Exploring syntactic and semantic acceptability: A case study on semantic restriction violations and aspectual mismatches

Jana Häussler[1]
Bielefeld University

Tom S. Juzek[1]
Kauz Linguistic Technologies and Ruhr University Bochum

Abstract

The present study investigates the interaction between syntax and semantics, and its effects on acceptability. The study compares ratings from two experiments – a syntactic rating task and a semantic one – the latter asking for meaningfulness/plausibility. The focus is on two phenomena: semantic restriction violations and aspectual mismatches with *for*-PPs. For comparison, the experiments also include two reference phenomena: resumptive pronouns, which are ungrammatical but in principle meaningful/plausible, and semantic contradictions, which are not meaningful/plausible but grammatical. Further, we include anchor items of various degrees of grammaticality and meaningfulness/plausibility, in order to calibrate the scale and probe the rating space. The results for the resumptive pronouns and the semantic contradictions, as well as the anchor items, indicate that our participants struggled to distinguish between the two tasks to some degree. Semantic deviations seem to drag down syntactic acceptability, and syntactic anomalies drag down perceived meaningfulness/plausibility. Importantly, however, the results remain interpretable. We observe that the impact of semantic anomaly on syntactic acceptability differs across phenomena, as did the impact of syntactic deviations on semantic acceptability. Furthermore, the semantic restriction violations seem to affect semantic acceptability more than syntactic acceptability. By contrast, the *for*-PPs received reduced ratings in both tasks. Our findings further substantiate the notion that the border between syntax and semantics is not clear-cut and that the interface between the two is complex.

Keywords: experimental syntax, semantics, grammaticality, acceptability, plausibility

[1] Corresponding authors. Both authors contributed equally.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Finnish Journal of Linguistics 34 (2021), 39–76

1 Introduction

The distinction between grammaticality and acceptability is widely accepted in syntax. Grammaticality is a pure reflection of the competence grammar, abstracting away from performance factors such as memory limitations and the like. Acceptability, by contrast, is the joint product of the competence grammar and various performance factors (cf. e.g., Schütze 1996).

The examples in (1), modelled after Gibson & Thomas (1999), illustrate this distinction, where an asterisk indicates ungrammaticality and the caret degraded acceptability. Example (1a) is a sequence that is both grammatical and acceptable. Structures such as the one illustrated in (1b) are ungrammatical and unacceptable, as either the main or relative clause is missing a verb phrase. (1c) is grammatical but degraded in acceptability, as signalled by the caret. Items with multiple embeddings, such as (1c), typically receive reduced acceptability ratings due to their complexity. This combination of grammatical but not (fully) acceptable is in contrast to acceptable ungrammatical structures as in (1d). Like (1b), the example in (1d) involves multiple embeddings, but unlike (1c), it is ungrammatical, as it lacks a verb phrase. Despite this violation, incomplete double embedded relative clauses such as (1d) are perceived as more acceptable and easier to comprehend than (1c) (Frazier 1985; Gibson & Thomas 1999; Christiansen & MacDonald 2009; Gimenes et al. 2009; Frank & Ernst 2019), especially when the final verb is compatible with both the first and the second NP in the string (Christiansen & MacDonald 2009; Frank & Ernst 2019). Similarly, reading times at the final verb are faster than in ungrammatical counterparts; that is, reading times for sentences such as (1d) lack the slowdown typical for grammatical violations (Vasishth et al. 2010; Frank et al. 2016). Ratings and reading times suggest that readers occasionally fail to notice the missing verb. Examples such as (1d) are instances of the "missing-VP effect" (Gibson & Thomas 1999) or more generally of a so-called grammaticality illusion (Phillips et al. 2011).

(1) a. The patient the nurse admitted met Jack.
b. *^ The patient the nurse met Jack.
c. ^ The patient the nurse the clinic had hired admitted met Jack.
d. * The patient the nurse the clinic had hired met Jack.

Note that our diacritic for (1d) is not meant to indicate that (1d) is fully acceptable. Instead, they are meant to indicate a contrast between (1c) and (1d) in perceived acceptability. Alternatives such as adding a caret qualified by a question mark, or using carets of different sizes, would probably increase ambiguity and inconsistencies in the use of diacritics rather than clarify the status of items such as (1d) (for a detailed discussion of diacritics, see Bard et al. 1996; Schütze 1996; Cowart 1997).

Various extra-grammatical factors contribute to (un)acceptability. They broadly fall into two categories: factors that are regarded as undesired noise and factors that are of interest to linguists. An example of undesired noise is experimental artefacts like scale effects, which "are undesirable differences due to the choice of scale" (Häussler & Juzek 2021: 106). For example, when the experimental items are imbalanced, this can lead to anchoring effects, leading to "a tendency towards higher ratings if one includes a lot of degraded items, and vice versa" (Häussler & Juzek 2021: 108). As to extra-grammatical factors that are of interest, those include cognitive restrictions like memory limitations (Chomsky & Miller 1963; Gibson & Thomas 1999; Keller 2000), semantic influences like real-world implausibility (Sprouse 2008), interpretation (Etxeberria et al. 2018), and ambiguity (e.g. Myers 2009: 409). Crucially, the various grammatical and extra-grammatical factors are hard to disentangle. Previous research has focused on the impact of grammatical versus cognitive factors, whereas semantic factors have received less attention so far. The present study examines the question of how semantic anomalies affect perceived syntactic well-formedness, as well as how far syntactic degradedness affects perceived meaningfulness/plausibility. Insights into these questions might also advance our understanding of further factors that affect syntactic acceptability.

Our research question relies on the assumption of a strict separation of syntax and semantics, for example, in the sense of Chomsky (1957). As we will see, the results of the present study pose a challenge to that assumption. Consequently, we will return to this question in § 4.4, where we also point to alternative frameworks that do not posit a strict separation of syntax and semantics.

As test cases for our research question, we chose four phenomena: contradictions as an instance of semantic anomalies, resumptive pronouns in relative clauses as an instance of syntactic violations, and aspectual mismatches and semantic restriction violations as exemplifications of interface phenomena. Although our results provide some insights into the

examined phenomena themselves, the concrete phenomena are secondary to our goals. The experiments did not test specific hypotheses regarding the four phenomena – and in that, the present study is to some degree explorative. We cannot do justice to the phenomena because each is so rich that an in-depth discussion goes beyond the scope of this paper.

In the following section, we introduce the phenomena used in the study, and we justify their relevance. § 3 presents two experiments in which we collected semantic and syntactic ratings for the same set of stimuli. The results are discussed in § 4, with special emphasis on contrasting syntactic versus semantic ratings. § 5 concludes the paper.

2 Phenomena under investigation

The experiments compared four anomalies that differ with respect to the locus of the violation (syntax and/or semantics). At the centre of the study are semantic restriction violations and aspectual mismatches with *for*-PPs. They have in common the fact that they manifest themselves as incompatibilities in the lexical semantics of the verb and some other sentence constituent. The other two phenomena serve as reference points as their degradedness stems from only semantic factors (semantic contradictions) or from only syntactic factors (resumptive pronouns).

2.1 Semantic restriction violations

Verbs and other predicates impose requirements on their arguments (e.g. Chomsky 1965). Such requirements concern syntactic properties, such as the syntactic category of an argument (c-selection), as well as semantic properties of the arguments (s-selection). It has been a matter of debate whether c-selection and s-selection form autonomous subsystems of the grammar (e.g. Grimshaw 1979; Pollard & Sag 1987) or can be reduced to a single component (e.g. Pesetsky 1982). Proponents of the latter view typically argue that c-selection can be derived from s-selection, for instance via canonical structural representations: propositions are realised as CPs or NPs, agents as NPs/DPs, and so on.

Semantic restriction violations (SRVs) arise when s-selectional requirements conflict with the argument's semantic features. For example, the verb *drink* requires a liquid entity as its internal argument; hence, *water* is fine as the internal argument, while *salt* is not (*John drinks water*/$^{\#}$*John drinks*

salt). Psychological predicates (*love*, *enjoy*, *comfort*, *anger*, etc.) require an animate, perhaps even human, referent for the experiencer argument.

(2) a. # Jill's dog comforted her old truck.
b. Jill's dog comforted her little girl.
c. Jill's dog monitored her old truck.
d. Jill's dog monitored her little girl.

The sentence in (2a) is odd because *comfort* is an object-experiencer verb and requires the internal argument (the experiencer) to refer to an animate entity. The oddity disappears once this animacy requirement is satisfied, for example, as per (2b). Such requirements are verb (class) specific, as demonstrated in (2c). A verb such as *monitor* imposes no animacy requirement on the internal argument; therefore, (2c) and (2d) are equally good.

It should be noted that the term *s-selection* has been used for a variety of (interrelated) semantic properties: lexical semantic features (e.g. animate, human, liquid), thematic roles (agent, patient, experiencer, etc.), and semantic types (e.g. proposition, interrogative, exclamative). In our study, we concentrate on animacy-related violations, as in (2a). Semantic restriction violations in this sense are very similar to implausibilities, raising the question of whether it is world knowledge rather than lexical knowledge that triggers the perception of anomaly. Psycho- and neurolinguistic studies have shown that implausibilities yield different effects compared to semantic restriction violations (e.g. Hagoort et al. 2004; Warren & McConnell 2007; Pylkkänen et al. 2009; Warren et al. 2015).

