Tom S. Juzek & Jana Häussler

2019

Semantic Influences on Syntactic Acceptability Ratings

In

A. Gattnar, R. Hörnig, M. Störzer & S. Featherston (Eds.)

Proceedings of Linguistic Evidence 2018: Experimental Data Drives Linguistic Theory

Tübingen: University of Tübingen

https://publikationen.uni-tuebingen.de/xmlui/handle/10900/87132

In A. Gattnar, R. Hörnig, M. Störzer & S. Featherston (Eds.) (2019). *Proceedings of Linguistic Evidence 2018: Experimental Data Drives Linguistic Theory*. Tübingen: University of Tübingen. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/87132

Semantic Influences on Syntactic Acceptability Ratings

Tom S. Juzek[1] & Jana Häussler[2]
[1]Saarland University, [2]University of Bielefeld
tom.juzek@posteo.net, jana.haeussler@uni-bielefeld.de

1 Introduction

The existence of gradience in syntactic acceptability has been acknowledged at least since Chomsky (1965) and is widely assumed in contemporary grammar theory (cf. Myers, 2009; Sorace & Keller, 2005). However, its true prevalence has only become clearer recently, with Featherston (2005) showing a linear pattern for a huge data set (1000 items) and Häussler et al. (2016) for a data set that is controlled as far as possible for a number of extra-grammatical factors and therefore makes explanations based solely on known performance factors less likely. Proponents of categorical grammars would explain the observed gradience mainly through performance factors (cf. Fanselow & Frisch, 2006; Hawkins, 2006; Newmeyer, 2003; Sprouse, 2007), while proponents of a gradient grammar allow for some of the observed gradience to come from the grammar itself (cf. Featherston, 2005; Keller, 2000; Manning, 2003; Sorace & Keller, 2005; Wasow, 2009). In the present paper, we look into the question of whether another extra-grammatical factor, semantic influences, could account for parts of the observed gradience. We conducted an experiment comparing the effect of syntactic and semantic deviations on acceptability ratings, viz. agreement violations and two types of semantic anomalies. Our results suggest that semantic influences can have a degrading effect on the acceptability of grammatical items. However, we did not observe that they had an ameliorating effect. A second experiment tests the opposite direction, i.e., the effect of syntactic violations on semantic ratings. Since we do not observe a grammaticality effect in this experiment, we conclude that the effects in the first experiment are not due to a general inability to distinguish the two rating tasks.

In the remainder of this section, we provide further background. We revisit the most relevant concepts of the debate and related literature, including the question of where the gradience in acceptability stems from. In Section 2 and Section 3, we present our experiments. The paper is concluded by a general discussion in Section 4.

1.1 Acceptability judgements

Syntacticians are interested in the grammatical status of sentences, the common unit of syntactic enquiry, and then use this information to describe and explain the structure of human language(s). To infer the grammatical status of a linguistic expression, syntacticians make use of various methods. Arguably, the most common methods include researcher introspection, where the investigating linguist is their own informant, corpus analyses, where language use is analysed, and acceptability judgement tasks. In such a judgement task, one or more third party informants are asked to judge the acceptability of some linguistic unit. Typically, these informants are unaware of the goal of the study, which is one of the motivations to use a judgement task instead of researcher introspection. In researcher introspection, the roles of informant and researcher are conflated and hence the informants know about the aim of the study and are thus potentially biased. When conducting an acceptability judgement task, the researcher has to

make various choices like how to administer the test, which scale to use, how many participants to include, and so on.[1]

1.2 Grammaticality, acceptability, and grammatical reasoning[2]

There is no and cannot be any method that observes the grammaticality of a linguistic sequence directly. The grammar is a mental construct and can only be observed indirectly. In the case of a judgement task, the researcher collects acceptability judgements rather than grammaticality judgements (though often labelled as such). These judgements are an outcome of the competence grammar and various extra-grammatical effects, in particular performance factors. Common performance factors are memory load (Keller, 2000: 28), real-world implausibility (Sprouse, 2013), and ambiguity (Myers, 2009: 409), all of which can make a grammatical sequence seem less acceptable, sometimes even unacceptable.

To illustrate the distinction between grammaticality and acceptability, consider (1), which is an example of a multiple center-embedding, taken from Chomsky & Miller (1963: 286). (1) does not violate any known grammatical rule, but to many it seems unacceptable because of its complexity.

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

Although less common, there are also examples that are ungrammatical and yet acceptable. Frazier (2008) discusses examples of what she calls *acceptable ungrammaticality* and refers to Otero (1972) and Langendoen & Bever (1973) for early research on the topic. Acceptable ungrammaticality can come in various forms and a more recent example is the comparative illusion, illustrated in (2), taken from Phillips et al. (2011).

(2) More people have been to Russia than I have.

Crucially, it takes grammatical reasoning to get from the output of any method to the grammatical status of a linguistic sequence. This applies to all methods, including researcher introspection, corpus analyses, and acceptability judgement tasks. In the case of acceptability judgements, the researcher's task is to make sense of the judgements, to contrast judgements, to generalise, and to theorise. This includes to identify the source of (1)'s markedness in acceptability. The inference is not always that straightforward as evidenced in the debate on island violations (for an overview see Boeckx, 2012) or superiority effects in multiple *wh*-questions (cf. Häussler et al., 2015: Hofmeister et al., 2013). Unfortunately, performance factors are sometimes stated without specifying them. Postulating concrete factors and testing them experimentally seems crucial to us.

1.3 Gradience in acceptability

While it is disputed whether grammaticality is a categorical or a gradient concept, there is consensus that there is gradience in acceptability. Gradience in acceptability has been acknowledged at least as early as Chomsky (1965) and has been discussed further in, e.g., Featherston (2005), Keller (2000), Newmeyer (2003), Sorace & Keller (2005), Sprouse (2007), Wasow (2009), to name a few. Featherston (2005) provided the first comprehensive illustration of the pervasiveness of gradience in acceptability. Featherston plotted about 1000 acceptability ratings and observed that they line up in a near-linear fashion as shown in Figure 1.

It is surprising how prevalent gradience in acceptability is, which is expressed in (Q). To explain (Q), proponents of gradient grammars would allow that parts of the gradience stem

[1] Cowart (1997) is an excellent introduction with respect to such methodological questions. For a reflection on the empirical issues in general, see, e.g., Schütze (1996) or contributions in Schindler et al. (forthcoming).
[2] This subsection follows in parts the discussion of the same topic in Juzek & Häussler (ms).

from the grammar itself, but they would also look at extra-grammatical factors as causes. Proponents of categorical grammars have to fully resort to extra-grammatical factors for this.

