Tom S. Juzek* and Jana Häussler*

Data convergence in syntactic theory and the role of sentence pairs

https://doi.org/10.1515/zfs-2020-2008
Received July 31, 2019; accepted January 20, 2020; published online March 27, 2020

Abstract: Most acceptability judgments reported in the syntactic literature are obtained by linguists being their own informants. For well-represented languages like English, this method of data collection is best described as a process of community agreement, given that linguists typically discuss their judgments with colleagues. However, the process itself is comparably opaque, and the reliability of its output has been questioned. Recent studies looking into this criticism have shown that judgments reported in the literature for English can be replicated in quantitative experiments to a near-perfect degree. However, the focus of those studies has been on testing sentence pairs. We argue that replication of only contrasts is not sufficient, because theory building necessarily includes comparison across pairs and across papers. Thus, we test items at large, i. e. independent of counterparts. We created a corpus of grammaticality judgments on sequences of American English reported in articles published in *Linguistic Inquiry* and then collected experimental ratings for a random subset of them. Overall, expert ratings and experimental ratings converge to a good degree, but there are numerous instances in which ratings do not converge. Based on this, we argue that for theory-critical data, the process of community agreement should be accompanied by quantitative methods whenever possible.

Keywords: experimental syntax, data convergence, grammaticality judgments, acceptability judgments tasks, introspection

1 Introduction

Linguists being their own informants is one of the main means of data collection in syntactic theory. This practice is commonly referred to as *researcher introspection*. In the wake of Schütze's seminal work (1996), there has been a debate about

***Corresponding author: Tom S. Juzek,** Sonderforschungsbereich 1102, Universität des Saarlandes, Saarbrücken, Germany, e-mail: tom.juzek@posteo.net
***Corresponding author: Jana Häussler,** Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Bielefeld, Germany, e-mail: jana.haeussler@uni-bielefeld.de

the adequacy and reliability of researcher introspection (among others Bard et al. 1996; Schütze 1996; Edelman and Christiansen 2003; den Dikken et al. 2007; Culicover and Jackendoff 2010). While several researchers have voiced their concerns about the reliability of it (Wasow and Arnold 2005; Featherston 2007; Gibson and Fedorenko 2010; Gibson and Fedorenko 2013; Gibson et al. 2013), others have defended researcher introspection, citing that it has proven itself to be reliable for most purposes and that there are no reasons to assume that quantitative methods give better results (Phillips and Lasnik 2003; Bornkessel-Schlesewsky and Schlesewsky 2007; Grewendorf 2007; Phillips 2010; Sprouse and Almeida 2012; Sprouse and Almeida 2013; Sprouse et al. 2013).

The present paper argues that in the case of American English and other well-represented languages, researcher introspection is best thought of as what we call a *process of community agreement*. The paper provides quantitative evidence for why over-relying on this process leads to issues with data convergence, that is the question to which degree judgments from the process of community agreement and non-expert judgments from quantitative judgment tasks agree. Sprouse et al. (2013) reported a near-perfect match between author judgments and judgments crowdsourced by Sprouse et al. However, Sprouse et al. focused on the analysis of sentence pairs, directly comparing each pair of an ungrammatical and/or unacceptable sentence (marked with "*", "?", etc.) and its grammatical and/or acceptable counterpart.[1] Testing sentence pairs comes with certain other methodological choices and we argue that such choices can have quite an impact on one's results.

Our primary aim is to provide a different perspective on the debate. We also compare expert and non-expert judgments, but we compare items *at large*, i. e. we do not restrict our analyses to the analysis of sentence pairs. We argue that a comparison at large is an appropriate way of assessing data convergence. For this, we depart from Sprouse et al. (2013) in various other methodological aspects. We carefully control our items for confounding factors, that is we only test items that do not involve ambiguity, extreme complexity and the like, as such factors might affect acceptability judgments. This allows us to treat perceived acceptability by the non-experts as a reasonable approximant of grammaticality. Our data include a considerable number of cases in which author judgments and crowdsourced judgments do not converge.

1 We say "and/or", because under the current standard use, an asterisk can denote any of the following three possibilities: ungrammatical and unacceptable; grammatical but unacceptable; or acceptable but ungrammatical. For a discussion of this issue, see Section 6.

Overview
Section 2 provides the necessary background. In Section 2.1, we provide terminological and conceptual clarifications and argue that one of the main methods of data collection is a process of community agreement, at least for languages like English. In Section 2.2, we compare the process of community agreement to quantitative methods, and acceptability judgments tasks in particular. Section 2.3 discusses the notions of acceptability, grammaticality and grammatical reasoning, and reviews their relation to performance. In Section 2.4, we examine the role that sentence pairs play in syntactic theory. As we will see in the following sections, whether one assumes that syntactic research is restricted to the analysis of sentence pairs has severe implications for how one approaches the debate on data convergence. Sprouse et al. (2013) had a considerable impact on that debate and we revisit their work in Section 3. We have created a corpus of author judgments and in Section 4, details about the corpus are given. Section 5 presents our experiment and our findings, whose validity and implications we discuss in Section 6. Section 7 concludes this paper.

2 Core aspects in the debate on acceptability judgments and data convergence

In this section, we review some background needed to better understand the current discussion on the empirical foundation of linguistics. This includes the distinction between the process of community agreement and quantitative methods, the distinction between acceptability, grammaticality, and grammatical reasoning, and the question of what role sentence pairs play in syntactic research.

2.1 Introspection and the process of community agreement

One of the main methods of data collection in syntactic theory is commonly called *introspection* and a key feature is that the roles of researcher and informant are conflated. However, using the term "introspection" like this is problematic because, strictly speaking, every acceptability judgment stems from introspection, no matter whether produced by a linguist being their own informant or by a participant in a quantitative experiment. Sometimes "researcher introspection" is preferred as a term, but we do not think that it is a better term than just "introspection", because the process is more complex than a single researcher providing their own data (as also noted by Phillips 2010; Linzen and Oseki 2018). In reality, and certainly for well-represented languages, self-generated data are just the

starting point. Syntactic data will undergo what we call the *process of community agreement*: Most linguists will ask colleagues and students before submitting a paper, probably even before writing it. If their intuition lacks support, the data will often not make it into a paper. Reviewers, and later-on readers will voice their concerns when they do not share certain intuitions. Subsequently, any judgment drawn from researcher introspection will be confirmed or questioned by further members of the linguistic community, or their respective sub-communities. The notion of a sub-community is important in this context: Arguably, for the process, there is not a single community; and even within the different linguistic subfields, there might be several communities that discuss and publish in separate venues.

The process of community agreement can go as far as that a judgment in a given paper might completely draw on the literature and there is little judgment by the paper's author(s). At times, authors simply *sanction* or *reinforce* the commonly accepted status of a phenomenon. To a good extent, the same process applies to sentences or syntactic constructions that are being discussed for the first time in the literature. While the judgment might be based on researcher's introspection, it is likely that this was only a first step. Again, the process will probably involve checks by colleagues, students, and reviewers.

To illustrate this process, consider the following example of superiority violations. Our corpus based on *Linguistic Inquiry* 2001–2010 contains among others the following examples taken from Clifton et al. (2006).

(1) a. **What did who buy there?*
b. **What do you expect who to buy?*

The asterisks in (1a) and (1b) do not (exclusively) reflect Clifton et al.'s introspective judgments. Clifton et al. (2006) credit the observation in (1a) to Kuno and Robinson (1972) and Chomsky (1973), the one in (1b) to Hendrick and Rochemont (1982) and Hornstein (1995). Over the years, many members of the linguistic community have reinforced these judgments on similar and not so similar items (e. g., Bolinger 1978; Lasnik and Saito 1984; Pesetsky 1987; Ginzburg and Sag 2001; Pesetsky and Torrego 2001).

2.2 Comparing the process of community agreement and quantitative methods

The main advantage of the process of community agreement is that data of this type are readily collected, whilst weeding out questionable data to a good extent. However, it also has several disadvantages.

First, it is comparably opaque: On encountering a judgment in the literature, one does not exactly know how a rating came about. Thus, the process is difficult to replicate. Second, the process of community agreement can involve only very few participants, in the extreme case just the researcher her-/himself. Due to the low number of participants, scale biases, i. e. inter-speaker differences in the application of the scale, may be more pronounced than in quantitative methods. Further, there are a limited number of items, no fillers, no distribution across lists (i. e. a participant might see an item in multiple conditions), and no randomization. Since typically only a few instances of a particular phenomenon are involved, the process of community agreement may lack generalizability. Common statistical methods are not applicable. Most importantly, the purpose of the study is hardly concealed, if at all. That is, a participant can easily guess the research question and hypothesis – or might outright know about it. All these factors might have a negative impact on the reliability of the results. Furthermore, the process in its entirety is relatively time-consuming. For concise discussions of these issues, see Gibson and Fedorenko (2010) and Schütze (1996).

Finally, the process requires that authors and reviewers/readers share the language under investigation and hence best works for English and a few more well-represented languages. For other languages, many reviewers and readers are unable to confirm or correct author judgments, because they lack the intuitions to do so. As a consequence, judgments on less well-represented languages might reflect the intuitions of just a very few researchers and there might not even be agreement among them. This might have an impact on the quality of those judgments. Linzen and Oseki (2018) provide a more detailed discussion of how this process causes problems with less well-represented languages and demonstrate their point for Hebrew and Japanese. About half of the contrasts asserted by authors could not be replicated in a quantitative experiment (Linzen and Oseki 2018).

Quantitative methods, in contrast, require more effort – in terms of time and potentially costs as well as in terms of skills necessary to conduct an experiment and to analyze the outcome statistically. In the following, our focus is on quantitative *experimental* methods. Corpus analyses are, of course, also quantitative and many of the points below also apply to corpus analyses.

Quantitative experiments are typically characterized by systematically varying one or more factors (independent variables) and registering the effect of this variation, in our case the effect on the acceptability rating (dependent variable). At the same time, possible confounding factors are controlled for as far as possible. To this end, the roles of researcher and participant are strictly separated, and participants are not informed about the exact aims of the experiment and the hypotheses. Researchers make an effort to conceal the independent variables, typically by including filler items and distributing items across lists such that each list

contains each item in only one condition. For external validity, more than just a few participants are involved as well as several lexicalizations of the phenomenon under investigation. For a discussion of these standards applied to acceptability judgment experiments, see e. g. Cowart (1997) or Gibson and Fedorenko (2013). For a recent study on the reproducibility of acceptability judgments, see Langsford et al. (2018). There are various scales available; in linguistics, the following are common: 2-point scale, N-point scale (where N > 2; this is a gradient scale), scales produced in a magnitude estimation task (e. g. Stevens 1946; Sorace 1992; and Bard et al. 1996), scales produced in a thermometer method task (e. g. Kilpatrick and Cantril 1960; Nugent 2004; and Featherston 2008).