We chose semantic restriction violations because they clearly instantiate a semantic anomaly, but they are also bound to entities that are syntactically related. Semantic restriction violations share with two of the other phenomena under investigation that they involve conflicts between syntactic elements. In the case of semantic restrictions, combinatorial restrictions concern a syntactic head and its dependents. This sets them apart from both the aspectual mismatches with *for*-PPs, which concern the compatibility of verbs and adverbial modifiers, and contradictions, whose anomalous status is entirely derived from the contradictory meaning of their lexical parts (as in *married to a bachelor*; see below for further details).

The semantic nature of semantic restriction violations should result in low semantic ratings and might also result in lower syntactic ratings when participants consider the semantic violation a syntactic one.

2.2 Aspectual mismatches with *for*-PPs

The second phenomenon that we focus on is sequences in which anomalies occur with certain combinations of verbs and prepositional phrases with *for*, hence the shorthand *for*-PPs. The sequences in (3) illustrate this phenomenon.

(3) a. # The front window broke for at least three minutes.
b. The front window broke almost instantaneously.
c. The front window vibrated for at least three minutes.
d. The front window vibrated almost instantaneously.
e. The front window vibrated for a very short time.
f. #? The front window broke for a very short time.
g. The front window broke over the course of at least three minutes.

The degradation in (3a) comes from a conflict between the event structure of the verb (*broke*) and the temporal modifier (*for at least three minutes*). In the cases that the present paper focuses on, the verb implies a sudden change of state, while the *for*-PP especially expresses a longer duration. The nature of this clash is exemplified by the contrast between (3a) and (3b). In (3b), the modifier is compatible with the verb's event structure. However, for verbs that allow for durational readings, durational temporal modifiers do not cause any degradation, as illustrated by (3c). Note that while the modifier in (3b) can specify the duration of the event denoted by the verb, the same modifier has a different function in (3d). In (3d), the more prominent reading is that something has caused the window to vibrate and that there is very little time between cause and effect. However, a slightly different modifier can be used to specify the duration of the verb, as per (3e), which makes use of a slightly modified *for*-PP. Applying this *for*-PP to *broke*, as in (3f), results in a sentence of a questionable status, and a modifier with *in* would be preferable. We will return to (3g) below.

The phenomenon first received considerable attention in the literature on aspectuality and verb semantics (e.g. Vendler 1957; Dowty 1979). In an early paper on the phenomenon, Vendler (1957) used *for*-PPs as a means to distinguish between achievements (*win a race*) and accomplishments (*run a mile*) on the one hand, and states (*love somebody*) and activities (*push a cart*) on the other hand. While states and activities are compatible with *for*-PPs, achievements and accomplishments are not. This is because states and activities can last for a longer duration, while achievements and accomplishments are typically very restricted in their duration, as they are instantaneous changes of states.

The difference between achievements and accomplishments versus states and activities also reflects a difference in telicity. Achievements and accomplishments are typically telic, while states and activities are typically atelic. This was noted early on, so that since the early work by Vendler (1957) and Dowty (1979), the use of *for*-PPs has been established as a common test to check for telicity.

Aspectual mismatches with *for*-PP come in degrees. Some verbs are strictly instantaneous (*the tyre burst*), while others allow for the event to span some duration (*the volcano erupted*). Further, sometimes repairs are possible. For example, (3a) could be made more plausible by further modifying *broke*, for example, along the lines of *slowly broke* or *broke bit by bit*. Iterative readings can also be used as a repair strategy (*the balloons popped for three minutes*) or as in (3g), and sometimes the sequence can be repaired by linking the *for*-PP to a resultative state (instead of linking it to the event itself, as in *the lake froze for three days*). For an introduction to event structures, including some discussion on *for*-PPs, see Pustejovsky (1991).

Aspectual mismatches also figure prominently in psycholinguistic and neurolinguistic theories of compositionality in semantic processing. A phenomenon that is of special interest in this context is aspectual coercion, that is, resolving the aspectual mismatch by reinterpretation (for an overview, see Brennan & Pylkkänen 2008). Such adjustments can, for instance, result in iterative readings (e. g. *the patient sneezed for ten minutes* interpreted as a series of sneezing events). The present study, however, does not primarily aim to contribute to the discussion of aspectual coercion and how to best capture the underlying mechanism. Instead, the study pursued a more general question, that is, how the aspectual mismatch affects syntactic versus semantic acceptability. We therefore explicitly aimed to construct items that are not easily repaired by aspectual coercion (for details, see § 3.1.2).

The relevance of aspectual mismatches for the present study comes from the fact that the phenomenon has both a semantic component, the event structure, and a syntactic one, namely, the syntactic encoding of event structure and the syntactic integration of the *for*-PP. That is, it is hard to pinpoint the exact source of the perceived degradation: Does it come from grammar constraints or from semantic considerations? As a consequence, our expectation is that a sentence such as (3a) could receive reduced ratings both in a semantic judgement task and in a syntactic one.

2.3 Two reference phenomena: semantic contradictions and resumptive pronouns

For comparison, we chose two reference phenomena for which the source of the deviation is clear: contradictions, which are purely semantic violations not involving any syntactic aspect, and resumptive pronouns, which are syntactically deviant (in Standard English).

The contradictions are constructed such that they include some semantic clash arising from the meaning of its lexical parts (as in *married to a bachelor*). The expectation is that the semantic contradictions should receive high ratings in the syntactic task and low ratings in the semantic task.

Cross-linguistically, resumption in relative clauses is a widespread phenomenon (McCloskey 2006; Asudeh 2012). In Standard English, however, resumption is only marginally acceptable, though acceptability increases with the depth of embedding (cf. Erteschik-Shir 1992; Dickey 1996). Many speakers of Standard English consider instances such as (4a) unacceptable, and Erteschik-Shir (1992) marks (4a) as ungrammatical.

(4) a. *^ This is the girl that John likes her.
b. This is the girl that John likes *t*.

Given that resumptive pronouns are acceptable in some languages and commonly produced in spoken informal registers of English (e.g. Cann et al. 2004; Radford 2019), and given that the inclusion of a resumptive pronoun results in a locally coherent string (*John likes her*), inferring a meaning for (4a) should not be too difficult. In fact, it has been argued that resumptive pronouns, though ungrammatical, may facilitate the processing of relative clauses. The evidence for this claim is somewhat mixed, with most studies failing to find such an effect (e.g. McDaniel & Cowart 1999; Heestand et al. 2011; Keffala & Goodall 2011; Keffala 2013; Polinsky et al. 2013). If at all, facilitating effects show up when the resumptive pronoun occurs inside an island and when processing load is high due to deep embedding (Hofmeister & Norcliffe 2013; Beltrama & Xiang 2016). In our materials, the resumptive pronouns did not occur in an island and there was only one level of embedding. We therefore expected no facilitating effect but no aggravating effect either. Given the syntactic violation, we expected reduced syntactic ratings. For the semantic ratings, we expected no effect or only a weak effect, possibly reflecting some additional effort.

3 Experiments

We ran two experiments: a semantic judgement task and a syntactic judgement task. In the semantic judgement task, participants were asked to judge the stimuli with respect to how meaningful/plausible the stimuli were. In the syntactic judgement task this was done with respect to how natural/grammatical they were. The two experiments tested instantiations of the four phenomena discussed in § 2. The resulting ratings are our dependent variable, with the four phenomena and the factor +/− Violation as independent variables (see § 3.2 for details).

3.1 Methodology

3.1.1 Participants and exclusions

For each experiment, we recruited 80 participants (40 per list) through Prolific,[2] who were then redirected to our own website for the actual experiment. Only self-declared native speakers of English with British nationality were allowed to participate. We assumed that most, if not all, of them are native speakers of British English. Regarding their gender, 111 participants chose "female", 45 chose "male" and 3 selected "other/do not want to say". Their ages ranged from 18 to 70, with a median of 31 (mean age 32.6). For unknown reasons, demographic data were missing for one participant.

The participants who completed the questionnaire were approved to Prolific and were paid for their participation with an hourly rate of effectively £13. For technical reasons, which we do not fully understand, two participants had missing data. We approved and paid them but excluded their data from further analyses. We also excluded data from participants who failed on control items ($N = 1$), and/or had extreme response times (extremely slow or extremely fast, $N = 3$). Unusual ratings for control items and unusual response times are a strong indicator of non-cooperative behaviour (cf. Häussler & Juzek 2016).

As to the control items, we included four items for which we had clear expectations, based on a pilot experiment, a strategy comparable to comprehension questions, as advocated by Gibson et al. (2011). If a participant deviated considerably, we excluded that participant. For details

[2] https://www.prolific.co/ (accessed 2019-08-01).

on the use of our "gotcha"/control items, we also followed Häussler & Juzek (2016). The exact control items can be found in the complete list of items (see below for details).