(Q) How can the high prevalence of gradience in acceptability be explained?

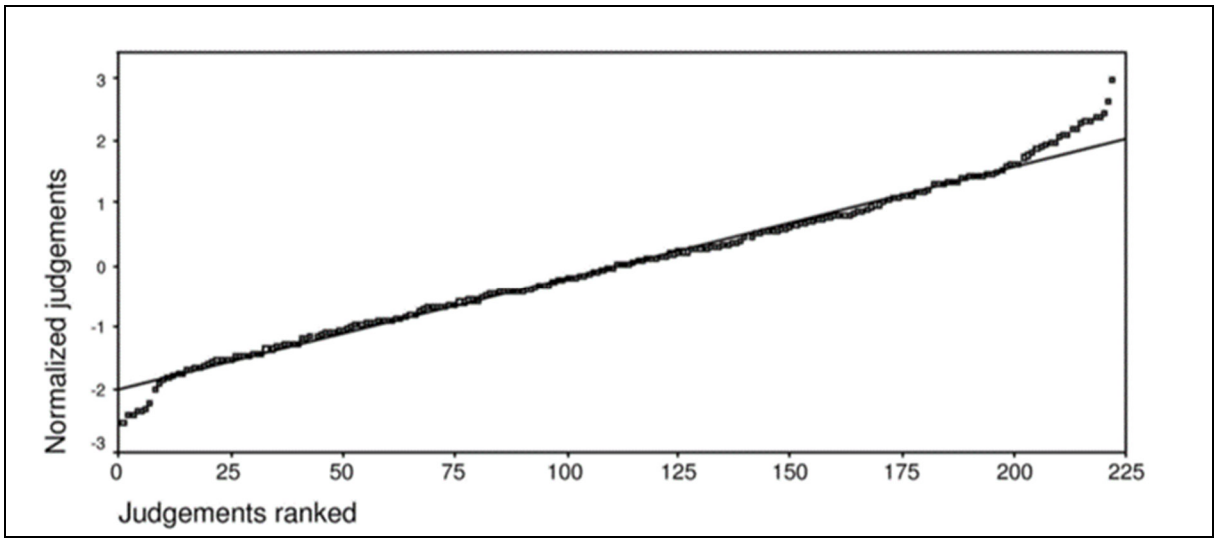

Figure 1 [taken from Featherston (2005: 5)]. Original caption: "Judgements elicited under controlled conditions produce a linear pattern of well-formedness."

Häussler et al. (2016) contributed to the ongoing discussion of (Q). They sampled 100 items from articles published in *Linguistic Inquiry* 2001-2010, 50 items marked with an asterisk in the paper (*-items) and 50 items that were unmarked in the paper (OK-items). All items were

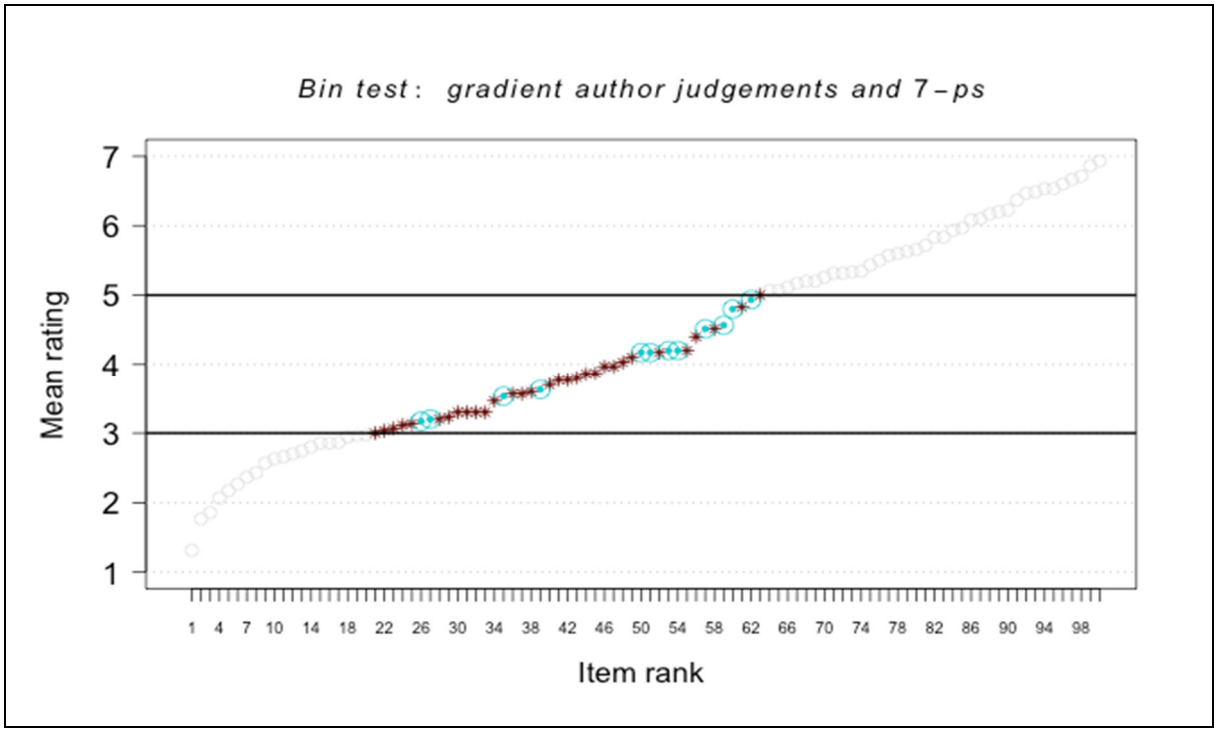

Figure 2 [taken from Häussler et al. (2016: 1), figure layout slightly adjusted]. Mean ratings by the non-linguists (y-axis) for 100 items extracted from articles published in *Linguistic Inquiry* (x-axis; in ascending order). Items with a mean rating of ≥ 3 and ≤ 5 are in red (star marked in the corresponding *LI* article) or blue (unmarked in the *LI* article)

sampled from papers in which the authors use three or more judgement categories, which is typically indicated by some form of '*?', '??', '?', etc., but only examples with '*' or no diacritic were selected, in other words: endpoints. Häussler et al. (2016) then collected acceptability judgements by non-linguists in an online experiment, using a 7-point-scale. Importantly, Häussler et al. (2016) had pre-screened their experimental items for potentially confounding factors like garden paths, linguistic illusions, unresolved references, and alike, and also checked their results for length effects, which were negligible. Häussler et al. (2016) then compared author judgements to mean ratings by the non-linguists and they divided the resulting acceptability space into three bins, where the mid-bin included items with a rating ranging from 3 to 5.[3] As evident in Figure 2, ratings form a linear pattern and the majority of items in the mid-bin are items that were marked with an asterisk in the corresponding *LI* article.