The process of community agreement and quantitative judgment tasks have one major thing in common: Everyone involved – the researchers, their colleagues, their students, as well as the reviewers and readers – has made their judgment introspectively. The same is true for participants in a quantitative judgment task.

Some researchers effectively combine experimental linguistics and theoretical linguistics. An earlier example for this is Pinker and Birdsong (1979) in the context of second language acquisition, and more recent examples are, to name a just a few, Keller (2000), Bresnan (2007) or Hofmeister and Sag (2010). As we will argue below, the process of community agreement is particularly useful to generate novel hypotheses in theoretical linguistics, while quantitative methods are useful to test predictions derived from these hypotheses.

2.3 Acceptability, grammaticality, and grammatical reasoning

The ultimate goal of linguistic theory is to understand the nature of human language. This includes describing and explaining the structure of human language(s). To this end, linguists wish to know which sequences of linguistic units are licit linguistic expressions. For syntax, this means knowing which sequences of words form grammatical sentences. Grammar, however, is a mental construct. Hence grammaticality cannot be observed directly. We only have access to intuitions about the perceived acceptability of sentences and these are inevitably affected by further factors in addition to grammaticality. Acceptability is the joint product of the competence grammar and the performance mechanisms which make use of the competence grammar during actual language use. Under this view, *grammaticality judgment* in the context of a judgment task is actually a misnomer and should be replaced by *acceptability judgment* (for a succinct discussion along these lines see Bard et al. 1996). Nevertheless, both terms are often used synonymously (e. g. in Schütze 1996).

Performance factors such as memory load and real-world (im)plausibility may make a sentence unacceptable despite being grammatical. Multiple center-embedding is a typical example. A sentence like (2), from Chomsky and Miller (1963: 286), does not violate any grammatical rule but it appears unacceptable because it is simply too complex to parse.[2]

(2) *The rat the cat the dog chased killed ate the malt*.

Examples like (2) are more easily accepted when presented visually compared to auditory presentation. Moreover, while examples like (2) are almost completely absent in corpora of spoken language, they do occur in written corpora (Karlsson 2007). In the auditory mode, acceptability is modulated by prosodic properties (Fodor et al. 2017). Both aspects point to a processing explanation. In general, grammaticality and processing effects can be disentangled by showing that acceptability of construction can be improved by manipulating aspects of prosody, context and/or lexical content, without changing the structure.

The opposite effect, *acceptable ungrammaticality* (Frazier 2008) or *grammaticality illusions* (Phillips et al. 2011; we prefer the term *linguistic illusions*, as they sometimes lie between syntax and semantics), also exists but is less common. Linguistic illusions have been observed for a range of phenomena including verb agreement, ellipsis, comparatives, negative polarity items and again multiple center-embeddings. Consider the following example from Christiansen and MacDonald (2009). (3a) is in fact ungrammatical, as it lacks a verb, but nevertheless received higher acceptability ratings than its complete counterpart (3b). This effect was first described in Frazier (1985); Gibson and Thomas (1999) coined the term *missing-VP effect*.

(3) a. **The chef who the waiter who the busboy offended frequently admired the musicians*.
b. *The chef who the waiter who the busboy offended appreciated admired the musicians*.

Regardless of which method is used to investigate a linguistic sequence, the direct output of that method is not a direct observation of the grammaticality of the linguistic sequence in question. In a judgment task, the researcher collects acceptability judgments typically by non-experts. In researcher introspection, the researcher collects, in a first step, their own acceptability judgment. In the case of the process of community agreement, the researcher considers several of such

2 Some linguists argue that the limits of center-embedding are encoded as grammar rules, e. g. Reich (1969) or Karlsson (2007).

judgments made by colleagues. In a corpus analysis, occurrences/frequencies are collected, and so forth. Then, through careful consideration and contrasting, this information, i. e. acceptability judgments, occurrences/frequencies, etc., is used to determine the grammatical status of a linguistic sequence. Such grammatical reasoning requires sufficient syntactic knowledge and logical reasoning, both of which require training. Consider example (2) again. When presented to non-experts, they find (2) unacceptable. When presented to syntacticians, their initial reaction is the same, they, too, find (2) unacceptable. However, in contrast to non-experts, syntacticians are better able to examine (2), see past its acceptability, and determine its true grammatical status.

Arguably, the process of community agreement combines acceptability intuitions and grammatical reasoning. When considering an item, linguists collectively query their introspective judgments and then apply grammatical reasoning to it. This often results in a reflection on both, an item's acceptability status and its grammaticality status.

2.4 The role of sentence pairs

In theoretical syntax, phenomena are typically researched by considering sentence pairs. A sentence pair consists of a sentence of degraded grammaticality/acceptability and a grammatical/acceptable counterpart. Pairs play an important role in linguistic argumentation and experimentation, not only in syntax but in other subfields of linguistics as well. They do so for a good reason: In linguistic argumentation, pairs allow for isolating grammatical factors while keeping all other factors constant (as far as possible), which allows linguists to extract the grammatical status of a sequence. In experimentation, pairs reduce the impact of confounding factors, thus eliciting "purified" acceptability judgments.

Two members of a pair should vary in only one feature. In reality, this is impossible to achieve since factors are typically coupled, even in a simple case like (4a) and (4b). The difference in agreement features is inevitably accompanied by a morpho-phonological difference (*is* vs. *are*).

(4) a. **It are raining*.
b. *It is raining*.

Furthermore, most linguists break with sentence pairs when analyzing complex phenomena. In the following, we will discuss various phenomena as examples for how syntax is not limited to the analysis of sentence pairs. Typically, such phenomena are analyzed using n-tuples, where $n > 2$ – with the exception of single grammatical/acceptable items.

Syntacticians often discuss single grammatical/acceptable items, where there is no relevant ungrammatical/unacceptable counterpart. Consider the example (5), taken from Stroik (2001: 367). In Stroik (2001), (5) is important for other contrasts and yet, it has no direct ungrammatical/unacceptable counterpart. As we will see in Section 4.1, single grammatical/acceptable items are surprisingly common.

(5) *Chris helped Pat, but Sam didn't help her.*

The discussion of (syntactic) blends, i. e. combining two or more violations, also involves three or more items. Consider the following example from Wellwood et al. (2018; without their emphases).[3]

(6) a. *Fewer people have been to Russia than I would have thought.*
b. **People have been to Russia fewer than I have.*
c. ?*Fewer people have been to Russia than I have.*

The same is true for super-additivity effects, where one has to consider at least a 3-tuple, see for example Hofmeister et al. (2014). Hofmeister et al. (2014) argue that super-additivity can be used to disentangle processing effects and grammatical constraints. Processing effects may combine in such a way that the resulting penalty is greater than the sum of the corresponding effects in isolation. Having two grammatical violations within a single sentence does inflate their effect beyond the sum. Crucially, testing for super-additivity requires an n-tuple: A version with no violation serving as the baseline, a version with violation A, a version with violation B and a version combining violations A and B. An example combining that-trace violation and agreement violation is given in (7d).

(7) a. *I was shocked to see which manufacturers the consumer report indicated make reliable and safe automobiles.* (baseline)
b. *I was shocked to see which manufacturers the consumer report indicated that make reliable and safe automobiles.* (that-trace violation)
c. *I was shocked to see which manufacturers the consumer report indicated makes reliable and safe automobiles.* (agreement violation)
d. *I was shocked to see which manufacturers the consumer report indicated that makes reliable and safe automobiles.* (that-trace violation and agreement violation)

3 The question mark in (6c) is meant to indicate a prediction for their experiment rather than the authors' judgment.

The discussion of island phenomena typically also involves three levels: (i) strong islands, (ii) weak islands, and (iii) non-islands.[4] Ratings for sentences without extraction out of an island will be higher than for sentences involving extraction out of a strong island. Ratings for sentences not violating any island will also be higher than for sentences with extraction out of a weak island. Finally, extraction out of a weak island will typically get higher ratings in comparison to extractions out of a strong island.[5] Consider the following examples from Szabolcsi (2006; Szabolcsi's diacritics).

(8) a. *Which topic do you think that I talked about?* (baseline)
b. ?**Which topic did John ask who was talking about?* (weak island)
c. **Which topic did you leave because Mary talked about?* (strong island)

Further, from such a series of ratings, viz. (iii) > (i), (iii) > (ii) and (ii) > (i), we can derive an ordered chain, viz. (iii) > (ii) > (i). Such chains are at odds with a notion of doing only pairwise comparisons, especially since one can build chains of chains. One way to deal with this conundrum would be to strip ">" and "<" of their transitivity – which is not an attractive option. The better alternative is that we do not restrict ourselves to only pairwise comparisons.

3 Sprouse et al. (2013)

The exact role of sentence pairs in syntactic enquiry is important for the debate on the reliability of syntactic data. Sprouse et al. (2013) had a considerable impact on the reliability debate, as they were the first to provide a quantitative comparison of author judgments and experimental results from several acceptability judgment tasks. To do so, Sprouse et al. (2013) created a corpus of items discussed in *Linguistic Inquiry* for the years 2001 to 2010, where an item is any sentence or sentence fragment that includes some kind of goodness judgment by the author(s). Sprouse et al. (2013) only extracted syntactic judgments from papers that were not "predominantly" syntactic. They then randomly sampled 300 items degraded in grammaticality/acceptability, extracted their good counterparts or constructed such counterparts, and constructed another 7 variations for each of the

4 The terms *weak* and *strong* do not primarily qualify effects size but rather the generality of the constraint. Weak islands are selective while strong islands are absolute. However, the penalty for extraction out a weak island is typically smaller compared to extraction out of a strong island.

5 Of course, such items can still be *tested* in pairs, for instance in a forced choice experiment. This is different, though, to strictly restricting one's grammatical reasoning to pairs.