Unusual response times were determined as follows. We averaged the means of all participants per sub-experiment. Any participant whose median response time was lower than 1.5 standard deviations below the grand average was regarded as unusually fast, while 4 standard deviations above the grand mean was regarded as unusually slow. In total, we excluded 3 participants, giving us 79 participants in the semantic task and 78 participants in the syntactic task.

3.1.2 Materials

The stimuli, which were identical in the two experiments, comprised the two phenomena introduced in § 2 – the semantic restriction violations and aspectual violations with *for*-PPs – and two reference phenomena: the contradictions and resumptive pronouns. Each item appeared in two versions – one involving the respective violation and a good counterpart.

For the semantic restriction violations, the good counterparts are derived by exchanging the verb, as in (5a) and (5b) adapted from Warren et al. (2015).

(5) a. # My dog reassured the old waterbottle and went to sleep.
b. My dog buried the old waterbottle and went to sleep.

The violation condition in (5a) contains an object-experiencer verb, while the baseline condition (5b) contains an agentive verb. The argument NPs are held constant: The subject NP is always [+animate], the object NP is always [−animate], resulting in a semantic restriction violation with object-experiencer verbs in the violation condition and no such violation in the baseline condition. In total, we adapted eight item pairs from Warren et al. (2015).

The items involving *for*-PPs are loosely based on items that we have encountered in various sources but mainly Pustejovsky (1991). They always consist of a noun followed by a punctual verb, followed by an optional object, followed by a *for*-PP that specifies a longer duration. In the literature, verbs modified by a durational *for*-PP are typically in the simple past (e.g. *The window vibrated*/# *broke for three minutes*), whereas in our experiments, the verb occurred in the past progressive. The rationale for this modification was to strengthen the aspectual conflict. In (6a), both the grammatical form of the

verb and the durational modifier conflict with the verb's lexical semantics. The grammatical form (the past progressive) and the modifier (durational *for*-PP) dictate a process reading, whereas verb semantics dictates a punctual reading. The baseline condition in (6b) involves the same grammatical form (past progressive) and the same durational modifier but an atelic verb and hence lacks aspectual conflict. Other examples include *Will's toy plane was blowing up* / *was descending for 20 minutes* and *Helen's phone was shattering into pieces* / *was charging for one hour*. For further discussion of the progressive, see § 4.3.

(6) a. # The front window was breaking for at least three minutes.
b. The front window was vibrating for at least three minutes.

Aspectual conflicts can be solved by coercion, that is, reinterpretation as a result of shifting operations. Rather than modifying the punctual event, the *for*-PP can be taken to modify the time span after the event (as in *The door was closed for three hours*), the time span during which the event was iterated (*The cat jumped for the door handle for about ten minutes … until it finally succeeded*), or the time span during which the event was repeated as a habit. When constructing our stimuli, we tried to hamper such aspectual coercion: The subject of the punctual verb was in most cases realised by an NP in the singular to hamper an interpretation in which the time-frame adverbial specifies the duration of a series of punctual events as in *Balloons were popping for twenty minutes* (one after the other). Further, the time specification was chosen such that habitual readings, possibly a special case of iterative readings (as in *The organizers invited Albert Einstein for several years*), are unlikely as well. The *for*-PPs in the experiment specified time frames ranging from three minutes to two hours. Habits, by contrast, are a matter of days, if not years. Because iterative readings are easily available for punctual verbs with reversible or unstable resultant states, for example, *knock*, *cough*, *blink*, etc., our experimental stimuli contain mainly verbs with irreversible resultant states (*break*, *pop*, *blow up*, *snap*, etc.). There was one notable exception: *The two removers were dropping the safe for ten minutes*. In this case, world knowledge makes an iterative reading unlikely, although it does not rule out such a scenario, for example, in a cartoon. Irreversibility also hinders readings in which the temporal adverbial specifies how long the resultant state persists. In *Max fell asleep for almost 2 hours*, the *for*-PP does not modify the duration of the transition but the duration of the subsequent sleep. Such a reading is not

available for (6a) or the other items in our materials.

In addition to semantic restriction violations and aspectual violations with a *for*-PP, we included two reference phenomena: semantic contradictions and resumptive pronouns. An example of a contradiction is given in (7). (7a) is the critical item, and (7b) is the good baseline.

(7) a. #My sister Jane is married to a bachelor.
b. My sister Jane is married to a lawyer.

The semantic contradictions are partly inspired by the discussion in Horn (2018). They represent a variety of conflict types. The example in (7a), for instance, involves a conflict between lexical meanings – being *married* is incompatible with being a *bachelor*. Other examples involve world knowledge in addition to lexical knowledge, as in *Having won the 100 m final of the 2017 world championship, Justin Gatlin was awarded the bronze medal*, which requires world knowledge about the reward for the winner of a championship, or logical inferences, as in *My new Volkswagen is emitting a lot of carbon dioxide, but it is not emitting any* CO_2, which requires understanding the negation and knowledge about the formula for carbon dioxide. This heterogeneity would be problematic for experiments directly addressing the acceptability or processing of contradictions, as it adds variance and mixes constructions. Contradictions as such are, however, not the main concern of the present experiment. Instead, our focus is on comparing form-based (syntactic) and meaning-based (semantic) acceptability. We therefore think that testing a less homogeneous set of items than usual is legitimate, if not beneficial. We return to this issue of heterogeneity in § 4. Note at this point that all critical items in this subset share the fact that they are syntactically well-formed but semantically clearly deviant.

The items including resumptive pronouns are modelled after examples in the literature (Prince 1990; Erteschik-Shir 1992; McKee & McDaniel 2001; Cann et al. 2004: 1554, through Keffala 2013; Ferreira & Swets 2005; Hofmeister & Norcliffe 2013; Keffala 2013; and Radford 2018: 96). We constructed eight item pairs with a resumptive pronoun in the violation condition. In one half of the item pairs, the good counterpart contains a relative clause with a gap, as in (8b). In the other half, the relative clause is replaced by a coordinated main clause, as in (9b). Again, such heterogeneity would be objectionable in an experiment on resumptives but should be acceptable for the purpose of the present study (for further discussion, see § 4.1).

(8) a. *^ Mary said that this is the girl that John likes her.
b. Mary said that this is the girl that John likes *t*.

(9) a. *^ This is a donkey that I don't know where it lives.
b. This is a donkey but I don't know where it lives.

In total, we have 32 item pairs, 8 per phenomenon and each involving a marked version and a good counterpart. These 64 sentences were distributed over two counterbalanced lists, so that per item, each participant only saw either the marked version or the good counterpart. We also included 32 filler/anchor items, so that the rate of fillers per list is 50%. Those items were based on the anchor items in Gerbrich et al. (2019). Gerbrich and colleagues designed their item set as a yardstick for comparisons across experiments that all use this filler set in addition to their critical items. The items were carefully selected from a larger set such that they represent a range of acceptability (on a five-point scale) and exhibit high interrater agreement. The set contains mainly declarative sentences but also a few exclamatives and questions, probably to increase variation (of surface form and violation type) and to allow for use in experiments testing exclamatives or questions. We further modified the items to include semantic deviance, such that they are marked, somewhat marked, and unmarked with respect to grammar and meaning, including all possible combinations. Items in Gerbrich et al. (2019) have five degrees of syntactic acceptability (for which we use "^", "^???", "^??", "^?", and "OK"), to which we added three degrees of semantic acceptability/plausibility ("#", "#?", "OK"). We left it at three degrees of semantic acceptability, as we found it hard to reliably introduce further degrees. Examples are given in (10–12), with the first diacritic indicating the syntactic status and the second indicating the semantic status. Note that in the ungrammatical sentences, the semantic status was evaluated ignoring/repairing the syntactic deviation.

(10) a. ok/ok The winter is very harsh in the North.
b. ok/#? There's a statue in the middle of the ocean.
c. ok/# The patient fooled the chair by pretending to be in pain.

(11) a. ^??/ok Hannah hates but Linda loves eating popcorn in the cinema.
b. ^??/#? Most people like very much a cup of sparkling wine in the morning.
c. ^??/# What my hamster wants to know is which student which exam failed.

(12) a. ^/ok Historians wondering what cause is disappear civilization.
b. ^/#? Old man he work garden grow many cat food and breads.
c. ^/# Backers must continue much planets for they become hairy.

Further, the first four items of any questionnaire were calibration items, which were added to give the subjects an idea of the endpoints of the scale. Two of the calibration items are semantically and syntactically unmarked, and the other two are semantically and syntactically marked.[3]

Two final remarks on the items: First, compared to many other experiments, our items exhibit some variation. This variation is on purpose, and the items vary more than one would normally allow in an experiment. For example, in an experiment on resumptive pronouns one would control for syntactic function of the relative pronoun and properties of the clause containing the gap. The reason for this is that our main comparison is not between conditions (+/−deviant) but between constructions (representing different types of anomalies – syntactic vs. semantic) and between experimental tasks (form-based, that is, syntactic, judgement vs. semantic judgement), since we were primarily interested in how well participants differentiate between syntactic and semantic deviations. We therefore aimed for a variety of violations and a range of rating options.