1.4 The source of gradience

Häussler et al. (2016) observe that 43 % of all items fell into the mid-bin, 77 % of which were marked with an asterisk in the paper from which they were sampled. Their observations lead to (P), which is a modification of (Q).

(P) How can the high prevalence of gradience in acceptability be explained, considering that (i) authors were in principle using a third category like '?' but decided to not use it for the items in question, and considering that (ii) the majority of the gradient items were marked with an asterisk in the paper from which they were sampled.

In approaching (P), we make the following assumptions.

(A1) Grammaticality is categorical. This assumption is not necessary, but (P) takes its strongest form if a categorical grammar is assumed.

(A2) (P) cannot be explained by scale biases, i.e., that the authors from which the items were sampled conceive '*' and OK differently. Scale biases are likely to occur, but they are unlikely to have a negative effect, because authors had at least one more category between '*' and OK available, viz. some form of '?'. This should have led to a clustering at the bottom and the top of the scale, mitigating scale effects.

(A3) The authors are not wrong, i.e., it is not the case that the *-items in the mid-bin should have been OK-items that are degraded by performance factors.

(A4) The observed gradience is not due to aggregation, i.e., it is not the case that the individual participants gave endpoint ratings which are concealed when analysing means. Juzek & Häussler (ms) inspected the distribution of the ratings by individual participants and found that the majority of them do indeed give in-between ratings.

(A5) Scale effects cannot fully explain the observed gradience. Scale effects occur when certain parts of a scale are originally not used, but then other items fill in that space. In Häussler & Juzek (ms), we followed up on the possible impact of scale effects and found that scale effects can explain some of the observed gradience, but they cannot explain it to a sufficient degree.

If one endorses (A1) to (A5), then one has to find other extra-grammatical factors to explain (P). Crucially, such factors should explain why ungrammatical items ameliorate. As mentioned, semantic influences, e.g., real-world implausibility, can degrade grammatical items. And Häussler & Juzek (2017) speculate whether semantic influences could ameliorate ungrammatical items. Inspired by this, we think that there is a possibility that plausibility could ameliorate the acceptability of ungrammatical items, in addition to the degrading effect of implausibility

[3] For details about their methodology, see Chapter 1 in Juzek (2016).

on grammatical items. This leads to the first experiment, in which we check for semantic influences in a syntactic acceptability judgement task. In a second experiment, we then check for syntactic influences in a semantic acceptability judgement task.

2 Experiment 1

Semantics might be one of the extra-grammatical factors contributing to the well-observed gradience in acceptability. Grammatical sentences involving a semantic anomaly might appear less acceptable than semantically non-deviant counterparts. Likewise, ungrammatical sentences involving a semantic anomaly might appear even less acceptable than ungrammatical but semantically non-deviant sentences. Furthermore, intelligibility might have an ameliorating effect on otherwise unacceptable (because ungrammatical) sentences.

The experiment addresses the question whether participants are able to ignore semantic anomalies when asked to do so and to concentrate on form for judging the acceptability of a sentence. It examines this question by comparing ratings for semantically deviant sentences and non-deviant counterparts and crossing the semantic anomaly with a syntactic anomaly. We chose two types of semantic anomalies: logical inconsistencies as in (3) and incompatibilities as in (4).

(3) a. #San Diego is both entirely in California and not entirely in California.
b. #Mary and John are atheists and they believe in God.
c. #Annie is my mother and she is my twin sister.

The example in (3a) involves a canonical contradiction. The two conjuncts are formally identical except for the negation. As a result, the truth of the first proposition implies the falsity of the second proposition and vice versa. The contradiction in (3b) is less obvious. Unpacking the lexical meaning of *atheist* as 'person who does not believe in God' makes the negation visible. Example (3c) is an apparent contradiction. The two statements are contrary, i.e., they cannot be true simultaneously, but they are not contradictory in a strict sense because they both can be false, e.g., if Annie is the aunt of the speaker. The examples have in common that they involve an inconsistency due to conflicting meaning contributions. From a semanticist's perspective, the type of inconsistency varies, but for the average participant they will appear equally contradictory (in a non-technical sense).

The second type of semantic anomaly examined in the study is illustrated in (4).

(4) a. #All of the windows were breaking for at least three minutes.
b. #Mary was winning the national lottery for three weeks.

The anomaly in (4) is tied to the *for*-PP. The sentences are perfectly normal once we remove the temporal modifier. The conflict between the *for*-PP and the verbal predicate is rooted in the event structure. The compatibility of a verbal expression with a temporal, more precisely a durative *for*-PP figures prominently in research on aspectuality and verb semantics (e.g., Dowty, 1979; Vendler, 1957). Vendler (1957) introduced *for*-PPs as diagnostics for distinguishing between achievements (*reach the summit, win a race*) and accomplishments (*run a mile, write a letter*) on the one hand, and states (*know something, love somebody*) and activities (*run, push a cart*) on the other hand. Only the latter two are compatible with *for*-PPs (*know something for quite some time, love somebody forever; run for an hour, push a cart for half an hour*) because only the latter last for a period of time. Achievements in contrast are instantaneous changes of state. Though the event presupposed by a verbal predicate like *to win a race* might last a while, the actual event of winning the race occurs in the single moment the winner is crossing the finish line. Hence, the event has no duration and therefore cannot be modified by a durative adverbial.