300 sentence pairs. Sprouse et al. (2013) then tested those pairs in several online acceptability judgment tasks and checked whether acceptability judgments by the online participants concurred with the grammaticality judgments by the experts in a test of directionality. For each pair, a match occurred if participants judged the ungrammaticality/unacceptability item to be less acceptable than its OK-counterpart. Sprouse et al. (2013) collected judgment with several tasks and applied several tests for determining convergence. For their main condition and their main test, they reported a near-perfect match rate of 95 % (and argue the remaining 5 % could be false positives).[6]

In our view, a test of directionality is a rather weak test. For any sentence pair, it takes a lot of disagreement between authors and non-experts before the directionality test indicates a mismatch (cf. Langsford et al. 2018, who call this kind of task a *targeted contrasts task*). Of course, Sprouse et al.'s choice of test is determined by the assumption that syntactic research is restricted to the analysis of sentence pairs. As argued in the previous subsection, we think that this assumption is too strong. A good deal of syntactic judgments are made with consideration of a wider context. Thus, the question arises, how would the judgments of linguists compare to judgments by non-experts if one shifts the focus from analyzing sentence pairs to analyzing items at large, i. e. when analyzing not only sentence pairs, but comparing a variety of items at the same time?

One could try to reanalyze Sprouse et al.'s results, in order to answer this question. However, their methodological choices are motivated by the importance they attribute to sentence pairs. For example, Sprouse et al. (2013) restricted themselves to sampling items of degraded grammaticality/acceptability and then extracted the respective counterparts, or, if none was present, created counterparts themselves. Further, they did not consider OK-items without an ungrammatical/unacceptable counterpart – but as we will see below, such items are pretty common. Such sampling choices restrict Sprouse et al.'s sample, which might affect the results. Further, Sprouse et al. (2013) consider in-between categories with respect to author judgments, for instance a ??-item vs. an OK-item. As we will argue in Section 6, such in-between items are likely to weaken the results of an at large comparison. As a consequence, if one were to reanalyze Sprouse

6 Effectively, Sprouse et al. (2013) report a null result: There is no difference between expert and non-expert judgments. However, no observed difference does not mean that there is no underlying difference, i. e. absence of observed divergence does not necessarily mean underlying convergence (cf. Altman and Bland 1995, who make this point in a more general context). We emphasize this point, because we are worried that some could draw conclusions from Sprouse et al. (2013) that are too strong. In their paper, Sprouse et al. (2013) themselves are cautious about possible inferences.

et al.'s results, using other tests, the results of such tests are likely to be difficult to interpret.[7]

Therefore, the aim of this paper is to also compare grammaticality judgments by linguists to acceptability judgments by non-expert participants, but by considering items at large. Similar to Sprouse et al. (2013), we randomly sample items from the literature, which we then use to compare author judgments and experimental ratings. However, we depart from their approach in several methodological aspects, so that our methodology works best for testing and analyzing a variety of items at the same time. We sampled items across the board, including OK-items without counterparts. However, we do not consider in-between items and items that come from authors who only use ["*", OK], in order to not weaken our analyses by having adjacent categories. We also do not create counterparts, as our analyses can handle a lack of variants. Other differences include that we do not create variants for our items, that we also extract syntactic judgments from papers that were not "predominantly" syntactic, and that we accepted other minor differences in creating our corpus. These differences imply that the present paper cannot and does not want to be a replication study of Sprouse et al. (2013). Rather, Sprouse et al. (2013) viewed the debate from one angle and we offer another one.

4 Our *LI* corpus

We wish to compare expert judgments from the process of community agreement to experimental ratings from a quantitative experiment. Ideally, we would have a complete list of syntactic phenomena discussed in the literature, check how these were evaluated by syntacticians, and then obtain experimental ratings on those phenomena. However, such a list does not exist and it would require resources beyond our means to create it and to then gather experimental ratings for all the items on that list. To get as close to this scenario as possible, we chose to construct a corpus of phenomena discussed in the literature, from which we then randomly sample sentences. We base our corpus on items discussed in *Linguistic Inquiry* (*LI*) for the years 2001 to 2010, where an item is any sentence or sentence fragment that includes some kind of goodness judgment by the author(s). This procedure is similar to the one by Sprouse et al. (2013). However, there are various minor differences between our corpus and the corpus by Sprouse et al. (2013). Arguably, the most no-

7 Sprouse et al. (2013) also include a test that tests items at large, i. e. in a broader context, as a secondary analysis. However, because of the fact that Sprouse et al.'s setup was designed for testing sentence pairs, it is not clear how to interpret the results of their secondary test.

Table 1: The categories that we use in our *LI* corpus and the number of items that fall into each category, including the percentage of total number of items. Each category includes the non-testable items.

Judgment category	No. of items
Standard acceptability judgments	2619 (60 %)
Coreference judgments	1184 (27 %)
Interpretation judgments	412 (9 %)
Few lexical items	103 (2 %)
Prosodic judgments	43 (1 %)

table difference is that we only consider items that were judged by authors who are (likely to be) native speakers of American English. The reasoning behind this choice is that we wished to focus on the most reliable judgments, as judgments by non-native speakers might be less reliable or not their judgments but taken from the literature or based on some kind of survey. Our categorization of nativeness will include wrong categorizations. For us, it was most important to have transparent and objective criteria. For further details and further differences and their motivations, see Appendix A that is available in the online version of this article.

4.1 The structure of our *LI* corpus

In total, there were 335 papers in *LI* in the years 2001 to 2010. 160 papers were authored by at least one native speaker of American English and thus considered in our extraction process. We went through those 160 papers and extracted items from them that contain some kind of judgment, whether of syntactic, semantic, or other nature. However, only 103 of those 160 papers feature in our corpus, because some of the papers do not include items that contain a judgment. And only 90 papers contained at least one syntactic acceptability judgment.

We extracted 4334 items that include some kind of judgment and categorized them according to the categories in Table 1. The categories are adopted from Sprouse et al. (2013: 221). 2619 items are standard acceptability judgments. According to Sprouse et al. (2013), sequences with a standard acceptability judgment are sequences that can be tested in a simple judgment task. Sequences with few lexical instantiations are typically very short sequences that are hard to vary lexically. Such sequences have their own category, "few lexical items" in Table 1, but arguably, this category was more relevant for Sprouse et al. (2013), as they created variants of the sampled items. We added a category for non-testable items, which are items that would have been hard to test in our experimental design for

Table 2: The number of judgment categories used in the corresponding paper from which an item was extracted, given for the 2619 standard acceptability judgments in our *LI* corpus. We include both absolute numbers and their percentages, the latter in parentheses.

Number of judgment categories	No. of items
2	486 (19 %)
3	654 (25 %)
4	460 (18 %)
5	936 (36 %)
6	83 (3 %)

one or several of the following reasons: They include deictic references, they include strong language, there are unintended alternative readings available such as repair readings or garden paths, or they include colloquial language that might be stigmatized. Of the 2619 items, we deemed 2539 testable (97 %). Table 1 gives the categories and the number of items that fell into the different categories.

Out of the 2619 standard acceptability judgments, 486 items come from authors whose judgments were binary, i. e. throughout their paper they used only two judgment categories, ungrammatical/unacceptable items, typically marked with a "*", and grammatical/acceptable items, which we refer to as OK. This is the first row in Table 2. The other 2133 items, the sum of the second to fifth row, come from authors whose judgments were gradient and include some form of "?", e. g. "?*", "??", etc. See Table 2 for details.[8]

8 Note that the concrete judgment categories are not always the same across papers. One paper with three judgment categories might use ["*", "?", OK], while another paper might use ["*", "?*", OK]. To give an example, we extracted 36 items with a standard acceptability judgment from Stroik (2001). We checked how many judgment categories Stroik (2001) uses. It's two: ["*", OK]. Thus, extracting from Stroik (2001) increased the counter for "2 judgment categories" by 36. Further: In principle, any "binary paper" could be an underlyingly "gradient paper" in which no in-between category was used. In such a case, we would not be able to tell the difference between a true binary paper, where the authors indeed intended to use a binary scale, and a seemingly binary paper, where the authors in principle endorse a gradient scale, but they only discussed *- and OK-items. It is likely that this reasoning applies to some of the binary papers, because the binary papers include on average considerably fewer items than gradient papers: 12 vs. 41 items, respectively.
Related to this is the observation that some authors appear in both categories. Such authors might have changed their minds over time; or the questionable papers only appear binary because too few phenomena were discussed. We decided to accept that ambiguity, as we did not want to decide whether an author "underlyingly" assumes a gradient scale or whether an author has changed his/her view over the years.

Table 3: The distribution of the rating categories for the 2619 standard acceptability judgments in our *LI* corpus and the number of items that fall in each category, including the percentage of total number of items. The "Other" category includes "#", "#?", "OK/*", "%", "*?", "(?)", "(??)", "??/*", "???/*" (where "#", "#?", and "%" seemed to be used for syntactic evaluations in the instances in question).

Rating category	No. of items
OK	1501 (57 %)
*	965 (37 %)
?	61 (2 %)
??	49 (2 %)
?*	23 (1 %)
Other	20 (1 %)

The majority of items come from authors who provide gradient judgments, viz. 2133 of 2619 items (81 %). 486 of 2619 items (19 %) come from authors who provide only binary judgments. This proportion is reflected in the scales used in the papers that we extracted from. 90 papers in our corpus include at least one standard acceptability judgment, 30 papers use a binary scale and 60 use some form of a gradient scale.

However, while most authors use three or more categories, the in-between categories only play a minor role in the discussion of syntactic phenomena and in theory building. Only 153 of the 2619 standard acceptability judgment items were judged as being in-between, indicated by "?*", "??", etc. The vast majority of items were either judged as OK, viz. 1501 items, or as *-marked, viz. 965 items. (5) from above is an example of an item that was judged as OK in the paper from which we have extracted it. So (5) would increase the counter for the OK rating category by one. See Table 3 for details.

The results in Table 3 also imply that there were many OK-items without a counterpart that is fully or somewhat ungrammatical (marked with "*", "?", etc.). The lack of counterparts further undermines the idea that syntactic enquiry should be primarily carried out based on the analysis of sentence pairs.

However, it is not immediately clear what in-between categories like "?" denote exactly. There are two main interpretations. First, in-between categories could be used to denote a gradience in grammar, i. e. the in-between status of the linguistic sequence in question comes directly from the grammar. Second, they could be used to denote extra-grammatical influences, i. e. the linguistic sequence in question is underlyingly (un)grammatical but extra-grammatical factors degrade or ameliorate its acceptability. Arguably, for most authors in *LI*, in-between categories denote the latter. As this point will be particularly relevant for the interpretation of our results, we decided to look into this question systematically.