Second, following an anonymous reviewer's suggestion, we checked our materials for gender imbalances and stereotypes. In contrast to the strong imbalance reported in Kotek et al. (2021) for examples in linguistic papers, we found an even distribution of female and male referents in our items (14 female, 14 male, 1 conjoined female + female, 1 conjoined female + male). There was also no gender difference in the likelihood of occurring in subject position (9 out of 14 females, 8 out of 14 males). However, the absolute number of arguments referring to humans specified for gender ($N = 30$) is too low for a substantial quantitative analysis.

3.1.3 Procedure

We ran two separate experiments, one asking for semantic ratings and the other collecting syntactic ratings. The two experiments included exactly the same materials, distributed over the same two lists described above, but with different participants, to avoid revealing the purpose of our study. In each

[3] A list of all items, including filler items, can be found at https://zenodo.org/record/5546040 (published 2021-07-04).

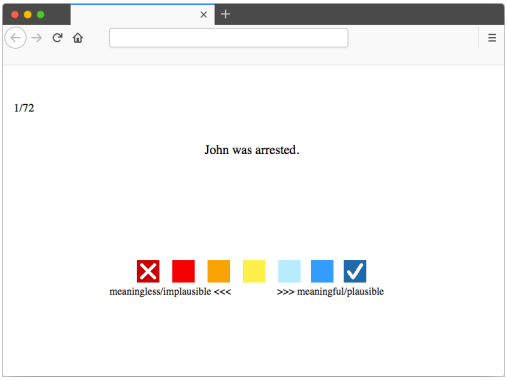

Figure 1. The rating interface used in the semantic judgement task

experiment, the participants rated the stimuli one at a time using a 7-point scale on a dedicated website for the experiments.[4] As an alternative for building an experimental website, we recommend "jsPsych",[5] which is also JavaScript-based and provides a library of building blocks as well as an active community willing to help at any time.

In the semantic judgement task, participants were asked to judge how "meaningful/plausible" sentences appeared to them. We instructed participants to concentrate on meaning and ignore grammaticality or spelling. To illustrate this point, we added the following examples. *Jack did his job goodly* was given as an example of a meaningful and intelligible, but also not fully grammatical sentence. *The storm intentionally broke the window* was introduced as fully grammatical but implausible. The scale and rating interface were introduced through an example. A red button with a cross-mark represented a low rating, and a blue button with a check-mark represented a high rating, with other buttons in between denoting in-between ratings. The endpoints of the scale were labelled with "meaningless/implausible" and "meaningful/plausible", to ensure that participants were always aware of the meaning of the buttons. The rating interface is illustrated in Figure 1.

[4] The code for the web interface is available at Zenodo: https://zenodo.org/record/5546040 (published 2021-07-04).
[5] https://www.jspsych.org/ (accessed 2021-12-25).

In the syntactic judgement task, participants were asked to judge how "natural" or "unnatural" the items were with respect to the items' grammaticality. The instructions explicitly asked participants to not be bothered with meaning or spelling. We also provided the same examples mentioned above. The scale and the interface were introduced in a similar fashion, however with labels adjusted to "unnatural/ungrammatical" and "natural/grammatical".

Both experiments tracked participants' response times and applied a warning mechanism to discourage fast clicking through the experiment. A threshold was defined per stimulus: (225 ms + 25 ms per character) divided by 2. Whenever a participant's response time went below that threshold, a warning message was displayed (this was motivated by the discussion in Häussler & Juzek 2016).

3.2 Analyses

Since we are interested in the impact syntactic anomalies have on semantic acceptability, semantic anomalies, and syntactic acceptability, our focus is on the comparison of the ratings in the two tasks. For the first indication of how well participants distinguished between the two tasks, we explored how much of the two-dimensional rating space (syntactic by semantic ratings) was used. Our main analysis is a point-biserial correlation measure in which we assessed the degree to which semantic and syntactic mean ratings per item correlate. As a baseline, we first correlated the expected semantic ratings with the expected syntactic ratings. As to the expected ratings, for each item, including the anchor items, we defined expectations between 0 and 1. We did this twice, once for each task. The expectations were based on introspective judgements from the literature and/or were defined by us. For instance, item (10a), categorised as OK/OK, had an expectation of 1 for the syntactic task and another 1 for the semantic task. Item (10c), categorised as OK/#, had expectations of 1 and 0, respectively. In line with the diacritics in (5–9), we categorised our two critical constructions (semantic restriction violations and aspectual mismatches with *for*-PPs) as well as the contradictions as syntactically 1 and semantically 0. Resumptive pronouns were categorised as syntactically 0 and semantically 1. For all constructions, the good counterparts were categorised as 1 and 1. We then correlated semantic expectations and syntactic expectations and obtained a correlation coefficient of -0.04. This represents our baseline, to which we compared the correlation measure based on the experimental ratings.

Based on item means, we correlated the observed semantic ratings with the observed syntactic ratings. In a last step, we compared the outcomes of the two correlation measures (expected vs. observed).

In addition to the comparison analysis, we ran separate analyses for each rating task using linear mixed effects models. In those models, ratings are our dependent variable, with the phenomena in interaction with Violation (violation vs. baseline) as fixed effects, and item pairs and participants as random factors (random intercept). Here, *item pair* is used to denote what is sometimes called "lexicalisation" or even "item". For instance, (7a) and (7b) were an item pair in our experiment.

3.3 Experimental results

In what follows, we first report separate analyses for the two rating tasks and then our main analysis, in which we compare the syntactic versus semantic ratings.[6] We used R for analysing the data and creating graphs (R Core Team 2021). For mixed effects models, we used the R-packages "lmerTest" (Kuznetsova et al. 2017) and "MuMIn" (Bartoń 2020).

3.3.1 Results of the semantic rating task

The average ratings for the four phenomena and their baselines are illustrated in Figure 2. Semantic ratings are given in dark grey (two leftmost bars in each plot). As expected, contradictions received a strong penalty in the semantic ratings. However, the other three phenomena exhibit penalties as well, although of different sizes.

3.3.2 Results of the syntactic rating task

The light grey bars in Figure 2 illustrate the syntactic ratings. They show the expected penalty for resumptive pronouns in relative clauses (the second plot in Figure 2). Substantial penalties in the syntactic ratings are also visible for semantic restriction violations (third plot in Figure 2) and for the aspectual mismatch in sentences with a punctual verb modified by a durational *for*-PP (rightmost plot in Figure 2). Contradictions, by contrast, show no substantial penalty in syntactic ratings. A linear mixed effects model with

[6] All results, including individual ratings, can be found on Zenodo (https://zenodo.org/record/5546040, published 2021-07-04). Our complete R-code is also available on Zenodo.

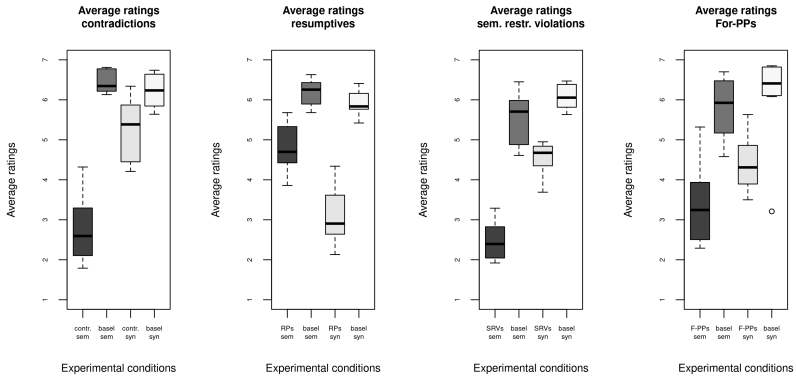

Figure 2. The median acceptability ratings, y-axis, for the four phenomena and their good baselines ("basel"); darkest grey (first bar from the left per plot) = the semantic ratings for a given phenomenon; dark grey (second bar) = the semantic ratings for the good baselines; light grey (third bar) = the syntactic ratings for a phenomenon; lightest grey (fourth bar) = the syntactic ratings for the good baselines

participants and item pairs as random factors (random slopes) confirms this visual impression (cf. Tables 1 and 2). The intercept is the baseline of the contradictions. In both experiments, these had the highest estimates. The subsequent three rows are the baselines for the other constructions in relation to the baseline of contradictions. *Violation* denotes the penalty for the critical items of the contradictions in relation to the intercept. The subsequent three rows are in relation to Violation. In the semantic experiment, significant differences were expected for Violation and SRV $\times$ Violation. In the syntactic experiment, these two effects should not be significant.

Our models include Violation and Construction Type as fixed effects, and the question arose as to whether the latter factor was necessary at all. In what follows, we compare our models to simpler models that do not include Construction Type as a factor.[7]

Construction Type did affect the ratings. For the syntactic ratings, as per ANOVA comparing the two models, a simple model without Construction Type as a factor was significantly different from a model including that factor ($\chi^2 = 20.34$, $p < 0.001$). The conditional R-squared for the simple

[7] Many thanks to one of the reviewers for this suggestion.