Semantic anomalies of the first type might be easier to ignore when asked to give a form-based acceptability judgement in comparison incompatibilities between verbal predicate and

modifier. The conflict between the verb and the durative *for*-PP, though being semantic in nature, is related to a particular expression in the sentence and hence to the form, viz. the *for*-PP. Therefore, participants might be less certain as to whether the perceived conflict is a meaning issue or a form issue. As a result, their syntactic ratings might be more vulnerable to the semantic anomaly resulting from the incompatibility of verb and durative *for*-PP.

Furthermore, the conflict with *for*-PPs can be repaired to some extent by means of coercion. In principle, the conflict between a non-durative verb and a durative modifier can be solved by deriving an iterative meaning, in which case the *for*-PP would modify the duration of the series of events, or by linking the *for*-PP to a resultative state rather than the event itself. We tried to reduce both possibilities when constructing the materials. However, if participants arrive at an interpretation in which the conflict is solved in at least some of the trials, we expect higher semantic acceptability ratings for *for*-PPs in comparison to logical inconsistencies.

2.1 Method

2.1.1 Participants

Participants were recruited online using *Prolific* (https://prolific.ac/) and then redirected to a website that hosted the experiment. In order to be eligible, participants had to be native speakers of British English. We excluded data from participants who submitted incomplete results, failed on control items, or had extreme response times, indicating fast clicking through (following Häussler & Juzek, 2016). Data from 37 participants entered statistical analyses.

2.1.2 Materials

The experiment included 32 experimental items plus 32 fillers as well as four calibration items and four control items, making a total 72 items.[4] Experimental items appeared in four versions and were distributed across four lists according to a Latin square design. Each participant saw each experimental item in only one of its versions and an equal number of items in each version. The four versions resulted from fully crossing the two within-items factors *Grammaticality* (+/- grammatical) and *Plausibility* (+/- plausible). A third, between-items factor varied the type of semantic anomaly.

The grammatical status of a sentence was varied by means of subject-verb agreement. In half of the items, the subject was singular; in the other half, the subject was plural. As a result, agreement violations involved a singular verb in half of the cases and a plural verb in the other half. The factor *Plausibility* was varied in two ways. In half of the items, the semantic anomaly was a logical inconsistency, in the other half, it was an incompatibility of verb semantics and adverbial.

We constructed 16 sentences involving a contradiction in the condition [-consistent]. The presence/absence of a logical inconsistency was fully crossed with the syntactic factor *Grammaticality*, yielding four conditions. Accordingly, each item appeared in four versions as exemplified in (5).

(5) a. My sister Jane is married to a lawyer.
b. *My sister Jane are married to a lawyer.
c. #My sister Jane is married to a bachelor.
d.*#My sister Jane are married to a bachelor.

The version in (5a) contains no anomaly, the sentence is both grammatical and semantically normal. The version in (5b), in contrast, is ungrammatical because of the plural verb. Both (5c) and (5d) are semantically deviant. They contain an inconsistency due to the conflict between *married* and *bachelor*. Jane is married to X implies that X is married to Jane. Specifying X as

[4] Due to typos, two items had to be excluded from analysis.

being a bachelor conflicts with the lexical meaning of *bachelor* as being unmarried. (5c) and (5d) differ in their grammatical status. While (5c) is logically impaired but grammatical, (5d) is both inconsistent and ungrammatical.

Logical inconsistency was achieved in several ways. Some of the items involve a contradiction of two statements expressed as conjuncts in a coordinate structure. Others involve two contrary statements. Finally, some inconsistencies occur clause internally, resulting from a semantic conflict between the lexical semantics of two expressions, e.g., the conflict in (5c) and (5d) between *married* and *bachelor*.

Another set of 16 items included a durative *for*-PP which was or was not compatible with the verbal predicate. The *for*-PP was present in all versions of an item. Its compatibility with the verbal predicate was varied by varying the verb. Semantically deviant versions contained a punctual verb, while semantically normal sentences contained a durative verb.[5] Again, the semantic status was fully crossed with the factor *Grammaticality*, yielding four versions as exemplified in (6).

(6) a. The two removers were inspecting the safe for ten minutes.
b. *The two removers was inspecting the safe for ten minutes.
c. #The two removers were dropping the safe for ten minutes.
d.*#The two removers was dropping the safe for ten minutes.

2.1.3 Procedure

Participant recruiting took place via *Prolific*, but for the actual experiment, participants were redirected to a separate website. They first read an instruction and then rated sentences on a 7-point scale with respect to their well-formedness. Experiment 1 asked participants to rate how 'natural' or 'unnatural' sentences appear to them, with respect to the items' grammaticality. The instruction explicitly stated that participants should not be bothered with meaning. To illustrate the distinction, examples were given: *Jack did his job goodly* (meaningful and intelligible, but also not fully grammatical) and *The storm intentionally broke the window* (fully grammatical but implausible). The scale was explained in the instructions and had labels pointing towards the two endpoints ('unnatural/ungrammatical' and 'natural/grammatical') to remind participants of the task throughout the experiment. In addition to labels, we used colour and pictograms: red for the lower part of the scale and blue for the upper part of the scale,[6] plus a cross-mark for the lower bound (for later analysis coded as '1') and a check-mark for the upper bound ('7'). The first four items were calibration items to make participants familiar with the task and to ensure that they use the full scale. Each item was displayed on a separate page together with the rating scale. A warning mechanism displayed a warning message when participants became too fast (cf. Häussler & Juzek, 2016).

2.2 Results

The mean ratings per condition are given in Table 1 and Table 2. For visual inspection see Figure 3 (left plot) in Section 3. The rightmost columns and the bottom row provide the results of pairwise comparisons using two-sided *t*-tests. To adjust for multiple comparisons, we applied Bonferroni correction.

The data show a general grammaticality effect. Regardless of plausibility, acceptability ratings are lower for ungrammatical sentences involving an agreement violation. The factor *Plausibility* has a modulating effect: Grammatical but semantically deviant sentences received a penalty (mean ratings for plausible sentences: 5.85, for implausible sentences: 4.58). Ratings

[5] For a comprehensive discussion of the two verb classes, see Engelberg (2000).
[6] In prior studies we used red and green in analogy to traffic lights and common use, but we switched to red and blue to make the scale accessible for colour-blind participants as well.

for ungrammatical sentences, in contrast, are independent of their semantic status (mean ratings for plausible sentences: 2.75, for implausible sentences: 2.48).

Table 1. Mean acceptability ratings per condition in Experiment 1 for the subset of items involving a logical inconsistency. The rightmost column and the bottom row provide results of *t*-tests for pairwise comparisons.