To determine an author's use of in-between categories, we checked two things. First, we checked their publication list and whether there is any indication that the author assumes gradience in grammaticality. Second, we checked whether the *LI* paper from which our items were sampled discusses gradience. In this assessment, we focused on the authors from whom our experimental items where sampled (see Section 5.1). We found that none of the original authors explicitly support the idea of a gradient grammar, thus, we assume the authors use the in-between categories to mark extra-grammatical factors influencing the acceptability of an item. A list of this assessment is attached as Appendix C, available in the online version of this article.

A similar reasoning applies to the distinction between items from authors whose judgments were binary and items from authors whose judgments were gradient. One can make the case that such items require slightly different measurement techniques and arguably, they require different analyses. In the present paper, however, we focus on the more common case, i. e. the items that come from authors whose judgments were gradient.

5 The experiment

Since we are dealing with a meta-issue, i. e. the degree to which author judgments and non-expert judgments converge, we are not interested in concrete syntactic constructions. This allows us to randomly sample items from our *LI* corpus. The sentences come with an expert judgment and we obtain the quantitative results through an online experiment.

We focus on items that come from papers in which authors use three or more judgment categories, i. e. "gradient papers". There are two reasons for this. First, such items make up the majority of all extracted standard acceptability judgments, viz. 81 % (see above for details). Second, for items from "gradient papers", we can better motivate a strong null hypothesis and the results become easier to interpret, as well. See Section 5.2 for details.

In the experiment, we use a 7-point scale. We choose an N-point scale for reasons of simplicity. In contrast to other methods, like scales produced in a magnitude estimation task (e. g. Stevens 1946; Sorace 1992; and Bard et al. 1996) or a thermometer method task (e. g. Kilpatrick and Cantril 1960; Nugent 2004; and Featherston 2008), an N-point scale requires little introduction and the method is as well validated as other common tasks (cf. Weskott and Fanselow 2011; Langsford et al. 2018). We chose a 7-point scale, because participants typically prefer five or more degrees (cf. Bard et al. 1996: 45) and accommodating their choice reduces

quantization effects, i. e. errors caused by "rescaling" one's preferred scale to the scale that one is asked to use.

We then check to which degree results from the process of community agreement and quantitative results agree. We test items at large, i. e. our analyses are not restricted to sentence pairs, for the reasons discussed above. To estimate the degree of convergence between expert judgments and non-expert judgments, we report convergence measures such as Cohen's kappa as well as the results of a general mixed-effect model. A reasonable null hypothesis could be as follows.

(H0) The process of community agreement and quantitative methods concur; i. e. there is no considerable difference between the expert judgments and quantitative experimental judgments.

5.1 Experimental setup

Materials

We randomly sampled 100 sentences from our *LI* corpus, without replacement, using R's *sample* function (R Core Team 2015). All items are standard acceptability judgments as described in Section 4 and all 100 items come from papers in which the author(s) used a gradient scale. We sampled those 100 items such that we have 50 *-items and 50 OK-items.[9] Those are not paired – we sampled each item independently from potential counterparts (as mentioned above many *LI*-items do not even have a counterpart). A list of the sampled items is attached as Appendix B, available in the online version of this article.

We could have also sampled items such that they had reflected the real proportions of *-items and OK-items in our corpus. However, we preferred a 50-50 split between *-items and OK-items to not cause any imbalances that we then would have had to offset by additional fillers; including too many bad or too many good items can have an impact on the ratings (cf. Cowart 1997; Sprouse 2009).

We focus on endpoints, that is *-items vs. OK-items, and ignore in-between categories, e. g. "?", "??", etc. This way, any observed difference will make the strongest possible point about differences between expert judgments and quantitative results.

9 One problem that we were facing upon further inspection of the items is that some items were miscategorized or not testable after all in the sense described in Section 4.1. As a consequence, we sampled slightly more items than needed, two times 55 items, and took the first 50 items that were true standard acceptability items and that were also testable. In total, we recategorized 5 items.

Finally, one might be concerned that the authors' different application of their scales makes a comparison of items at large and across authors invalid. However, the impact of such scale biases should be mitigated by the fact that the expert judgments went through the process of community agreement and thus represent not only the judgments of the authors but build on previous literature and include the judgments by their students, colleagues, and reviewers. Secondly, the effect of scale biases should be further mitigated by the fact that all of the authors from which we sampled were using three or more judgment categories and that we are looking at endpoint categories. Further, the judgments come from experts who have a motivation to use a scale that is accessible to their peers. If the meaning of "*" and OK varied to such a degree that this causes systematic violations in our test, despite the mentioned mitigating effects, then this would not reflect well on the discipline.

Fillers

Our experiment does not contain fillers. Fillers typically have two purposes: First, to obscure the purpose of the study. Second, to offset imbalances within the set of critical items (cf. Cowart 1997: 51–52). With respect to the first point, fillers would not further obscure the purpose of our study, since we are dealing with a meta-issue, viz. data convergence. As to the second point, our items are balanced already, because of the mentioned 50-50 split.

Procedure

Any sampled item comes with an expert judgment, as per sampling procedure. We obtained the quantitative judgments in an online acceptability judgment task, using a 7-point scale. See Figure 1 for an illustration.

Instructions

The instructions read as follows: "In the following, you will read certain sentences. Please evaluate how natural those sentences are with respect to their grammaticality." Then, we introduced the rating scale. Participants made their ratings by pressing buttons. The experimental interface is shown in Figure 1. When introducing the scale, we told participants that the lowest value corresponds to "fully unnatural/ungrammatical", represented by the red button, and the highest value to "fully natural/grammatical", represented by the blue button. The remaining buttons denote values "in between". Participants were asked to rate a sample sentence, and were then instructed further: "The question you should keep in mind when rating the following sentences is: How natural do they sound to you with respect to the grammaticality?" And continued: "Please do

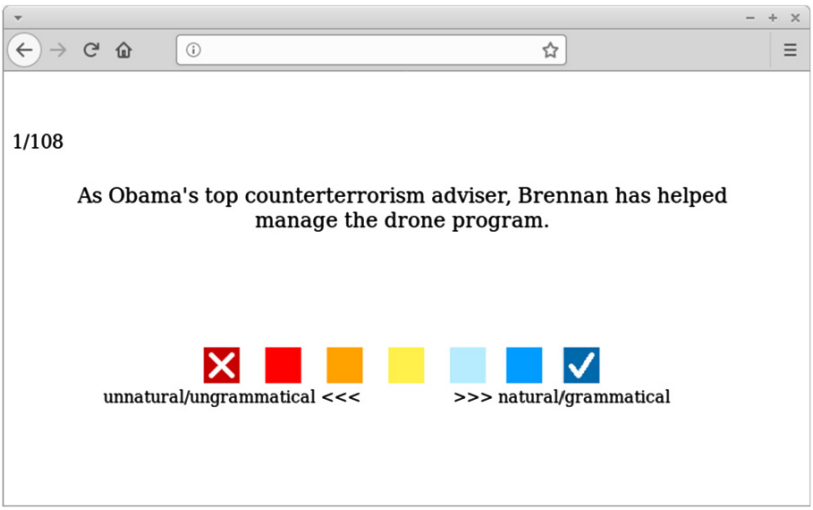

Figure 1: An illustration of the experimental interface that we used in our experiment.

not be bothered with punctuation, spelling variants, or meaning. As to meaning: For example, while *Jack did his job goodly* is meaningful and plausible, it is also not fully grammatical. Thus, it should receive a low rating. On the other hand, *The storm intentionally broke the window* is not completely plausible but fully grammatical. Thus, it should receive a high rating." The final piece of guidance was: "Further, this is about your intuition and there is no right or wrong. This also means that there is no need to brood over sentences."

The first four items were calibration items, presented in a random order. These items were included to give participants an idea of the range of possible goodness and badness they might encounter. The calibration items comprised two bad items, (9a) and (9b), and two good ones (10a) and (10b). The two bad items are modeled after sentences in Ferreira and Swets (2005) and received low ratings therein. The good items are from *USA Today*.

(9) Bad items:

a. *This is a donkey that I don't know where it lives.*
b. *This is the man that I don't know where he comes from.*

(10) Good items:

a. *As Obama's top counterterrorism adviser, Brennan has helped manage the drone program.*
b. *Iran has proposed restarting talks as early as next month.*

After the calibration items, the critical items followed in a random order, shuffled anew for each participant. We randomly interspersed four *gotcha* items, which we used to detect non-cooperative behavior; see below for details. Besides ratings, we also collected response times, in order to detect non-cooperative participants. The response time of an item is the time from loading the item until the time at which the rating is given.[10]

We also included an on-line warning mechanism that warned participants when they became potentially non-cooperative. Non-cooperative behavior can be a problem in crowdsourced tasks (cf. Downs et al. 2010; Kazai et al. 2011); and in our experience, such a warning mechanism can bring down rates of non-cooperative behavior by more than 50 %.

For any given item, we deem a response time below 400 ms as non-cooperative. If a participant slipped below this threshold four times, then a pop-up window appeared, asking the participant to pay attention. A second pop-up appeared after the twelfth violation.

Participants

We recruited 120 participants through Amazon Mechanical Turk, which were then redirected to an external website for the actual task. As entry requirements, we asked participants to be "master workers". Such participants have a high number of positive evaluations from previous tasks and have finished a high number of other tasks. This is to ensure that only reliable participants would take part. We are only interested in native speakers of American English, which we stated in the description. To accommodate participants from the US, recruiting took place between 19:00 and 22:00 GMT (between 15:00 and 18:00 EST). Payment was such that it resulted in an hourly rate of about $12. In the introduction, we mentioned that only non-linguists could participate, i. e. we trusted linguists to not lie to us and to not take part nonetheless. We also mentioned that our study was approved by and followed the guidelines of the ethics committee of our university.

After the experiment, we anonymously gathered the following information: age, gender, and whether the participant considers themselves as native speakers of American English. We asked participants to answer truthfully and stressed that whatever they answer, they will still receive payment. We did this, because technically, non-native speakers of American English could still participate.

Our experiment had the following demographics (this is after exclusions; see below for details): 111 participants included; mean age: 40.64 years (SD: 11.93); gender distribution: 54 female and 57 male participants.

10 Response times were calculated on the client's side, so latency should not be an issue.

Exclusion criteria

In total, we excluded 9 out of 120 participants. We excluded participants for the following reasons: not being a native speaker of American English (0 exclusions), returning incomplete results (5 exclusion), having extreme response times (3 exclusions), or failing on gotcha items (1 exclusions). Our exclusion criteria were applied in the order listed below. We will briefly discuss the criteria in the following.