Table 1. A summary of the general mixed-effects model for the semantic ratings, including coefficient estimates, standard errors, and the *t*-values. The asterisk marks significance with a level of < 0.05.

Fixed effects	Estimate Std.	Error	df	*t*-value	*p*
(Intercept) [baseline ctrd.]	6.45	0.18	48.12	35.64	$< 0.001*$
For-PPs [baseline]	−0.65	0.25	41.01	−2.66	$0.011*$
Resumptives [baseline]	−0.26	0.25	41.01	−1.08	0.29
SRV [baseline]	−0.93	0.25	41.01	−3.79	$< 0.001*$
Violation [ctrd.]	−3.69	0.14	2327.12	−25.54	$< 0.001*$
For-PPs × Violation	1.26	0.20	2327.12	6.17	$< 0.001*$
Resumptives × Violation	2.31	0.20	2327.12	11.32	$< 0.001*$
SRV × Violation	0.63	0.20	2327.12	3.09	0.002*

Random effects	Variance	Std. Dev
Participant	0.32	0.56
Item pair	0.16	0.40
Residual	2.91	1.71

Table 2. A summary of the general mixed-effects model for the syntactic ratings, including coefficient estimates, standard errors, and the *t*-values. The asterisk marks significance with a level of < 0.05.

Fixed effects	Estimate Std.	Error	df	*t*-value	*p*
(Intercept) [baseline ctrd.]	6.23	0.22	41.97	28.36	$< 0.001*$
For-PPs [baseline]	−0.14	0.30	34.65	−0.47	0.64
Resumptives [baseline]	−0.31	0.30	34.65	−1.06	0.30
SRV [baseline]	−0.15	0.30	34.65	−0.52	0.61
Violation [ctrd.]	−0.98	0.13	2352.09	−7.38	$< 0.001*$
For-PPs × Violation	−0.70	0.19	2352.09	−3.73	$< 0.001*$
Resumptives × Violation	−1.84	0.19	2352.09	−9.77	$< 0.001*$
SRV × Violation	−0.55	0.19	2352.09	−2.91	$0.004*$

Random effects	Variance	Std. Dev
Participant	0.35	0.59
Item pair	0.28	0.53
Residual	2.72	1.65

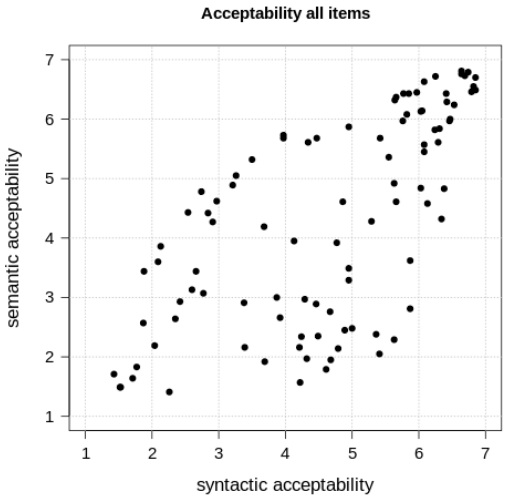

Figure 3. The average ratings for all items, including fillers items. Average syntactic acceptability is given on the *x*-axis, and average semantic acceptability is given on the *y*-axis.

versus complex model was 0.407 versus 0.410, respectively, calculated with the "MuMIn" package (Bartoń 2020). A comparison of the semantic models yielded a similar outcome ($\chi^2 = 25.45$, $p < 0.001$). The conditional R-squared was 0.517 for the simple model and 0.519 for the complex model. In summary, while the complex models were significantly different from the simpler models, they were only minimally more capable.

3.3.3 Comparison of the two rating tasks

For a better idea of how well the participants were able to distinguish the two tasks, we plotted the rating space of both the critical items and the fillers, with syntactic acceptability on the *x*-axis and semantic acceptability on the *y*-axis. The lowest average ratings in the semantic and syntactic tasks are approximately 1.5, and the highest ones are close to 7. The resulting graph in Figure 3 indicates that participants used almost the entire rating space; that is, differentiated degrees of acceptability in both rating tasks. The distribution in

Figure 3 suggests a moderate and not zero correlation. Notably, the second quadrant (top left part of the plot) is rather empty. Items that received low syntactic ratings hardly received high semantic ratings. The fourth quadrant (bottom right), by contrast, is populated to some extent: Items that received low semantic ratings sometimes received higher syntactic ratings.

As mentioned above, the distribution visible in Figure 3 suggests a non-zero correlation. In fact, the coefficient of the point-biserial correlation measure for the experimental ratings – that is, semantic ratings versus syntactic ones, is 0.78. This is considerably different from our expectation of a near-zero correlation (the −0.04 from above).

4 Discussion

When constructing our items, there was a tension. On the one hand, there must be some lexical variation to prevent a poor lexical combination permeating through all items of a phenomenon. On the other hand, there cannot be too much variation, as this could mean that items differ in quality and that they might not even test the same thing. However, the low variances for item pairs in Tables 1 and 2 underline that our items, although differing in lexical choices, were constructed with a sufficient degree of consistency.

4.1 Participants' ability to distinguish tasks

To be sure that the participants were able to distinguish the two tasks, we examined the reference phenomena, that is, the contradictions and the resumptive pronouns, as well as the filler items.

For the contradictions and the resumptive pronouns, the results show the expected tendencies. The contradictions received low ratings in the semantic judgement task and reasonably high ratings in the syntactic judgement task. However, the syntactic ratings are somewhat reduced, whereas the ratings for the baseline items are, as one would expect, near the top of the scale. As to the resumptive pronouns, they received reasonably reduced ratings in the syntactic judgement task. However, the semantic ratings for the resumptive pronouns are also considerably depressed.

Where does the degradation in the ratings come from? In our view, the intended meaning of the items with resumptive pronouns is relatively straightforward, so they should have received high semantic ratings and reduced syntactic ratings. Yet, their semantic ratings are also somewhat

degraded, suggesting some kind of "syntactic drag", that is, a reduction in perceived meaningfulness caused by a syntactic anomaly. This reduction is against our expectations. We can think of two major ways to interpret this, which are not mutually exclusive but may contribute simultaneously to the observed effect. First, it is possible that the participants struggled to distinguish between the two tasks to some degree. They were asked to judge semantic acceptability, but instead of doing so, they fell back to some degree to judging some kind of general (un)naturalness. Second, participants might have struggled to parse the sentence properly because of the syntactic anomaly. As a result, they struggled to extract its meaning under the reasonable use of resources. This difficulty is predicted by syntax-first models of sentence comprehension (Frazier & Fodor 1978; Frazier 1979; Frazier & Clifton 1996). Under this modular view, autonomous syntactic processing precedes semantic processing, and syntactic representations are a prerequisite for interpretation. Since resumption is ungrammatical in our materials, parsing might have failed to provide a syntactic representation for the resumptive pronoun, which in turn hampered interpretation.

We expected that surface-oriented resort strategies, in the psycholinguistic literature referred to as *heuristics*, would help readers overcome the syntactic conflict. Those heuristics seem easy to apply in the case of resumptive pronouns. Since the resumptive pronoun occurs where otherwise a gap would occur, it could be treated as a spell-out of that gap. Furthermore, resumption results in a locally coherent string (*This is the girl that John likes her* includes *(that) John likes her*). In line with the good-enough approach (Ferreira 2003; Ferreira & Patson 2007), we expected that this local string would be sufficient for the purpose of interpreting the ungrammatical sentence. Perhaps it was, but it did not show in our results, possibly because the rating task triggered deeper (more than "good enough") processing. Notably, facilitating effects of resumptive pronouns in island contexts have only been found with online measures, not with acceptability ratings.

Further, the penalty in semantic ratings might reflect an increased processing effort to arrive at a meaning for the ungrammatical sentence. Beltrama & Xiang (2016) collected comprehensibility ratings and observed that resumption outside an island reduces comprehensibility compared to a gapping strategy. Although comprehensibility is different from meaningfulness/plausibility, it is very possible that comprehensibility affects semantic ratings just as ease of processing affects acceptability ratings in general.

Table 3. Mean ratings for the two subsets in the resumptive pronoun construction

		Mean ratings	
		syntactic	semantic
Subset 1	Baseline condition *Mary said that this is the girl that John likes.*	5.92	6.29
	Violation condition *Mary said that this is the girl that John likes her.*	2.80	4.49
Subset 2	Baseline condition *That's the girl but I don't know what she did.*	5.92	6.07
	Violation condition *That's the girl that I don't know what she did.*	3.90	5.13

To gain further insights into these outcomes, we would have to follow up on this with other kinds of experiments, namely, experimental methods providing online measures that tap into the ongoing processes of comprehension, for example, electrophysiological methods or eye tracking.