	+ consistent	- consistent	pairwise comparison
+ grammatical	5.80	4.99	$t = 2.25$, $p = .03$
- grammatical	2.66	2.36	$t = 1.23$, $p = .23$
pairwise comparison	$t = 10.38$, $p < .001$	$t = 8.33$, $p < .001$	

Table 2. Mean acceptability ratings per condition in Experiment 1 for the subset of items involving incompatibilities with *for*-PPs. The rightmost column and the bottom row provide results of *t*-tests for pairwise comparisons

	+ compatible	- compatible	pairwise comparison
+ grammatical	5.91	4.10	$t = 5.80$, $p < .001$
- grammatical	2.83	2.61	$t = 0.90$, $p = .38$
pairwise comparison	$t = -11.01$, $p < .001$	$t = 5.42$, $p < .001$	

2.3 Discussion

Acceptability ratings in Experiment 1 exhibit susceptibility to semantic influences but only in one direction. Implausibility results in degraded acceptability. The effect was restricted to grammatical sentences. Ungrammatical sentences showed a slight tendency in the same direction but not a substantial effect. However, since we chose a strong and obvious grammatical violation, the lack of an additional penalty could be simply a floor effect. Psycholinguistic research has shown that readers are sensitive to agreement violations, even in a language like English, in which agreement is functionally less important for parsing than word order (e.g., Pearlmutter et al., 1999). Agreement violations are often easy to detect and lead to immediate rejection of the corresponding sentence.[7] Garden-path sentences with disambiguation by number agreement exhibit strong effects, stronger than with disambiguation by other means, e.g., case (Meng & Bader, 2000). We consider it therefore likely that the agreement violation dragged acceptability close to the bottom, leaving little room for further degradation by semantic anomaly.

Although it would be interesting to see whether semantics might add to decreased acceptability for other types of syntactic violations, the potential floor effect does not affect our research question. Our point of departure was the consideration that intermediate acceptability could be the result of an ameliorating effect of plausibility. As implausibility can decrease acceptability, plausibility in turn might increase acceptability. Note that the pattern produced by amelioration due to plausibility would be almost indistinguishable from an additional penalty

[7] Agreement errors are less easy to spot in the context of attraction. Sentences such as *The key to the cabinets are rusty* are occasionally processed as if being grammatical. For attraction effects in language comprehension see Pearlmutter et al. (1999), Wagers et al. (2009) and Häussler (2009).

for implausible ungrammatical sentences. In both cases, plausible ungrammatical sentences would receive higher ratings than implausible one. However, an ameliorating effect would not be obscured by a floor effect as one might argue for the additional penalty. The current data corroborate the existing evidence for a degrading effect of implausibility but do not support the conjecture that plausibility has an ameliorating effect in ungrammatical sentences.

Ratings for grammatical but implausible sentences are on average lower than ratings for corresponding semantically normal sentences. Notably, ratings in this condition do also show more variance. This could indicate that participants are less certain how to convert their impression of degradedness to a syntactic acceptability judgement.

Experiment 1 demonstrates susceptibility of acceptability ratings to semantic effects despite the explicit instruction to ignore meaning issues. Does this mean that participants cannot comply with the task? This would be a worrying conclusion. However, the decrease in acceptability for implausible sentences was rather minor. Grammatical sentences suffering from an implausibility did receive upper-intermediate ratings, they were not completely rejected (mean rating for inconsistent sentences: 4.99, mean ratings for sentences with a *for*-PP incompatible with verb semantics: 4.17). Experiment 2 takes up on this worry and collects plausibility ratings for the same set of items. If participants indeed messed up the tasks, we expect the same response pattern as for Experiment 1, and an accordingly high correlation between the two experiments.

3 Experiment 2

Experiment 1 collected acceptability ratings targeting the underlying grammaticality. To reduce the impact of semantic factors, participants were asked to ignore issues related to meaning. Experiment 2 is the mirror image of Experiment 1. This time, participants were asked to rate plausibility and ignore grammatical issues. This way, we address the worry pointed out above that participants who are not linguists might be struggling to follow the instructions of Experiment 1.

3.1 Method

3.1.1 Participants

Similar to Experiment 1, participants were recruited via *Prolific*. Participation was restricted to people who had not participated in Experiment 1. Experiment 2 applied the same exclusion criteria as Experiment 1. Statistical analyses include data from 37 participants.

3.1.2 Materials

Experiment 2 employed the same stimuli as Experiment 1.

3.1.3 Procedure

Experiment 2 collected ratings with respect to the meaningfulness/plausibility of sentences. Scale labels changed accordingly: ‘meaningless/implausible’ and ‘meaningful/plausible’. The instruction asked participants to concentrate on meaning and ignore grammaticality. The same examples as in Experiment 1 were given to illustrate the distinction between grammatical and meaningful. In all other respects, Experiment 2 applied the same procedure as Experiment 1.

3.2 Results

Mean ratings are given in Table 3 and Table 4, for visual inspection see Figure 3 (right plot).

The results show a main effect of *Plausibility*. Implausible sentences received lower ratings than plausible sentences regardless of the type of semantic anomaly and also regardless of their

grammatical status. Pairwise comparisons with two sided *t*-tests, including a Bonferroni correction, confirm this impression of a general effect of *Plausibility*. The penalty for implausibility is slightly less pronounced in the case of *for*-PPs.

Table 3. Mean ratings per condition in Experiment 2, subset of items involving logical inconsistencies. The rightmost column and the bottom row provide results of *t*-tests for pairwise comparisons

	+ grammatical	- grammatical	pairwise comparison
+ consistent	5.67	5.24	$t = 1.63$, $p = .11$
- consistent	3.00	2.99	$t = 0.02$, $p = .98$
pairwise comparison	$t = 7.48$, $p < .001$	$t = 5.48$, $p < .001$	

Table 4. Mean ratings per condition in Experiment 2, subset of items involving incompatibilities with *for*-PPs. The rightmost column and the bottom row provide results of *t*-tests for pairwise comparisons

	+ grammatical	- grammatical	pairwise comparison
+ compatible	5.87	5.09	$t = 2.30$, $p = .03$
- compatible	3.98	3.33	$t = 1.97$, $p = .06$
pairwise comparison	$t = 4.95$, $p < .001$	$t = 5.74$, $p < .001$	

Grammaticality, on the other hand, had only selective effects on plausibility ratings, which failed significance. Ratings for the item set containing incompatibilities (*for*-PPs) show a corresponding trend. Note that Bonferroni-correction changed the significance level to .025.

