the lack of interaction in Figure 2a). If the factors do not sum (linearly) to the full effect (e.g., the super-additive interaction in Figure 2b), then there must be an explanation for the super-additivity. In the case of island effects, the super-additive interaction was explained by the assumption that the two processes draw on the same limited pool of working memory resources. The explanation for the super-additivity can then be tested by searching for correlations between the strength of the interaction and the relevant parsing properties (e.g., working memory capacity).

4.2 Gradient acceptability and the nature of syntactic theory

Although it is possible to categorize sentences as either acceptable or unacceptable in qualitative tasks such as the yes-no task, the results of numerical judgment tasks such as magnitude estimation have suggested that acceptability is better described by a continuous scale, with sentence types taking values at any point along the scale. The fact that acceptability is a continuous measure has led several researchers to investigate to what extent the nature of the grammar itself may be continuous (Keller 2000, Sorace and Keller 2005, Featherston 2005b, Bresnan 2007). It is not uncommon to encounter those who believe continuous acceptability necessitates a continuous (or gradient) syntactic system. However, there is no necessary link between the nature of acceptability and the nature of the syntactic system. The question in fact hinges on (at least) three complex theoretical issues: (i) What is the relationship between acceptability judgment data and syntactic theories? (ii) What is the correct level of abstraction for a computational theory? (iii) What is the continuous syntactic property that could give rise to continuous acceptability?

Up to this point, the discussion of the relationship between acceptability judgment data and syntactic theories has been one of inference from data to theory: when certain extra-syntactic factors (such as parsing strategies, parsing resources, and issues related to acceptability judgment tasks themselves) are held constant, modulations in acceptability judgments are interpreted (via a linking hypothesis) as evidence about the properties of the syntactic system. However, it is also possible to reverse the direction of the relationship and investigate how well the syntactic theory predicts the acceptability judgment data. While this approach is more common in the computational modeling literature, where the term *generative* is used to refer to

models that can be used to generate the observable data, it seems clear that the question of how best to account for the continuous nature of acceptability judgments is a predictive/generative question. This can be seen in the structure of a common argument that syntactic theories should be gradient: (i) acceptability is continuous, (ii) categorical grammars predict categorical acceptability, (iii) gradient grammars predict continuous acceptability, (iv) therefore the grammar must be gradient. Crucially, this style of argument assumes that the syntactic theory is the correct locus for the mechanisms that lead to continuous acceptability. It is clear that this is not a logical necessity: all of the extra-syntactic components of the language faculty that must be controlled in order to make careful inference from acceptability data to syntactic theory (such as parsing strategies, parsing resources, and the conscious mechanisms that underlie judgment tasks) may be contributing to the gradience in the acceptability judgment data, and therefore are potential sources of the mechanisms that generate continuous acceptability. In short, once one adopts a predictive/generative approach to modeling acceptability judgment data, the question is to what extent should the continuous mechanisms be part of the syntactic theory, and to what extent should the continuous mechanisms be part of the extra-syntactic components of the language faculty.

In order to determine to what extent the syntactic theory should predict continuous acceptability, we must be explicit about what a syntactic theory is a theory of, and which aspects of that theory can give rise to continuous acceptability. In section one, we presented a view of syntactic theory as a computational level description of a part of the human language faculty. In other words, syntactic theory is a formulation of the properties of the syntactic structure building mechanisms of the human parser that abstracts away from parsing strategies, parsing resources, and other issues that are specific to the real-time implementation of parsing algorithms. From this perspective, there are only two options for including gradient mechanisms within the syntactic theory. The first option is to actually abstract away form the algorithmic level less, such that one or more of the gradient mechanisms of the algorithmic level actually exists in the computational level description. Bresnan (2007) has advocated a probabilistic approach to the syntax of the dative construction in English that may be an example of this sort of 'weaker' abstraction, as the syntactic theory appears to include information about the probability of the dative construction under various morphosyntactic, semantic, and information-structure environments – information that has previously been of primary interest to algorithmic level sentence processing theories (see also Bresnan and Hay 2008, Bresnan and Ford 2010).