A side note on the resumptives: As mentioned in § 3.2, the items containing resumptive pronouns are heterogeneous. In one half of the items, the resumptive is an object, replaced by a gap in the baseline version; in the other half, the resumptive pronoun is the subject of the relative clause, and the subject pronoun of a root clause in the baseline version. We analysed the data for the two subsets. Interestingly, we found (almost) no difference in the baseline condition – that is, in the condition that was different in the two subsets – but a difference in the violation condition (Table 3). We leave it to future research to examine this difference in more detail.

The filler/anchor items that are clearly meaningful but syntactically marked exhibited similar ratings. Participants struggled to give high ratings to them in the semantic judgement task. This is in contrast to semantically marked items that are syntactically unmarked. Participants still gave reasonably high ratings to them in the syntactic judgement task. Hence, the syntactic deviations made it difficult to arrive at a meaningful interpretation, whereas the semantic deviations did not hamper recognition of the syntactic well-formedness (or did so only mildly).

These findings underline the importance of carefully controlling experimental items. This applies to both syntactic and semantic experiments.

For syntactic experiments, there is a need to control the items for potential semantic confounds, and for semantic experiments, for potential syntactic confounds.

4.2 All phenomena across tasks

We observed some common trends but also some differences between the four constructions. As expected, the contradictions received a strong penalty in the semantic ratings and only a mild penalty in the syntactic ratings. Conversely, resumptive pronouns in illicit positions received the strongest penalty in the syntactic ratings. Unexpectedly, they also received a substantial penalty in the semantic ratings, though a weaker penalty than the other three conditions. In the syntactic ratings, the contradictions were the best, and the resumptive pronouns the worst; in the semantic ratings, the reverse was true. The semantic restriction violations and the *for*-PPs gravitate in both tasks more towards the middle than the contradictions and the resumptive pronouns. Just like we observed a "syntactic drag" on semantic ratings, we note that there is a "semantic drag" on the syntactic ratings and that the extent varies with construction type. The stronger the semantic drag, the more tightly the corresponding semantic factor is connected to syntactic structure.

The contradictions in our experiment were independent of the syntactic structure. Similarly, the semantic restriction violations were well-formed in terms of their phrase structure. The situation is different for the *for*-PPs modifying a punctual verb. Under the assumption that the event structure is represented or limited by the phrase structure in the verbal domain, extended projections of punctual verbs lack a position that could host a time frame adverbial such as *for an hour*. In other words, the licensing of temporal *for*-PPs depends on the lexical semantics of the corresponding verb and the syntactic representation. Finally, the resumptive pronouns in our materials instantiate a proper syntactic violation, hence the strong penalty in the syntactic ratings and the comparatively low mean syntactic acceptability.

4.3 Discussion of the two critical phenomena: semantic restriction violations and aspectual mismatches with *for*-PPs

Comparing the results for semantic restriction violations and for the *for*-PPs in particular, we observe similarities as well as differences. Both phenomena received penalties in both tasks but to a different degree. The semantic

restriction violations received low ratings in the semantic task, and somewhat reduced ratings in the syntactic task. It seems that semantic restriction violations are violations that were perceived as mainly semantic in nature, with some syntactic impact. The *for*-PPs, on the other hand, seem to be a phenomenon right at the interface of syntax and semantics. They received low results in the semantic task and considerably reduced ratings in the syntactic task. In both cases, the ratings gravitate towards the middle of the rating scales.

It seems that for both phenomena, the source of the violations cannot entirely be located in just the grammar or in just semantic factors. This particularly applies to the *for*-PPs. They have an odd in-between status in two ways. First, the averaged ratings were in-between in both tasks. Second, the source that leads to the degradation is not clear, which also agrees with the discourse in the literature. *For*-PPs are sometimes identified as a semantic phenomenon, by marking it with a hash (e.g. Rothstein 2007), and sometimes as a syntactic phenomenon, by marking it with an asterisk (e.g. Vendler 1957; Dowty 1979; Engelberg 2000).

The two critical phenomena differ in several aspects, in particular the nature of the symptom indicating the violation and the means available for a repair. The *for*-PP items involve a syntactic symptom, namely the *for*-PP itself. Its incompatibility with the verb can be resolved by removing the *for*-PP or replacing it with a time-point adverbial, for example, *at 3pm sharp*. Neither repair option was available to the participants in our experiment simply because the *for*-PP was explicated in our stimuli. Theoretically, replacing the verb is another option, but this was not possible in our experimental setup with this kind of written presentation.

Another option to resolve the incompatibility is to shift the interpretation, that is, to apply aspectual coercion (Pustejovsky 1995; Jackendoff 1997; de Swart 1998; and, more recently, Dölling 2014; for an overview, see Lauwers & Willems 2011; for psycholinguistic evidence, see Piñango et al. 1999; Todorova et al. 2000; Bott 2008; 2010; Brennan & Pylkkänen 2008; and Townsend 2013; but Pickering et al. 2006). As described in § 3.1.2, we tried to hamper shifted readings, for example, by using singular subjects to make iterative readings less available. Nevertheless, we might have unintentionally facilitated aspectual coercion by presenting the verbs in the past progressive. The progressive might serve as an additional trigger for aspectual coercion. If so, it could have initiated aspectual coercion early in the sentence, at the verb, and thereby reduced the conflict arising at the clause-final *for*-PP. Consistent with this reasoning, coercion could be the reason why ratings in the violation

condition are not at the bottom. Further support for this suspicion comes from the observation that the semantic ratings exhibit a relatively wide dispersion. Possibly, coercion was easily available for some trials but not for all. Based on the current data, we cannot determine whether participants arrived at shifted readings or gave intermediate to higher ratings from time to time for a different reason. We simply do not have our participants' interpretations of the stimuli, so we are left guessing. Furthermore, our rating data are not suitable for investigating the time course aspectual coercion or inter-individual differences. We leave the thoughts outlined above to future research dedicated to the phenomenon itself.

Even if coercion occurred in some trials, the strong penalty visible in the semantic rating task suggests that coercion did not occur in all trials. Participants struggled to arrive at a meaningful interpretation. Yet, the semantic conflict does not (directly) explain the degradation in the syntactic rating task. Syntactic approaches to event structure assume that the event structure is specified by the syntactic representation, more precisely by functional projections such as TelicP, AspP, and other similar ones (cf. Borer 1994; 2005; Travis 1994; 2000; Ritter & Rosen 1998; Ramchand 2008; 2017). Under this view, the explanation for the syntactic penalty seems straightforward: The time-frame adverbial cannot be integrated into the syntactic representation when this representation lacks a corresponding projection. Durative verbs such as *wait* trigger the assembly of a syntactic structure, including an active AspP_{EM} projection (Borer 1994), or some equivalent, while punctual verbs do not. The missing syntactic structure can be added given the right context for aspectual coercion. Otherwise, the *for*-PP cannot be integrated, and the final syntactic structure is flawed. Participants in the experiment seemed to share the intuition that the licensing of time-frame adverbials was a matter of structure, not only meaning. We surmise that this intuition was based on parsing. The parser fails to compute a complete structure; this results in the impression that something is wrong with the structure and yields a degraded syntactic rating.

By contrast, semantic selectional restrictions are typically assumed to be purely semantic constraints not reflected in the syntactic structure. The assembly of phrase structure is independent of s-selection or, as Adger (2003: 89) puts it, "Merge does not inspect s-selectional properties". The animacy conflict can be resolved by personification coercion or by adjusting the world relative to which the felicity of the sentence is evaluated. Trucks, as in (2a), can be comforted when they are capable of feelings, for example, in

a cartoon or movie. Previous psycholinguistic and neurolinguistic research has shown that readers can shift their expectations regarding semantic compatibilities and easily adapt to anomalies in a supportive, for example, fictional context (e.g. Nieuwland & van Berkum 2006; Filik & Leuthold 2008; Bade & Buscher 2019). Participants in the current experiment apparently refrained from such an adjustment. We believe that three properties of our study contributed to their reluctance. First, in contrast to some previous studies, no fictional context was given. Second, the sentences in our experiments were not related to each other; that is, they did not form a story. Furthermore, most sentences in the study did not require such a fictional context. Taken together, the experimental setup provided little motivation, if any, to accommodate a fictional context. Therefore, the strong penalty in the semantic rating task is no surprise.