Not being a native speaker of American English: Only participants who stated that they consider themselves as native speakers of American English were included.

Returning incomplete results: If a participant returned incomplete data, then we excluded that participant.

Having extreme response times: If a participant had extremely low response times, i. e. the participant was extremely fast, or high response times, i. e. the participant was extremely slow, then we excluded that participant. Extremely low response times are a sign that a participant is just clicking his/her way through the questionnaire, without paying attention. Extremely high response times could be a sign that the participant is distracted, possibly affecting the quality of the responses, as well. We identified "extremely low/high" response times by means of two thresholds. The lower threshold is defined as the mean of all participants' median response times minus one and a half standard deviations of the median response times of all participants. The upper threshold is defined as the mean of all participants' median response times plus four standard deviations of the median response times of all participants.

Failing on gotcha items: Some participants have plausible response times but are distracted. This might affect the quality of their ratings. To exclude them, we include gotcha items, i. e. items for which we have a clear expectation with respect to the ratings to be given. We included two items that should sound bad to native speakers of American English, (11a) and (11b), modeled after Hansen et al. (1996), and two items that should sound good to native speakers of American English, (12a) and (12b), modeled after Brians (2014) and Kövecses (2000), respectively.

(11) Bad in AE:

a. *Peter wanted that we should come early*.

b. *My knowledges of chemistry are rather weak*.

(12) Good in AE:

a. *My son's grades have gotten better since he moved out of the fraternity*.

b. *The professor requested that Dillon submit his research paper before the end of the month*.

These items were randomly interspersed in the final two thirds of the questionnaire. If (11a) and (11b) received an averaged rating that is higher than those of (12a) and (12b), then we exclude that participant. Altogether, each participant rated 108 items: 50 *-items, 50 OK-items, 4 calibration items, and 4 gotcha items.

In total, we excluded 9 out of 120 participants. This exclusion rate of 7.5 % is somewhat lower than the exclusion rate reported in Munro et al. (2010). We think that this is because of the fact that we have only recruited participants with high approval rates and high task familiarity (so-called "master workers"). After exclusions, we have a total of 11,100 data points.

5.2 Analyses

We do not explicate formal criteria for the rejection of our null hypothesis, that is, we do not formally test our (H0). The reason is that different tests and different data structures will lead to different results and any setup will allow for different interpretations, making it hard to claim that a certain test setup is able to "prove" or "disprove" the reliability of syntactic data. So instead, we simply present the results of the experiment and highlight *-items that have received unexpectedly high ratings in the experiment and OK-items that have received unexpectedly low ratings. We also report the results of a general mixed-effects model, in which we model the online ratings as a function of the author judgments, using a Poisson distribution. Our random effects are item IDs and participants. The model was run in R (R Core Team 2015), using lme4 (Bates et al. 2015) and the pseudo-code of the model is given in (F1).

(F1) glmer(experimental_rating ~ author_judgment + (1|item) + (1|subject), family = poisson)

A considerable problem with the model is that it predicts values on a 7-point-scale ("experimental_rating") based on a binary variable ("author_judgment"), which means that the model cannot find a perfect fit. Therefore, we report further measures of data convergence, viz. Cohen's kappa, as well as precision and recall rates, and F1 scores.

To this end, we transform the experimental ratings to a binary variable, for which we group together ratings of [1, 2, 3] as "unacceptable" and ratings of [5, 6, 7] as "acceptable", dropping ratings of [4]. Cutting the scale into halves is of course an arbitrary decision. The boundary between grammatical/acceptable and ungrammatical/unacceptable could well be located below or above [4]. However, for the item set at hand, it seems appropriate given that the authors of the corresponding *LI* articles used a scale with more than two points and given that our

items represent the endpoints of that scale. Choosing [4] as the cutting point is reasonable as it makes the least additional assumptions.

As to Cohen's kappa, it is a widely used measure for inter-rater agreement. It can be interpreted as the proportion of agreement corrected for random agreement, with a value of 1 indicating perfect agreement, and a value of 0 denoting that the rate of agreement is equivalent to random (Cohen 1960).

Treating the transformed experimental ratings as actual values vs. the author judgments as predicted values, we also give the precision and recall rates, as well as the F1 scores.

William Snyder (personal communication) made the following important point: "(...) different U.S. English speakers can (...) have genuinely different judgments, as a result of regional and idiolectal differences in their grammars. Averaging across many participants should reduce the effects of noise in individuals' judgments, but it could also have the effect of introducing some noise, as a result of mixing the judgments of systematically different grammars." To account for this, we also illustrate individual ratings.[11] The input data, the R code to our analyses, and further files can be found on our Gitlab (https://gitlab.com/superpumpie/data_convergence).

5.3 Ratings and results

The expert ratings vs. experimental ratings are illustrated in Figure 2 (left) and items with unexpectedly high or low ratings are illustrated in Figure 2 (right). The concrete items, the corresponding average ratings, and standard deviations can be found in Appendix B. We analyze both non-normalized results and normalized results, normalized using Z-scores. Results normalized by using Z-scores can reduce scale biases, but they can also be harder to interpret. With respect to the number of items with unexpectedly high or low ratings, the number is somewhat lower for normalized results, viz. 15 (it is 20 when non-normalized). Otherwise, non-normalized and normalized results come out very similar. In the following, we use the non-normalized results.

Cohen's kappa comes out at 0.457, a value which is commonly considered as only moderate (Landis and Koch 1977). Precision, recall, and F1 scores are given in Table 5. They are also far away from perfect agreement.

11 As a reviewer pointed out to us, it would also be interesting to distinguish between different subfields. However, reliably annotating the items in our corpus for subareas is non-trivial and resource intense, so that we have to leave this issue for future research.

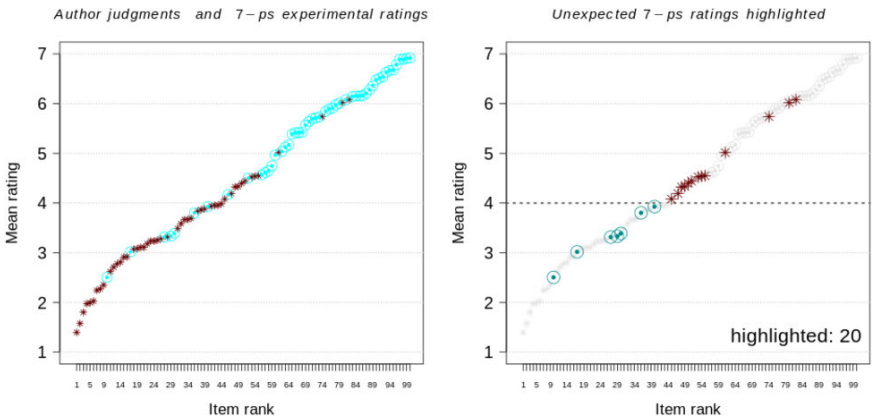

Figure 2: (left): The figure shows the averaged experimental ratings for items that come from *LI* articles in which authors use a gradient scale. Items, lined up on the x-axis, are ordered by their averaged ratings, given on the y-axis. Expert judgments for *-items are marked by red asterisks and judgments for OK-items are marked by blue circles. (right): The figure highlights *-items with an average rating above 4 and OK-items with an average rating below 4.

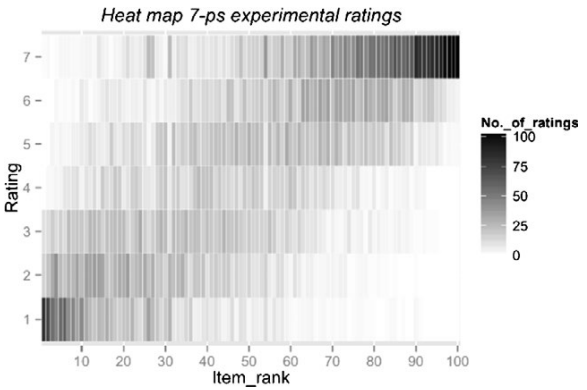

Figure 3: A heat-map for the individual ratings. Items are ordered in ascending order on the x-axis with respect to their averaged experimental ratings. For any x-y combination, the shading shows how often a certain rating was given, where darker hues represent increasing frequency. The results are given for non-normalized data.

There are two critical observations. First, *-items and OK-items, as per author judgments, mingle to a surprising degree. Second, instead of an S-curve, we observe a near-linear increase of the averaged experimental ratings, with many items having an average rating of 3 to 5. There are two possible explanations for the ob-

Table 4: A summary of the general mixed-effects model, including coefficient estimates, standard errors, and the z-values. The asterisk marks significance with a level of < 0.001.

Fixed effects	Estimate	Standard error	z
(Intercept)	1.19	0.04	29.02 *
Author judgment	0.48	0.06	8.50 *
Random effects	Variance	Standard deviation	
Item	0.08	0.28	
Participant	0.01	0.11	
Correlation of fixed effects: –0.685			

Table 5: Precision, recall, and F1 scores for actual ratings, i. e. the transformed experimental ratings, vs. the predicted ratings, i. e. the author judgments.

Rating bin	Precision	Recall	F1 score
Unacceptable [1, 2, 3]	0.771	0.636	0.697
Acceptable [5, 6, 7]	0.702	0.819	0.756
Weighted average	0.730	0.745	0.732

served gradience. First, the gradience could be due to aggregation across participants. Or second, it could be because participants indeed rated items in a gradient way.[12] As the heat-map in Figure 3 illustrates, the observed gradience is (at least to a good part) due to the fact that participants did rate the experimental items gradiently, i. e. using intermediate ratings. The amount of observed gradience is surprising. However, an in-depth discussion of where the gradience is stemming from is beyond the scope of the present paper. For a discussion of gradience more generally, see e. g. Keller (2000), Sorace and Keller (2005), Featherston (2005), Lau et al. (2016), and the contributions in Fanselow et al. (2006).

For the general mixed-effects model, we also use the non-normalized results, as the model is able to handle individual variation. The results of the general mixed-effects model can be found in Table 4. The author judgments are a fair predictor of the experimental ratings. However, the convergence rate is far from perfect. We also observe that there is very little variance across items and participants. Further, we do not expect much variance from order effects, as items were randomized for each participant individually. Hence, another factor would be needed to explain the somewhat low convergence rate.

12 The "or" is an inclusive "or".

6 Discussion

We have compared quantitative experimental ratings and expert judgments. In our experiment, participants rated items taken from *LI* papers which were either *-marked in the respective papers or unmarked. Participants rated each item on a 7-point scale. By and large, the experimental ratings agree with the expert judgments: Items marked by an asterisk in the *LI* paper tend to get low experimental ratings and items left unmarked in the paper tend to get high ratings. Yet, the results, in particular a kappa value of only 0.457, indicate that the degree of convergence is far from perfect.