3.3 Comparing Experiment 1 and Experiment 2

Experiments 1 and 2 obtained two types of ratings for the same set of items, but from two distinct groups of participants. Experiment 1 obtained syntactic acceptability ratings, targeting grammaticality, while Experiment 2 obtained semantic acceptability ratings, targeting plausibility. A direct comparison of the response patterns as given in Figure 3 reveals several differences. This non-uniform behaviour suggests that participants were by and large able to differentiate the two tasks.

Further evidence comes from the only rather weak correlations between syntactic and semantic judgements (for all experimental items $r = .33$). The strength of the correlation differs for the two types of semantic anomalies involved in the experiment (cf. Figure 4). It is lower for (in)consistencies compared to the correlation for the *for*-PPs. For the former, there is only a weak correlation of $r = .22$, for the latter, there is a moderate correlation of $r = .48$. We could interpret this as an indication for the interface character of semantic factors constraining the occurrence of *for*-PPs.

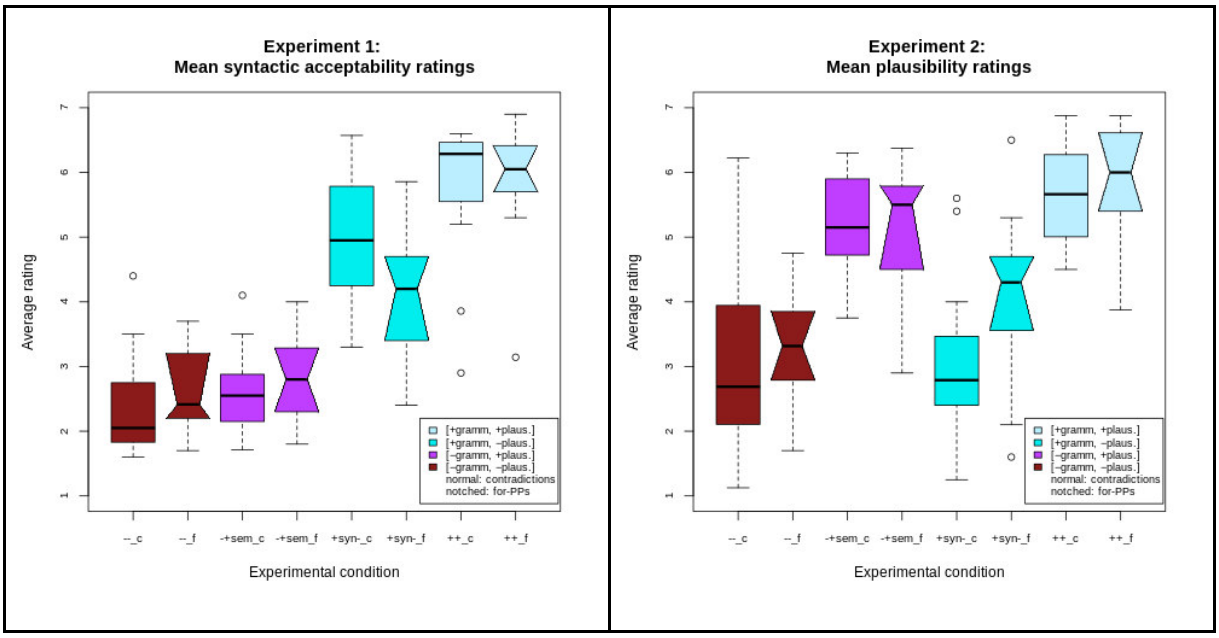

Figure 3. Boxplots for acceptability ratings in Experiment 1 (left plot) and Experiment 2 (right plot), broken down by condition. Signs in the condition labels on the x-axis indicate the level of the two main factors. The first sign indicates *Grammaticality* (+/- grammatical), the second sign indicates *Plausibility* (+/- plausible). The character following the understroke indicates the type of semantic anomaly (c = inconsistency, f = incompatibility with *for*-PP), which is also marked by different shapes (rectangular/normal bars for inconsistencies, notched bars for incompatibilities). To ease comparison, the corresponding semantically non-deviant condition received the same marking

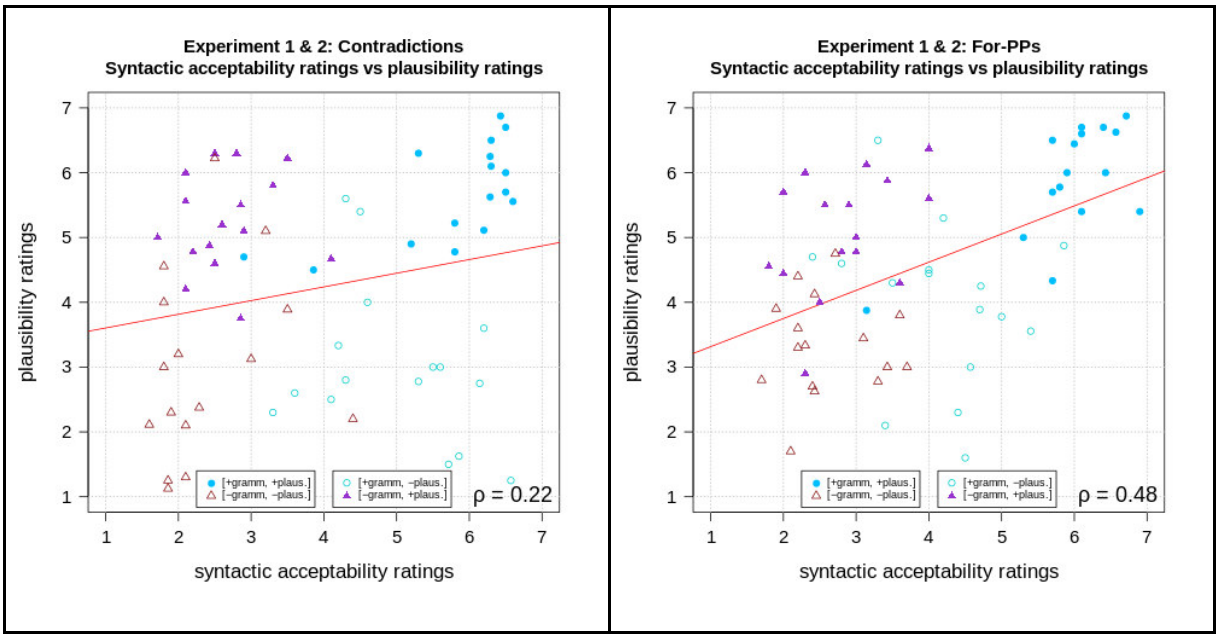

Figure 4. Correlation plots for ratings in Experiments 1 and 2. Left: Subset of items involving inconsistencies, right: subset of items with a durative *for*-PP