4.4 The distinction between syntax and semantics in grammatical frameworks

Our research question, as formulated in § 1, emanates from the premise that syntax and semantics are strictly separate, an assumption that is not shared by all grammar approaches. The relation between syntax and semantics is conceptualised in different ways. Some approaches such as various types of Construction Grammar (e.g. Goldberg 1995; 2006; Croft 2001; 2013; for an overview, see Hoffmann & Trousdale 2013), including Cognitive Grammar (Langacker 2008; 2010), assume that form and meaning are paired in lexicalised or templatic units (not only at the word level), while others conceptualise syntax and semantics as autonomous modules that communicate with each other via an interface. Modular approaches vary substantially in the design of the interface, the translation procedures and the principles of interpretation (for thorough discussions of the syntax-semantics interface, see Kuhn 2007; Lechner 2015). Constraint-based grammars such as Lexical Functional Grammar (Bresnan 1982; 2001; Bresnan et al. 2016), Head-driven Phrase Structure Grammar (Pollard & Sag 1994), Tree Adjoining Grammar (Joshi 1988), Categorial Grammar (Lambek 1958; Jacobson 1996), Combinatory Categorial Grammar (Steedman 1996; 2000; 2019; Steedman & Baldridge 2011), Parallel Architecture (Jackendoff 1997; 2002), and Synchronous Tree Adjoining Grammar (Nesson & Shieber 2006; Shieber 2014) posit a single representation from which operations in several modules are computed or several parallel representations that are linked in a

non-derivational way. Interfaces are sets of constraints on relations between modules. Syntax-centred models, in contrast, assume derivational relations between modules and give priority to syntax, from which representations in other subsystems are derived. Under this view, interfaces are unidirectional output-input relations. The syntax-semantics interface is the output of syntax and the input for semantics. All these theories would make different predictions for our experiment. A discussion of these predictions is beyond the scope of this paper. We have to leave it to the readers to decide whether the results agree with their expectations.

However, under the assumption of a strong link between form and meaning, evaluating one but not the other makes little sense. If so, the observed differences might be task-induced effects. After all, we explicitly asked the participants to differentiate between form and meaning, and they probably did the best they could do. For instance, they might have judged the overall likelihood of the contradicting parts in our contradictions holding simultaneously rather than judging the meaningfulness of the construction. Or they estimated the likelihood of the given form to occur with the intended meaning, for example the likelihood of durational *for*-PP with the given verb phrase under its conventional meaning.

The moderate correlation between syntactic and semantic ratings, which we found in our experiments, is compatible with the view that form and meaning are closely related to each other. The observed dissociations between the results of the two rating tasks, however, are an interesting finding that calls for an explanation. By contrast, the view that syntax and semantics are separated is challenged by the moderate correlation, but compatible with the observed dissociations. In other words, frameworks that assume a clear divide between syntax and semantics, and frameworks that assume a close mapping of the two make basically complementary predictions. Either syntactic and semantic ratings should deviate for the constructions under investigation here or they should adhere to each other. Our findings are not (fully) compatible with either view. Despite the mixed results, we believe that an experimental approach – as presented in this paper – is suitable for explicit theory testing. Paired experiments parallelly probing syntactic and semantic judgements provide a valuable tool for testing models of the interaction between syntax and semantics.[8]

[8] Many thanks to the reviewer who underlined the need for and the importance of this subsection.

4.5 Expert versus non-expert intuitions

Our participants mixed up syntactic and semantic factors to some degree. In light of the discussion of expert versus non-expert intuitions (e. g., Culbertson & Gross 2009; Devitt 2014), one could ask whether linguists would give judgements that distinguish more between semantics and syntax. This touches on the question of how linguistic intuitions come about (cf. Schindler et al. 2020). According to Culbertson & Gross (2009), no considerable difference between non-expert ratings and expert ratings is to be expected. This is in contrast to Devitt (2014), who would predict that experts are able to distinguish the two tasks better, as their intuitions are shaped by additional experiences. A recent study by Fanselow et al. (2019) provided data that could support Devitt's view. Fanselow et al. (2019) observed that when evaluating a sequence like *Who wonders who bought what*, syntacticians differ from other linguists in their interpretation. Such differences could come from theory-driven biases or from different processing strategies. Our experimental setup, when administered to non-experts versus experts, might produce data that offer further insights into the question of how intuitions come about and whether expertise makes a difference.

5 Conclusion

We examined two phenomena – semantic restriction violations and violations around *for*-PPs – and tested them in two experiments, a semantic judgement task and a syntactic judgement task. The experiments included two reference phenomena: semantic contradictions, which are syntactically well-formed but semantically marked, and resumptive pronouns, which are semantically licit; that is, their meaning is clear, but they are syntactically marked. Our results indicate that the participants somewhat struggled with the two tasks. Importantly, however, the participants distinguished the two tasks sufficiently well for the results to be interpretable. For the two critical phenomena, semantic restriction violations and aspectual mismatches with *for*-PPs, we observed that they received reduced ratings in both tasks. These findings call into question whether the source of the perceived degradation can be unambiguously pinpointed to either grammatical constraints or semantic factors. Rather, it seems that both structural properties and semantic restrictions contribute to their degradation. However, validation using other experimental methods is needed. It would also be interesting to determine whether linguists are better

at distinguishing between the two tasks. This would help to gain a deeper understanding of how linguistic intuitions come about.

As a final remark, we would like to emphasise the value of experimental methods for theoretical linguistics. Theories make predictions which experiments can test. The outcome of experiments can inspire adjustments in theoretical accounts, which in turn will create new predictions to be tested with experiments. In our view, such an interplay of theoretical and empirical (not only experimental) approaches is fruitful and of benefit for linguistics in general.

Acknowledgements

We wish to thank the reviewers for their valuable feedback. We also thank the participants of Linguistic Evidence 2018 for the discussions (the present paper builds in parts on our paper published in the proceedings of Linguistic Evidence 2018). Sam Featherston was so kind to share the anchor items and results with us before the group's work was published. Further, we deeply appreciate the feedback from Tom Wasow, Elaine Francis, and Oliver Bott.

References

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

Asudeh, Ash. 2012. *The logic of pronominal resumption*. Oxford: Oxford University Press.

Bade, Nadine & Buscher, Frauke. 2019. An experimental comparison of two reinterpretation strategies: Benefits and challenges of using fictional contexts in experimental studies. In Gattnar, Anja & Hörnig, Robin & Störzer, Melanie & Featherston, Sam (eds.), *Proceedings of Linguistic Evidence 2018: Experimental data drives linguistic theory*. Tübingen: University of Tübingen. DOI: 10.15496/publikation-32624.

Bard, Ellen G. & Robertson, Dan & Sorace, Antonella. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68.

Bartoń, Kamil. 2020. *Mu-MIn: Multi-model inference*. R Package Version 1.43.17. (http://R-Forge.R-project.org/projects/mumin/). (Accessed 2021-07-04).

Beltrama, Andrea & Xiang, Ming. 2016. Unacceptable but comprehensible: The facilitation effect of resumptive pronouns. *Glossa* 1(1). 29. DOI: 10.5334/gjgl.24.

Borer, Hagit. 1994. The projection of arguments. In Benedicto, Elena & Runner, Jeff (eds.), *Functional projections*. (University of Massachusetts Occasional Papers 17). Amherst, MA: Graduate Linguistic Student Association, University of Massachusetts.

— 2005. *Structuring sense, vol. 2: The normal course of events*. Oxford: Oxford University Press.

Bott, Oliver. 2008. Doing it again and again may be difficult, but it depends on what you are doing. In Abner, Natasha & Bishop, Jason (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 63–71. Somerville, MA: Cascadilla Proceedings Project.

— 2010. *The processing of events*. Amsterdam: John Benjamins.

Brennan, Jonathan & Pylkkänen, Liina. 2008. Processing events: Behavioral and neuromagnetic correlates of aspectual coercion. *Brain & Language* 106. 132–143.

Bresnan, Joan (ed.). 1982. *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Bresnan, Joan. 2001. *Lexical functional syntax*. Cambridge: Blackwell.

Bresnan, Joan & Asudeh, Ash & Toivonen, Ida & Wechsler, Stephen. 2016. *Lexical-functional syntax*. 2nd edn. (Blackwell Textbooks in Linguistics 16). Oxford: Wiley Blackwell.

Cann, Ronnie & Kaplan, Tami & Kempson, Ruth. 2004. Data at the grammar-pragmatics interface: The case of resumptive pronouns in English. *Lingua* 115. 1551–1578.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

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

Chomsky, Noam & Miller, George A. 1963. Introduction to the formal analysis of natural languages. In Luce, R. Duncan & Bush, Robert R. & Galanter, Eugene (eds.), *Handbook of mathematical psychology,* vol. 2, 269–321. New York City, NY: Wiley.

Christiansen, Morten H. & MacDonald, Maryellen C. 2009. A usage-based approach to recursion in sentence processing. *Language Learning* 59. 126–161.

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

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

— 2013. Radical construction grammar. In Hoffmann, Thomas & Trousdale, Graeme (eds.), *The Oxford handbook of construction grammar*, 211–232. Oxford: Oxford University Press.

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

de Swart, Henriëtte. 1998. Aspect shift and coercion. *Natural Language and Linguistic Theory* 16. 347–385.

Devitt, Michael. 2014. Linguistic intuitions and cognitive penetrability. *Baltic International Yearbook of Cognition, Logic and Communication* 9. DOI: 10.4148/1944-3676.1083.

Dickey, Michael Walsh. 1996. Constraints on the sentence processor and the distribution of RPs. In Dickey, Michael Walsh & Tunstall, Susanne (eds.), *Linguistics in the laboratory*, 157–191. (University of Massachusetts Occasional Papers 19). Amherst, MA: Graduate Linguistic Student Association, University of Massachusetts.