Our results point into a different direction than the results of the test of directionality in Sprouse et al. (2013). Sprouse et al. (2013) reported a match rate of 95 %, where the remaining 5 % could be false positives. However, it is important to note that our findings do not contradict Sprouse et al.'s findings. Their methodological choices were made in order to test sentence pairs. Our study tests items at large and therefore applied a somewhat different methodology. Hence, we interpret Sprouse et al.'s results and our results as different takes on the underlying linguistic reality.

This leaves the question of what could have caused the partial lack of convergence in our data. A crucial point to keep in mind is the difference in construct that is measured. Author judgments indicate formal/structural well-formedness (grammaticality) whereas non-expert judgments reflect perceived well-formedness (acceptability). The two constructs are of course related. Grammaticality judgments are based on perceived well-formedness and grammatical reasoning. Acceptability ratings, on the other hand, are the joint product of grammar and processing. Below, we therefore discuss performance factors and other extra-grammatical factors, as well as scale biases and questions regarding the test. Furthermore, we discuss the issue of sentence pairs and whether the mismatch could be a true divergence in judgments between author judgments and non-expert judgments.

Performance factors

One might argue that the mismatch between the expert judgments and the quantitative online ratings is caused by performance factors, like memory load, real-world (im)plausibility, ambiguities, etc., influencing the experimental ratings. Under this view, the online participants' acceptability judgments were confounded by processing factors to a considerable degree. The experts' judgments, on the other hand, are not affected, because for our experimental items, all the expert judgments are grammaticality judgments. And thus, by comparing grammaticality vs. acceptability, the results would be distorted.

We agree that performance factors can affect acceptability ratings. Arguably, this is why the example in (2), *The rat the cat the dog chased killed ate the malt*, is commonly viewed as unacceptable. Our materials, however, were screened[13] for a number of potentially confounding extra-grammatical factors: They do not involve any garden-path sentences, multiple center-embedding, ambiguities, agreement attraction configurations, idiomatic expressions, strong language etc. Removing such items mitigates the impact of performance factors, but of course it does not rule out the presence of other performance factors (we will discuss concrete examples below).

The fact that there are quite a few items that were judged ungrammatical in *LI* and that at the same time received relatively high ratings in the online experiment also argues against attributing the mismatch to performance factors. Though performance factors can ameliorate perceived acceptability and even create illusory grammaticality, cf. Frazier (2008) and Phillips et al. (2011), it is unlikely that they are the driving force in the current data set, as it does not exhibit any resemblance to known linguistic illusions.

Other extra-grammatical factors

There are other extra-grammatical factors that could have an influence on the quantitative experimental results, in particular scale effects and semantic effects. It is beyond the scope of the present paper to discuss these in detail and we leave this for further research. Crucially, though, in the present case, scale effects (i. e. differences in the response options)[14] could at best explain why items somewhat gravitate towards the center of the scale, but they cannot explain why the experimental judgments cover the whole rating space in a rather linear fashion.

Scale biases

Scale biases are inter-speaker differences in the application of the scale. Scale biases affect both expert judgments and experimental ratings. However, in both cases, their impact should be minor. The impact on the expert judgments is reduced, because we are looking at prime examples, that is example carefully chosen by the authors to illustrate some phenomenon. In some cases, the example might be taken from or modeled after examples in the existing literature. If so, the

13 We did so introspectively, i. e. the stimuli have not been tested independently for this.

14 Most authors use 3 to 5 response categories (cf. Table 2) while participants were requested to apply a 7-point scale. But it is important to remember that we extracted only end-of-scale categories on part of the author judgments – fully grammatical items (OK) and *-marked items.

given judgment does not only reflect the current authors' perception of the item but also the intuition of their predecessors. Furthermore, the authors' colleagues, students, and reviewers have probably contributed to the judgments, as well. The impact on the quantitative judgments is reduced by averaging over a high number of participants and can be further reduced by normalizing the individual ratings using z-scores. Furthermore, we were sampling endpoint items, i. e. *-items and OK-items, and all authors from which we sampled use in-between categories. This should have reduced scale biases further.

However, if one does take the view that the relative lack of convergence was caused by scale biases on the authors' side, then this would be very troubling. It would mean that, despite the rather lengthy and costly process of community agreement, the judgment categories varied so much across authors that the meaning of those categories is not easily accessible. If that was the case, then authors could not readily use each other's judgments for joint theory building, and they could not compare phenomena across languages. If that really was the case, then the field should move on to a standard that is easier to interpret (Schütze 1996 makes a similar point for in-between judgment categories).

Reliability of online ratings

One might question the reliability of online ratings in general. Various studies have found that crowdsourcing studies replicate results from experiments in which the participants are present in the lab (e. g. Munro et al. 2010; Schnoebelen and Kuperman 2010; Sprouse 2011; for other fields see e. g. Mason and Suri 2012 or Krantz and Dalal 2000). Still, non-cooperative behavior can be an issue for online studies (see e. g. Downs et al. 2010; Zhu and Carterette 2010; Kazai et al. 2011). The present study went to great lengths to mitigate the issue. We only included "master workers", i. e. participants with relative high task familiarity and high approval ratings, and we implemented the on-line warning mechanism and the exclusion criteria described above. Further, we have a relatively high N and used other measures to ensure high data quality, like a comprehensive introduction of the rating scale and calibration items. Because of this, we are positive that our results are reliable and reproducible.

Questions regarding variance

One might wonder whether the number of data points is sufficient to draw any conclusions from our data. Experimental noise could lead to irrelevant overlaps when testing neighboring categories, in our case judgments of ungrammaticality and grammaticality by the authors. However, our setup is different. First, as just argued, we expect relatively little measurement error, certainly in the expert

judgments. Crucially, though, we are not looking at neighboring categories. All items come from papers in which the authors use three or more judgment categories like "?" and "??". As argued in Section 4.1, those in-between categories are used to indicate extra-grammatical influences. So for any potentially in-between item, the authors could have used an in-between judgment category. Thus, it was reasonable to expect clustering towards the endpoints and an extremely high convergence rate.

Sentence pairs

Our argumentation depends on the question whether syntactic enquiry is, on the whole, limited to the analysis of sentence pairs. If one assumes that it is, then the results by Sprouse et al. (2013), echoed in our (H0), are an adequate reflection of the convergence rate of syntactic data. However, we think that this assumption would be too far-reaching. We agree that the analysis of sentence pairs plays an important role in syntactic theory, similar to the analysis of minimal pairs in phonology, but we do not think that syntactic theory is restricted to the analysis of sentence pairs. We provided evidence against such an absolute status above. First, certain phenomena are difficult to discuss in only pairs, e. g. blends, superadditivity effects, and island effects. Second, even if one assumes strict pairs, then there is still the loophole of forming chains of pairs, as discussed in Section 2.4. Further, a good deal of items presented in the literature are discussed in the absence of a counterpart – this applies to the analysis of well-formed sentences in particular. Thus, we do not think that the field is restricted to the analysis of pairs. Instead, judgments are made with consideration of a wider context.

However, we understand that this point is not without controversy. Therefore, we ask those readers who do not agree with our argumentation to treat the following as a *what-if*. What if the field wanted to move beyond the analysis of sentence pairs? Would there be issues with data convergence?

The actual judgments

As to the quantitative experimental ratings presented above, we think that they are a robust reflection of the underlying linguistic reality. We went to great lengths to ensure a solid methodology and execution, making sure that the results are as free of confounding factors as possible. Thus, we treat these acceptability ratings as a reasonable approximant of grammaticality.

Could the observed divergence be a true divergence between expert judgments and non-expert judgments? We approach this possibility by looking at the concrete items and the corresponding ratings. Consider the selection of OK-items with unexpectedly low ratings, (13a) to (13h), and the selection of *-items with

unexpectedly high ratings, (14a) to (14h). See Appendix B for more items, and the corresponding *LI* papers.

(13) OK-items with low ratings (including standard deviations):

a. [OK]*John is much taller than Mary than Bill is*. 2.50 (1.72)
b. [OK]*Who must Bill have said that Susan married*. 3.02 (1.63)
c. [OK]*Mugsy Boags wasn't very tall a basketball player*. 3.32 (2.18)
d. [OK]*There will arrive a man tomorrow*. 3.33 (1.58)
e. [OK]*Who thinks that Susan talked with who*. 3.39 (1.60)
f. [OK]*Every investigator of one of these languages seems to his supervisor to be brilliant, but won't tell you which of the languages*. 3.80 (1.87)
g. [OK]*Alice met each man taller than my father*. 3.93 (1.69)
h. [OK]*How did they believe, and Mary claim, that Peter had murdered John*. 4.17 (1.89)

(14) *-items with high ratings (including standard deviations):

a. **John beseeched for Harriet to leave*. 4.44 (1.76)
b. **If you want good cheese, you only ought go to the North End*. 4.52 (1.73)
c. **There are a cat and a dog in the yard*. 4.54 (2.31)
d. **John appears to hit Bill right now*. 4.55 (1.66)
e. **John said to take care of himself*. 5.02 (1.69)
f. **October 1st, he came back*. 5.74 (1.31)
g. **John pounded the yam yesterday to a very fine and juicy pulp*. 6.02 (1.24)
h. **I read something yesterday John had recommended*. 6.08 (1.19)

For OK-items with unexpectedly low ratings, there are two possibilities. Either they are ungrammatical after all and should have been marked with an "*". Or they are indeed grammatical but confounded by extra-grammatical factors. For the OK-items in (13a) to (13h), we find it hard to pinpoint the exact factors and we can only speculate about the reasons. In the example of (13a), our intuition is that it is not too hard to parse and that the meaning is clear. Comparing the size of differences is not very common, at least not with the structure in (13a). Therefore, the reader might expect a simple comparison (*John is much taller than Mary*) and experience a garden path when the sentence continues.[15] The low judgment probably reflects both – processing difficulties and low frequency of usage. Future research on the processing of comparatives could take this finding as a starting point and test whether (13a) indeed involves garden pathing. Low frequency, on the other

15 We would like to thank an anonymous reviewer for pointing out the potential garden path, which we missed in the screening of our experimental items.

hand, might also be the driving factor for comparatively low acceptability ratings for (13c). So, one might view it as grammatical but rather unacceptable. The current standard use of diacritics does not distinguish between (un)grammatical vs. (un)acceptable – the two dimensions are collapsed into one. We think that using separate diacritics to indicate grammatical status (a theoretical decision) versus acceptability (a decision based on perception) would be useful. (13c) is an example where the reader would benefit from disentangling the two dimensions. The same reasoning applies to (13f), which could be viewed as grammatical but as too complex to be fully acceptable. Again, this finding calls for follow-up studies to identify the exact source of the surmised processing difficulty and how it relates to structural properties of (13f).