4 General discussion and conclusions

We started with the observation that a substantial number of items marked with an asterisk by their authors received mediocre ratings from non-linguists in Häussler et al. (2016). We wondered whether semantic factors could have an effect and identified plausibility as a potential

ameliorant. The present study cannot answer this question for the data set in Häussler et al. (2016), but the current results cast an ameliorating effect of plausibility into doubt. For the current data set, we found no ameliorating effect of plausibility and no degrading effect of plausibility either, though the latter might be a floor effect.

Plausibility must be distinguished from intelligibility. Sentences in the current data set were all intelligible, even the ungrammatical ones, though readers might be uncertain how to repair the agreement violation. They could derive a grammatical counterpart by either adjusting the number specification of the subject or by changing the number marking on the verb. The situation is different for other types of syntactic violations. Take for instance island violations as in (7).

(7) Guess who Bill arrived after Mary said that John saw. (Boeckx, 2012: ix)

The intended interpretation of (7) is not immediately obvious. It requires to identify the gap for the fronted *wh*-phrase. Examples like (7) might be therefore unacceptable for two reasons: their grammatical status and the lack of an interpretation (that results from the grammatical status). This might make sentences worse compared to other ungrammatical sentences such as agreement violations or semi-sentences in the sense of Katz (1964). This difference between ungrammatical sentences would produce gradience in acceptability. In the current study, we only included agreement violations, which in principle could be easy to repair, though readers cannot know whether to change the number specification of the verb or of the corresponding subject NP. Psycholinguistic research on garden-path sentences disambiguated by number agreement suggest that readers often do not attempt to solve the apparent agreement violation by reanalysis. We leave it to future research to compare syntactic violations with respect to their effect on intelligibility and acceptability.

While the current study examined only one type of syntactic violation, it included two types of semantic anomalies: inconsistencies and incompatibilities. The results show that the two types differ. *For*-PPs incompatible with verb semantics received a stronger penalty in syntactic ratings but a weaker penalty in semantic ratings. This suggests that participants are not certain about the source of the conflict. This difficulty is not surprising if we consider the (in)compatibility of a verbal predicate and a modifier a syntax-semantics interface issue. This in-between status is also reflected in the way linguists describe corresponding contrasts. The incompatibility with a durative *for*-PP is often marked with an asterisk (e.g., Dowty, 1979; Engelberg, 2000; Vendler, 1957) while other authors use '#' to indicate semantic rather than syntactic anomaly (e.g., Rothstein, 2007). As argued in Sorace (2006), interface phenomena might be particularly prone to gradience in acceptability.

One worry was that participants who are not trained in distinguishing syntax and semantics might struggle to ignore semantic anomalies in an acceptability task. The current results show that participants can differentiate syntactic and semantic aspects of acceptability when instructed accordingly. The current study emphasised to ignore meaning and gave an example for a grammatical but implausible sentence. Furthermore, the current study provided labels at the endpoints of the scale to remind participants constantly about their task. Both methodological details might have helped participants to focus on grammar in Experiment 1.

As pointed out by a reviewer[8], the benefit of a very explicit instruction might fade away in the course of an experiment. The more detailed the instruction, the less we can guarantee that participants comply to it during the entire experiment, especially if that experiment is rather long. In our case, the instruction is not that complex, so we hope that participants were aware of it throughout the entire experiment. Nevertheless, we consider testing potential decline of awareness of the instruction dependent on its complexity and the presence of aids to keep the

[8] We would like to thank the reviewer for highlighting this and the subsequent methodological point.

instruction in mind, e.g., labels, an important issue. At the current point, we leave the details to future research.

Another methodological lesson to take from the present study concerns the construction of filler and control items. Agreement violations have been reliably detected by the participants, which makes them both suitable and not suitable as control items. They are effective to filter out very inattentive and non-cooperative participants – an issue that is particularly virulent in studies recruiting participants online (cf. Downs et al., 2010; Häussler & Juzek, 2016). Participants who fail consistently on such items should clearly be excluded from further analyses. At the same time, agreement violations are less effective to control for minor inattentiveness because of their conspicuity (cf. Häussler & Juzek, 2017). This makes them also less appropriate as control (baseline) condition for overall sensitivity to syntactic violations. However, it is exactly this property which makes them perfect calibration items to establish the lower end of the scale.

While the data show that participants can in principle differentiate form-based acceptability and plausibility, they also show that even with explicit instruction to ignore meaning issues, participants are still susceptible to degrading effects by implausibility. At this point, it would be interesting to see how untrained participants compare to trained linguists. Would linguists be able to distinguish the two tasks even better, as Devitt (2014) would predict, or would there be no significant difference between laypeople who are familiar with the task and experts, as Culbertson & Gross (2009) have observed? We could even ask for differences within the group of professional linguists. Do syntacticians, semanticists and phonologists draw the border at different points? In a recent study, Fanselow et al. (2019) found hints that syntacticians and non-syntacticians differ in their interpretation of *Who wonders who bought what*. This difference could reflect a theory-driven bias or different processing strategies.

To conclude, the present study provides further evidence for how an extra-grammatical factor, implausibility, can have an impact on acceptability. Critically, though, we only found effects in one direction, viz. the degrading of grammatical sentences. We did not observe that plausibility had an ameliorating effect. We therefore think that plausibility can only play a minor role for explaining gradience in acceptability judgements.

Acknowledgments

Many thanks to Tom Wasow and the participants of the Linguistic Evidence 2018 conference for their constructive and helpful feedback.

References

Boeckx, C. (2012). *Syntactic Islands*. Cambridge: Cambridge University Press.

Chomsky, C. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of Mathematical Psychology, Vol. 2* (pp. 269-321). New York City, NY: Wiley.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage.