The second option is to include an additional property in the syntactic theory that can capture gradience. Once again, this is a complex ontological issue, as any property that is

included in the syntactic (computational) theory must be mapped to the actual language faculty as it is implemented in the human brain (i.e., there is a *mentalistic* commitment). This leads to a stark contrast between syntactic theories in which *grammaticality* is a purely theoretical construct, and syntactic theories in which *grammaticality* is a mentalistic construct. In the former, *grammaticality* is not a property that is available to the mental system, but rather a label that theoreticians can apply to sentences rather than using the non-technical terms 'possible/impossible' (e.g., Chomsky 1957). In the latter, *grammaticality* is a property that is available to the mental system in some form, such as when different syntactic constraints are assumed to lead to different levels of unacceptability (e.g., Huang 1982, Chomsky 1986a), or when structures are assumed to be marked as ungrammatical (e.g., the star feature in Chomsky 1972a). As Keller (2000) demonstrates for Optimality Theory, and Featherston (2005b) demonstrates for the Decathlon model, syntactic theories that assume that grammaticality is a mentalistic construct can be used to directly predict continuous acceptability without altering the level of abstraction in the computational/algorithmic divide; however, the cost is assuming that grammaticality is a mentalistic construct rather than simply a theoretical one.

On the one hand, it is clear that experimental syntax techniques, especially numerical tasks like magnitude estimation and Likert scales, yield a new form of continuous acceptability data that provide researchers with the opportunity to reverse the normal direction of data-theory inference, and construct predictive/generative syntactic theories. On the other hand, the discussion in this subsection suggests that the interesting questions raised by this approach are not data-driven questions. In other words, the data enable this line of questioning, but the data don't determine the answer (see also Sprouse 2007b). The questions raised in this section (such as: What is the right level of abstraction for a computational theory? and Should syntactic theories include a gradient, mentalistic property called grammaticality?) are theoretical questions. And like all theoretical questions, they can only be answered through careful comparison of the empirical adequacy of competing theories.

5. Conclusion

Our goal in this chapter was to review the role of experimental syntax in the construction of an integrated cognitive science of language. While this role is undoubtedly still evolving, it seems clear that experimental syntax is well positioned to make substantial contributions to two questions that are central to the integration of syntactic theories with parsing theories: *Is the data underlying existing syntactic theories sound?* and, *What types of inference are licensed by the*

linking hypothesis between acceptability judgments and syntactic theories? Although both questions are scientifically relevant to the theory, the current state of evidence suggests that questions about the reliability of existing judgment data may have been a (historically driven) distraction: there appears to be no evidence that the existing data is faulty, and growing evidence that the traditional methods are appropriate for the majority of phenomena of interest to syntacticians. This suggests that the contribution of experimental syntax in the coming years will be as a tool for investigating what the acceptability judgment data reveals about the nature of syntactic theory. We have seen two examples of this approach in this chapter: the question of reductionist approaches to complex syntactic phenomena (e.g., island effects), and the question of gradient approaches to syntactic theory. Undoubtedly there are more questions waiting to be discovered as the field progresses toward an integrated theory of language.

NOTES

[i] Dabrowska (2010) interpreted this to be the result of different frequencies of exposure to specific kinds of unacceptable sentences. Since the training of generative linguists include reading scores of textbooks' and articles' examples of specific ungrammatical sentences, this could lead to a higher familiarity with them, which might lower generative linguists' sensitivity to the unacceptability of these sentences. It is important to note, however, that this kind of explanation systematically predicts the *opposite* of the cognitive bias hypothesis for the phenomena studied by syntacticians: judgments of generative linguists are going to be systematically *less* sensitive to the predicted contrasts than the judgments of naïve participants.