For *-items with unexpectedly high ratings, there are three possibilities. First, they are grammatical and should have been marked as OK. Second, they are grammatical, but their acceptability is degraded due to extra-grammatical factors. Third, they are linguistic illusions, i. e. they are ungrammatical, but their acceptability is ameliorated. However, our item set did not include any known linguistic illusion.[16]

Consider for example (14h), extracted from Fox (2002). Fox discusses the extraposition of *yesterday* in the context of a relative clause and how such an extraposition is grammatical when the *that*-relativizer is realized and ungrammatical when it is not realized. The status of such extrapositions has theoretical implications. While Fox judges the version without the *that*-relativizer to be ungrammatical, the majority of participants in the study found it OK. It is hard to say why the participants do so exactly and one would have to follow up on this. However, considering the ratings, it seems more plausible to view (14h) as a grammatical sequence that is slightly degraded by extra-grammatical factors. Either way, the observed divergence should give rise to further discussion and research. One might find an explanation of why grammaticality and acceptability diverge in this example or one might adjust the grammatical status of (14h). However, it is likely that a change in grammatical status will have theoretical repercussions.

Another example is (14f), from Landau (2007). Landau discusses restrictions on fronting of time adverbials and how punctual time adverbials like *October 1st* cannot be fronted. Accordingly, Landau views *He came back October 1st* as grammatical, but *October 1st, he came back* as ungrammatical. These restrictions are discussed in the wider context of Government-Binding Theory (Landau 2007). In

16 In principle, there is a fourth option, viz. that the items are of intermediate grammaticality. However, we have no evidence that the authors in our sample assume a gradient grammar, hence this possibility is very unlikely.

contrast to Landau, the majority of participants in our study were OK with (14f). Again, it would be interesting to follow up on the question of why exactly authors and non-experts diverge, i. e. why is it that non-experts license this use of punctual time adverbials? Landau argues that other punctual time adverbials, anaphoric ones (*That day, he met Jane*; Landau 2007), can indeed be fronted. Considering our experimental results, one could for example hypothesize that punctual time adverbials can be fronted in general. Either way, the observed divergence again presents an opportunity for further discussion and research.

An in-depth discussion of all the instances of divergence is beyond the scope of the present paper. The instances of discussed above, have one thing common: They represent input for future investigations, systematically testing the constructions, using different methods. As argued, it is likely that the field would benefit from considering such additional data. And to find such divergencies in the first place, a combination of different methods is needed. Therefore, we are worried that an over-reliance on the process of community agreement could have negative consequences for theory building, as it is likely that the resulting theories cannot reflect the depth of the underlying linguistic reality as adequately as theories built on richer data. In summary, looking at our experimental items and the corresponding expert judgments, we think that the observed divergence is not caused by other, secondary factors. Instead, it is a true divergence in judgments.

This is not saying that the linguists "got it wrong" and that the method of researcher introspection and the process of community agreement give "wrong" results. Instead, we interpret our results such that a combination of methods can reveal a level of complexity that is harder to access through the process of community agreement only. Each divergence is an interesting observation and we could use them as a starting point for further investigation.

Further, one should keep in mind that we have looked at sequences in American English. American English is arguably the language that has the strongest representation in the literature. Author judgments in other, less well-represented languages are likely to represent a truncated process of finding community agreement. Thus, the divergence rate between author judgments and experimental results is likely to come out higher in less well-represented languages (for experimental evidence regarding Hebrew and Japanese see Linzen and Oseki 2018).

The process of community agreement certainly has its place in linguistics. First, over the years, the process of community agreement has been used to collect data of good quality. Second, the process is useful to explore new ideas and phenomena. During the exploration phase, the process of community agreement provides an effective means to identify potential factors which later on can be enriched by data from quantitative methods. Feedback from other researchers who are informed and take an interest in the outcome is valuable during this phase.

Third, quantitative methods are very difficult to use in the context of field research and language documentation. Field researchers and language documenters primarily rely on a limited number of consultants. This also means that the roles of researcher and informant/consultant are not conflated. Thus, even if there was a potential researcher bias, it does not affect the results as much, which is likely to mitigate the effect of a low N.

Clear cut cases might not need experimental validation (Pullum 2003 makes a similar point with respect to corpus analyses), neither do items that are secondary to one's theory. However, if there are any doubts or if an item is essential to one's theory, then quantitative validation is desirable, because if things had been clear-cut, we would not have observed the relative lack of convergence in the first place. At the very least, unexpected quantitative results should be a starting point for further investigation.

We advocate a multi-method approach. Quantitative methods are not an end in itself. Quantitative methods without any initial reflection are probably less likely to succeed. And whatever the method of data collection, whether researcher introspection, the process of community agreement, a corpus analysis, an acceptability judgment task, reaction time measurements, etc., it takes an expert to prepare the queries for such tasks and to make sense of their output. Concretely, it requires comprehensive knowledge to apply grammatical reasoning to get from any method's output to the grammatical status of the linguistic sequence in question. From there, it again requires an expert to contrast, generalize, and critically, to theorize. An increase of quantitative data does not lead to a reduced need for theorizing.

7 Conclusion

We have argued that the term *introspection*, commonly used for data collection that conflates the roles of researcher and informant, is a misnomer because every acceptability judgment is an introspective judgment. Moreover, the judgments in the literature typically reflect more than a single person's evaluation of a given string. As an alternative, we therefore introduced the term *process of community agreement* to characterize the arguably most common method of data collection in syntactic enquiry.

Our study examined the degree to which data obtained via the process of community agreement converge with acceptability judgments obtained in a quantitative experiment. Our data consists of linguistic sequences of American English, arguably the language that is best represented in the literature. The results show

agreement between *LI* judgments and experimental judgments but there is also a considerable degree of divergence between the two. We argue that the relative lack of convergence is not caused by performance factors or other extra-grammatical factors that influence the judgments by non-experts in a different way compared to the experts. Arguably, scale biases also only play a minor role at best.

Instead, we think that the divergence comes from an overreliance on the process of community agreement. Overall, the process of community agreement provides good data. However, in our view, this process combined with quantitative methods results in even better data. Under this view, each divergence between methods could be seen as an opportunity for further investigation, systematically testing constructions, using a multi-method approach. Doing so could be a chance to advance the field one step further. Arguably, this point becomes even more important for less well-represented languages, as work by Linzen and Oseki (2018) indicates. However, further evidence is needed to make definite conclusions about other languages. As to American English, we conclude that a combination of the process of community agreement and well-administered quantitative methods are better able to capture the richness and complexity of the linguistic reality than just expert judgments alone.

Acknowledgment: We extend our deepest gratitude to the following for their invaluable feedback and support (in alphabetical order): Ash Asudeh, Mary Dalrymple, Gisbert Fanselow, Gerhard Jäger, Greg Kochanski, Elke Teich, and Tom Wasow. Further, we would like to thank the reviewers, especially Ingo Plag, for their constructive and helpful feedback.

References

Adger, David & Gillian Ramchand. 2003. Predication and equation. *Linguistic Inquiry* 34(3). 325–359. https://doi.org/10.1162/002438903322247515.

Altman, Douglas G. & J. Martin Bland. 1995. Absence of evidence is not evidence of absence. *British Medical Journal* 311. 485. https://doi.org/10.1136/bmj.311.7003.485.

Bard, Ellen G., Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68. https://doi.org/10.2307/416793.

Bates, Douglas, Martin Mächler & Ben Bolker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Bolinger, Dwight. 1978. Asking more than one thing at a time. In Henry Hiz (ed.). *Questions*. 107–150. Dordrecht: Reidel. https://doi.org/10.1007/978-94-009-9509-3_4.

Bornkessel-Schlesewsky, Ina & Matthias Schlesewsky. 2007. The wolf in sheep's clothing: Against a new judgment-driven imperialism. *Theoretical Linguistics* 33(3). 319–333. https://doi.org/10.1515/tl.2007.021.

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

Brians, Paul. 2014. *Common errors in English usage*. https://brians.wsu.edu/common-errors/ (accessed 20 December 2014).

Chomsky, Noam. 1973. Conditions on transformations. In Stephen R. Anderson & Paul Kiparsky (eds.). *A festschrift for Morris Halle*. 232–286. New York: Holt, Rinehart & Winston.

Chomsky, Noam & George A. Miller. 1963. Introduction to the formal analysis of natural languages. In Duncan R. Luce, Robert R. Bush & Eugene Galanter (eds.). *Handbook of mathematical psychology. Vol. II*, 269–321. New York: Wiley.

Christiansen, Morton H. & Maryellen C. MacDonald. 2009. A usage-based approach to recursion in sentence processing. *Language Learning* 59(1). 126–161. https://doi.org/10.1111/j.1467-9922.2009.00538.x.

Clifton, Charles Jr., Gisbert Fanselow & Lyn Frazier. 2006. Amnestying superiority violations: processing multiple questions. *Linguistic Inquiry* 27(1). 51–68. https://doi.org/10.1162/002438906775321139.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20. 37–46.

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. London: Sage Publications.

Culicover, Peter W. & Ray Jackendoff. 2010. Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences* 14(6). 234–235. https://doi.org/10.1016/j.tics.2010.03.012.

den Dikken, Marcel, Judy B. Bernstein, Christina Tortora & Raffaella Zanuttini. 2007. Data and grammar: Means and individuals. *Theoretical Linguistics* 33(3). 335–352. https://doi.org/10.1515/tl.2007.022.

Downs, Julie S., Mandy B. Holbrook, Steve Sheng & Lorrie F. Cranor. 2010. Are your participants gaming the system?: Screening Mechanical Turk Workers. *CHI'10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2399–2402. https://doi.org/10.1145/1753326.1753688.

Edelman, Shimon & Morten H. Christiansen. 2003. How seriously should we take minimalist syntax? *Trends in Cognitive Sciences* 7(2). 60–61. https://doi.org/10.1016/s1364-6613(02)00045-1.

Fanselow, Gisbert, Caroline Féry, Matthias Schlesewsky & Ralph Vogel (eds.). 2006. *Gradience in grammars: Generative perspectives*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199274796.003.0001.