Culbertson, J., & Gross, S. (2009). Are linguists better subjects? *British Journal for the Philosophy of Science*, *60*(4), 721-736.

Devitt, M. (2014). Linguistic intuitions and cognitive penetrability. *Baltic International Yearbook of Cognition, Logic and Communication*, *9*. https://doi.org/10.4148/1944-3676.1083

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Screening Mechanical Turk workers. In E. Mynatt (Ed.), *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2399-2402). Atlanta, GA: Association for Computing Machinery.

Dowty, D. R. (1979). *Word Meaning and Montague Grammar.* Dordrecht: Kluwer.

Engelberg, S. (2000). *Verben, Ereignisse und das Lexikon*. Tübingen: Niemeyer.

Fanselow, G., & Frisch, S. (2006). Effects of processing difficulty on judgments of acceptability. In G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel (Eds.), *Gradience in Grammars: Generative Perspectives* (pp. 291-316). Oxford: Oxford University Press.

Fanselow, G., Häussler, J., & Weskott, T. (2019). Who cares what who prefers? In K. Carlson, C. Clifton & J. D. Fodor (Eds.), *Grammatical Approaches to Language Processing: Essays in Honor of Lyn Frazier* (pp. 261-274). Cham: Springer.

Featherston, S. (2005). The Decathlon Model of empirical syntax. doi: 10.1515/9783110197549.187. Accessed online: https://www.researchgate.net/publication/255584351_The_Decathlon_Model_of_Empirical_Syntax/download

Frazier, L. (2008). Processing ellipsis: A processing solution to the undergeneration problem? In C. B. Chang & H. J. Haynie (Eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics* (pp. 21-32). Somerville, MA: Cascadilla Proceedings Project.

Häussler, J. (2009). *The Emergence of Attraction Errors during Sentence Comprehension*. Dissertation, University of Konstanz.

Häussler, J., Grant, M., Fanselow, G., & Frazier, L. (2015). Superiority in English and German: Cross-language grammatical differences? *Syntax*, *18*(3), 235-265.

Häussler, J., & Juzek, T. S. (2016). Detecting and discouraging non-cooperative behavior in online experiments using an acceptability judgement task. In H. Christ, D. Klenovsak, L. Sönning & V. Werner (Eds.), *Methods and Linguistic Theories* (pp. 73-99). Bamberg: University of Bamberg Press.

Häussler, J., & Juzek, T. S. (2017). Hot topics surrounding acceptability judgement tasks. In S. Featherston, R. Hörnig, R. Steinberg, B. Umbreit & J. Wallis (Eds.), *Proceedings of Linguistic Evidence 2016: Empirical, Theoretical, and Computational Perspectives*. Tübingen: University of Tübingen. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/77066

Häussler, J., & Juzek, T. S. (ms). Linguistic intuitions and the puzzle of gradience.

Häussler, J., Juzek, T. S., & Wasow, T. (2016). To be grammatical or not to be grammatical – Is that the question? Evidence for gradience. Poster presented at the 'Annual Meeting of the Linguistic Society of America', 7-10 January 2016, Washington, DC.

Hawkins, J. (2006). Gradedness as relative efficiency in the processing of syntax and semantics. In G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel (Eds.), *Gradience in Grammars: Generative Perspectives* (pp. 207-226). Oxford: Oxford University Press.

Hofmeister, P., Jaeger, T., Sag, I., Arnon, I., & Snider, N. (2013). The source ambiguity problem: distinguishing effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes*, *28*, 48-87.

Juzek, T. S. (2016). *Acceptability Judgement Tasks and Grammatical Theory*. PhD thesis, University of Oxford.

Juzek, T. S., & Häussler, J. (ms). Data reliability in syntactic theory and the role of sentence pairs.

Katz, J. (1964). Semi-sentences. In J. Fodor & J. Katz (Eds.), *Structure of Language: Readings in the Philosophy of Language* (pp. 400-416). Englewood Cliffs, NJ: Prentice-Hall.

Keller, F. (2000). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality.* PhD thesis, University of Edinburgh.

Langendoen, D. T., & Bever, T. G. (1973). Can a not unhappy man be called a not sad one? In S. R. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 392-409). New York City, NY: Holt, Rinehart and Winston, Inc.

Manning, C. D. (2003). Probabilistic syntax. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 289-341). Cambridge, MA: MIT Press.

Meng, M., & Bader, M. (2000). Mode of disambiguation and garden-path strength: An investigation of subject-object ambiguities in German. *Language and Speec*h, *43*(1), 43-74.

Myers, J. (2009). Syntactic judgment experiments. *Language and Linguistics Compass*, *3*, 406-423.

Newmeyer, F. (2003). Grammar is grammar and usage is usage. *Language, 79*, 682-707.

Otero, C. (1972). Acceptable ungrammatical sentences in Spanish. *Linguistic Inquiry*, *3*, 233-242.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language, 41*, 427-456.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner (Ed.), *Experiments at the Interfaces* (pp. 147-180). Bingley, UK: Emerald Group Publishing Limited.

Rothstein, S. (2008). Two puzzles for a theory of lexical aspect: The case of semelfactives and degree achievements. In J. Dölling, T. Heyde-Zybatow & M. Schäfer (Eds.), *Event Structures in Linguistic Form and Interpretation* (pp. 175-197). Berlin: Mouton de Gruyter.

Schindler, S., & tbd. (forthcoming) (Eds). *Linguistic Intuitions*. Oxford: Oxford University Press.

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago, IL: University of Chicago Press.

Sorace, A. (2006). Gradience and optionality in mature and developing grammars. In G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel (Eds.), *Gradience in Grammars: Generative Perspectives* (pp. 106-123). Oxford: Oxford University Press.

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

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, *1*, 118-129.

Sprouse, J. (2013). Acceptability judgments. In M. Aronoff (Ed.), *Oxford Bibliographies Online: Linguistics*. Oxford: Oxford University Press. Accessed online: http://www.socsci.uci.edu/~jsprouse/papers/Acceptability.Judgments.OUP.pdf

Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, *66,* 143-160.

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*, 206-237.

Wasow, T. (2009). Gradient data and gradient grammars. In M. Elliott, J. Kirby, O. Sawada, E. Staraki & S. Yoon (Eds.), *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society* (pp. 255-271). Chicago, IL: Chicago Linguistic Society.