Featherston, Sam. 2005. The decathlon model of empirical syntax. In Marga Reis & Stephan Kepser (eds.). *Linguistic evidence. Empirical, theoretical and computational perspectives*. 187–208. Berlin & New York: de Gruyter. https://doi.org/10.1515/9783110197549.187.

Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33(3). 269–318. https://doi.org/10.1515/tl.2007.020.

Featherston, Sam. 2008. Thermometer judgments as linguistic evidence. In Claudia M. Riehl & Astrid Rohte (eds.). *Was ist Linguistische Evidenz?* 69–89. Aachen: Shaker Verlag.

Ferreira, Fernanda & Benjamin Swets. 2005. The production and comprehension of resumptive pronouns in relative clause "island" contexts. In Anne Cutler (ed.). *Twenty-first century psycholinguistics: Four cornerstones*. 263–278. Mahwah, NJ: Lawrence Erlbaum Associates.

Fodor, Janet D., Stefanie Nickels & Esther Schott. 2017. Center-embedded sentences: What's pronounceable is comprehensible. In Roberto G. de Almeida & Lila R. Gleitman (eds.). *On concepts, modules, and language: Cognitive science at its core*. 139–168. Oxford: Oxford University Press.

Fox, Danny. 2002. Antecedent-contained deletion and the copy theory of movement. *Linguistic Inquiry* 33(1). 63–96. https://doi.org/10.1162/002438902317382189.

Frazier, Lyn. 1985. Syntactic complexity. In David R. Dowty, Lauri Karttunen & Arnold M. Zwicky (eds.). *Natural language parsing. Psychological, computational and theoretical perspectives*. 129–189. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511597855.005.

Frazier, Lyn. 2008. Processing ellipsis: A processing solution to the undergeneration Problem? In Charles B. Chang & Hannah J. Haynie (eds.). *Proceedings of the 26th West Coast Conference on Formal Linguistics*. 21–32. Somerville, MA: Cascadilla Proceedings Project.

Gibson, Edward & Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14(6). 233–234. https://doi.org/10.1016/j.tics.2010.03.005.

Gibson, Edward & Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1–2). 88–124. https://doi.org/10.1080/01690965.2010.515080.

Gibson, Edward & James Thomas. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes* 14(3). 225–248. https://doi.org/10.1080/016909699386293.

Gibson, Edward, Steven Piantadosi & Evelina Fedorenko. 2013. Quantitative methods in syntax/semantics research: A response to Sprouse & Almeida (2013). *Language and Cognitive Processes* 28(3). 229–240. https://doi.org/10.1080/01690965.2012.704385.

Ginzburg, Jonathan & Ivan A. Sag. 2001. *Interrogative investigations: The form, meaning, and use of English interrogatives*. Stanford, CA: CSLI Publications.

Grewendorf, Günther. 2007. Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics* 33(3). 369–381. https://doi.org/10.1515/tl.2007.024.

Hansen, Klaus, Uwe Carls & Peter Lucko. 1996. *Die Differenzierung des Englischen in nationale Varianten. Eine Einführung*. Berlin: Erich Schmidt Verlag.

Hendrick, Randall & Michael Rochemont. 1982. Complementation, multiple *wh*, and echo questions. Ms. http://twpl.library.utoronto.ca/index.php/twpl/article/download/6392/3380 (accessed 18 July 2017).

Hofmeister, Philip, Laura S. Casasanto & Ivan A. Sag. 2014. Processing effects in linguistic judgment data: (Super-)additivity and reading span scores. *Language and Cognition* 6(1). 111–145. https://doi.org/10.1017/langcog.2013.7.

Hofmeister, Philip & Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language* 86(2). 366–415.

Hornstein, Norbert. 1995. *Logical form*. Cambridge, MA: Blackwell.

Karlsson, Fred. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics* 43(2). 365–392. https://doi.org/10.1017/s0022226707004616.

Kazai, Gabriella, Jaap Kamps & Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. *Proceedings of the Twentieth International Conference on Information and Knowledge Management (ACM CIKM)*. 1941–1944. https://doi.org/10.1145/2063576.2063860.

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees*

of grammaticality. Ph. D. dissertation. University of Edinburgh.
Kilpatrick, Franklin P. & Hadley Cantril. 1960. *Self-anchoring scaling: A measure of individuals' unique reality worlds*. Washington, DC: The Brookings Institution.
Kövecses, Zoltan. 2000. *American English – an introduction*. Peterborough, ON: Broadview Press.
Krantz, John & Reeshad Dalal. 2000. Validity of Web-based psychological research. In Michael Birnbaum (ed.). *Psychological experiments on the Internet*. 35–60. San Diego, CA: Academic Press. https://doi.org/10.1016/B978-012099980-4/50003-4.
Kuno, Susumo & Jane J. Robinson. 1972. Multiple *wh*-questions. *Linguistic Inquiry* 3(4). 463–487.
Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159–174.
Langsford, Steve, Aamy Perfors, Andrew T. Hendrickson, Lauren A. Kennedy & Danielle J. Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics* 3(1). 37. http://doi.org/10.5334/gjgl.396.
Lasnik, Howard & Mamoru Saito. 1984. On the nature of proper government. *Linguistic Inquiry* 15(2). 235–289.
Lau, Jey H., Alexander Clark & Shalom Lappin. 2016. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41(5). 1202–1241. https://doi.org/10.1111/cogs.12414.
Linzen, Tal & Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: A Journal of General Linguistics* 3(1). 100. http://doi.org/10.5334/gjgl.528.
Landau, Idan. 2007. EPP Extensions. *Linguistic Inquiry* 33(3). 485–523. https://doi.org/10.1162/ling.2007.38.3.485.
Mason, Winter & Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1). 1–23. https://doi.org/10.3758/s13428-011-0124-6.
Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen & Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HTL 2010 Workshop Creating Speech and Text Language Data with Amazon's Mechanical Turk*. 122–130. Los Angeles, CA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W10-0719/.
Nugent, William R. 2004. A validity study of scores from self-anchored-type scales for measuring depression and self-esteem. *Research on Social Work Practice* 14(3). 171–179. https://doi.org/10.1177/1049731503257879.
Pesetsky, David & Esther Torrego. 2001. T-to-C movement: causes and consequences. In Michael Kenstowicz (ed.). *Ken Hale: A life in language*. 355–426. Cambridge, MA: MIT Press.
Pesetsky, David. 1987. *Wh*-in-situ: Movement and unselective binding. In Eric J. Reuland & Alice G. B. ter Meulen (eds.). *The representation of (in)definiteness*. 98–129. Cambridge, MA: MIT Press.
Phillips, Colin. 2010. Should we impeach armchair linguists? In Shoishi Iwasaki, Hajime Hoji, Patricia M. Clancy & Sung-Ock Sohn (eds.). *Japanese-Korean Linguistics*, vol. 17. 49–64. Stanford, CA: CSLI Publications.
Phillips, Colin & Howard Lasnik. 2003. Linguistics and empirical evidence: Reply to Edelman

and Christiansen. *Trends in Cognitive Sciences* 7(2). 61–62. https://doi.org/10.1016/S1364-6613(02)00045-1.

Phillips, Colin, Matthew W. Wagers & Ellen F. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In Jeffrey T. Runner (ed.). *Experiments at the interfaces (Syntax and Semantics 37)*. 153–186. Bingley: Emerald Publications. https://doi.org/10.1163/9781780523750_006.

Pinker, Steven & David Birdsong. 1979. Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior* 18(4). 497–508. https://doi.org/10.1016/s0022-5371(79)90273-1.

Pullum, Geoffrey K.. 2003. Corpus fetishism. In *Language Log*. http://itre.cis.upenn.edu/~myl/languagelog/archives/000122.html (accessed 20 December 2014).

R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org (accessed 31 May 2015).

Reich, Peter A. 1969. The finiteness of natural language. *Language* 45(4). 831–843. https://doi.org/10.2307/412337.

Schnoebelen, Tyler & Victor Kuperman. 2010. Using Amazon Mechanical Turk for linguistic research. *Psihologija* 43(4). 441–464. https://doi.org/10.2298/psi1004441s.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, IL: The University of Chicago Press. https://doi.org/10.26530/oapen_603356.

Sorace Antonella. 1992. Lexical conditions on syntactic knowledge: Auxiliary selection in native and non-native grammars of Italian. Ph.D. dissertation. University of Edinburgh.

Sorace, Antonella & Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115(11). 1497–1524. https://doi.org/10.1016/j.lingua.2004.07.002.

Sprouse, Jon. 2009. Revisiting Satiation: Evidence for an equalization response strategy. *Linguistic Inquiry* 40(2). 329–341.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167. https://doi.org/10.3758/s13428-010-0039-7.

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's 'Core Syntax'. *Journal of Linguistics* 48(3). 609–652. https://doi.org/10.1017/s0022226712000011.

Sprouse, Jon & Diogo Almeida. 2013. The role of experimental syntax in an integrated cognitive science of language. In Cedric Boeckx & Kleanthes K. Grohmann (eds.). *The Cambridge handbook of biolinguistics*. 181–202. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511980435.013.

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. https://doi.org/10.1016/j.lingua.2013.07.002.

Stevens, Stanley S. 1946. On the theory of scales of measurement. *Science* 103(2684). 667–688. https://doi.org/10.1126/science.103.2684.677.

Stroik, Thomas. 2001. On the light verb hypothesis. *Linguistic Inquiry* 32(2). 362–369. https://doi.org/10.1162/ling.2001.32.2.362.

Szabolcsi, Anna. 2006. Strong and weak islands. In Martin Everaert & Henk van Riemsdijk (eds.). *The Blackwell companion to syntax*, vol. 4. 479–532. Oxford: Blackwell.

Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115(11). 1481–1496. https://doi.org/10.1016/j.lingua.2004.07.001.

Wellwood, Alexis, Roumyana Pancheva, Valentine Hacquard & Colin Phillips. 2018. *Journal of Semantics* 35(3). 543–583. https://doi.org/10.1093/jos/ffy014.

Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273. https://doi.org/10.1353/lan.2011.0041.

Zhu, Dongqing & Ben Carterette. 2010. An analysis of assessor behavior in crowdsourced preference judgments. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*. 21–26. http://ir.ischool.utexas.edu/cse2010/materials/zhucarterette.pdf.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/zfs-2020-2008).The supplemental file provides an overview over the methodological differences to Sprouse et al. (2013), as well as further information about the sampled items.