Dölling, Johannes. 2014. Aspectual coercion and eventuality structure. In Robering, Klaus (ed.), *Events, arguments and aspects: Topics in the semantics of verbs*, 189–226. Amsterdam: John Benjamins.

Dowty, David R. 1979. *Word meaning and Montague grammar*. Dordrecht: Kluwer.

Engelberg, Stefan. 2000. *Verben, Ereignisse und das Lexikon*. Tübingen: Niemeyer.

Erteschik-Shir, Nomi. 1992. Resumptive pronouns in islands. In Goodluck, Helen & Rochemont, Michael (eds.), *Island constraints: Theory, acquisition and processing*, 89–108. Dordrecht: Springer.

Etxeberria, Urtzi & Tubau, Susagna & Deprez, Viviane & Borràs-Comes, Joan & Espinal, M. Teresa. 2018. Relating (un)acceptability to interpretation: Experimental investigations on negation. *Frontiers in Psychology* 8(2370). DOI: 10.3389/fpsyg.2017.02370.

Fanselow, Gisbert & Häussler, Jana & Weskott, Thomas. 2019. Who cares what who prefers? In Carlson, Katy & Clifton, Charles Jr. & Fodor, Janet Dean (eds.), *Grammatical approaches to language processing: Essays in honor of Lyn Frazier*, 261–274. Cham: Springer.

Ferreira, Fernanda. 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology* 47. 164–203.

Ferreira, Fernanda & Patson, Nikole D. 2007. The 'good enough' approach to language comprehension. *Language and Linguistics Compass* 1(1–2). 71–83.

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

Filik, Ruth & Leuthold, Hartmut. 2008. Processing local pragmatic anomalies in fictional contexts: Evidence from the N400. *Psychophysiology* 45(4). 554–558.

Frank, Stefan L. & Ernst, Patty. 2019. Judgements about double-embedded relative clauses differ between languages. *Psychological Research* 83. 1581–1593.

Frank, Stefan L. & Trompenaars, Thijs & Vasishth, Shravan. 2016. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science* 40. 554–578.

Frazier, Lyn. 1979. *On comprehending sentences: Syntactic parsing strategies*. University of Connecticut. (Doctoral dissertation).

— 1985. Syntactic complexity. In Dowty, David R. & Karttunen, Lauri & Zwicky, Arnold (eds.), *Natural language processing: Psychological, computational, and theoretical perspectives*, 129–189. Cambridge: Cambridge University Press.

Frazier, Lyn & Clifton, Charles Jr. 1996. *Construal*. Cambridge, MA: MIT Press.

Frazier, Lyn & Fodor, Janet Dean. 1978. The sausage machine: A new two-stage parsing model. *Cognition* 6. 291–325.

Gerbrich, Hannah & Schreier, Vivian & Featherston, Sam. 2019. Standard items for English judgement studies: Syntax and semantics. In Featherston, Sam & Hörnig, Robin & von Wietersheim, Sophie & Winkler, Susanne (eds.), *Experiments in focus: Information structure and semantic processing*, 305–328. Berlin: De Gruyter.

Gibson, Edward & Piantadosi, Steve & Fedorenko, Kristina. 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass* 5(8). 509–524. DOI: 10.1111/j.1749-818x.2011.00295.x.

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

Gimenes, Manuel & Rigalleau, François & Gaonac'h, Daniel. 2009. When a missing verb makes a French sentence more acceptable. *Language and Cognitive Processes* 24. 440–449.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.

— 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199268511.001.0001.

Grimshaw, Jane. 1979. Complement selection and the lexicon. *Linguistic Inquiry* 10(2). 279–326.

Hagoort, Peter & Hald, Lea & Bastiaansen, Marcel & Petersson, Karl Magnus. 2004. Integration of word meaning and world knowledge in language comprehension. *Science* 304(5669). 438–441.

Häussler, Jana & Juzek, Tom S. 2016. Detecting and discouraging non-cooperative behavior in online experiments using an acceptability judgement task. In Christ, Hanna & Klenovšak, Daniel & Sönning, Lukas & Werner, Valentin (eds.), *A blend of MaLT: Selected contributions from the Methods and Linguistic Theories Symposium 2015*, 73–99. Bamberg: University of Bamberg Press.

— 2021. Variation in participants and stimuli in acceptability experiments. In Goodall, Grant (ed.), *The Cambridge handbook of experimental syntax*, 97–117. Cambridge: Cambridge University Press.

Heestand, Dustin & Xiang, Ming & Polinsky, Maria. 2011. Resumption still does not rescue islands. *Linguistic Inquiry* 42. 138–152.

Hoffmann, Thomas & Trousdale, Graeme (eds.). 2013. *The Oxford handbook of construction grammar*. Oxford: Oxford University Press.

Hofmeister, Philip & Norcliffe, Elisabeth. 2013. Does resumption facilitate sentence comprehension? In Hofmeister, Philip & Norcliffe, Elisabeth (eds.), *The core and the periphery: Data-driven perspectives on syntax inspired by Ian A. Sag*, 225–246. Stanford, CA: CSLI Publications.

Horn, Laurence R. 2018. Contradiction. In Zalta, Edward N. (ed.), *The Stanford encyclopedia of philosophy*. Winter 2018 edn. Stanford, CA: Metaphysics Research Lab, Stanford University. (https://plato.stanford.edu/archives/win2018/entries/contradiction). (Accessed 2021-01-02).

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.

— 2002. *Foundations of language*. Oxford: Oxford University Press.

Jacobson, Pauline. 1996. Semantics in categorial grammar. In Lappin, Shalom (ed.), *The handbook of contemporary semantic theory*, 89–116. Oxford: Basic Blackwell.

Joshi, Aravind. 1988. Tree adjoining grammars. In Dowty, David & Karttunen, Lauri & Zwicky, Arnold (eds.), *Natural language parsing*, 206–250. Cambridge: Cambridge University Press.

Keffala, Bethany. 2013. Resumption and gaps in English relative clauses: Relative acceptability creates an illusion of 'saving'. In Cathcart, Chundra & Chen, I-Husan & Finley, Greg & Kang, Shinae & Sandy, Clare S. & Stickles, Elise (eds.), *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, 140–154. Berkeley, CA: Berkeley Linguistics Society.

Keffala, Bethany & Goodall, Grant. 2011. Do resumptive pronouns ever rescue illicit gaps in English? (Poster presented at CUNY 2011 Conference on Human Sentence Processing, Stanford University, 24–26 March 2011).

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Edinburgh: University of Edinburgh. (Doctoral dissertation).

Kotek, Hadas & Dockum, Rikker & Babinski, Sarah & Geissler, Christopher. 2021. Gender bias and stereotypes in linguistic example sentences. (Forthcoming paper, lingbuzz/005367). (https://ling.auf.net/lingbuzz/005367). (Accessed 2021-06-27).

Kuhn, Jonas. 2007. Interfaces in constraint-based theories of grammar. In Ramchand, Gillian & Reiss, Charles (eds.), *The Oxford handbook of linguistic interfaces*, 613–650. Oxford: Oxford University Press.

Kuznetsova, Alexandra & Brockhoff, Per B. & Christensen, Rune H. B. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. DOI: 10.18637/jss.v082.i13.

Lambek, Joachim. 1958. The mathematics of sentence structure. *American Mathematical Monthly* 65. 154–169.

Langacker, Ronald W. 2008. *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

— 2010. Cognitive grammar. In Heine, Bernd & Narrog, Heiko (eds.), *The Oxford handbook of linguistic analysis*, 87–109. Oxford: Oxford University Press.

Lauwers, Peter & Willems, Dominique. 2011. Coercion: Definition and challenges, current approaches, and new trends. *Linguistics* 49(6). 1219–1235.

Lechner, Winfried. 2015. The syntax-semantics interface. In Alexiadou, Artemis & Kiss, Tibor (eds.), *Syntax – theory and analysis: An international handbook,* vol. 2, 1198–1255. Berlin: Mouton de Gruyter.

McCloskey, James. 2006. Resumption. In Everaert, Martin & van Riemsdijk, Henk (eds.), *The Blackwell companion to syntax,* vol. 4, 94–117. Oxford: Blackwell.

McDaniel, Dana & Cowart, Wayne. 1999. Experimental evidence for a minimalist account of English resumptive pronouns. *Cognition* 70. B15–B24.

McKee, Cecile & McDaniel, Dana. 2001. Resumptive pronouns in English relative clauses. *Language Acquisition* 9(2). 113–156.

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

Nesson, Rebecca & Shieber, Stuart M. 2006. Simpler TAG semantics through synchronization. In Wintner, Shuly (ed.), *Proceedings of the 11th Conference on Formal Grammar*, 129–142. Stanford, CA: CSLI Publications.

Nieuwland, Mante S. & van Berkum, Jos J. A. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience* 18. 1098–1111.

Pesetsky, David. 1982. *Paths and categories*. Cambridge, MA: MIT. (Doctoral dissertation).

Phillips, Colin & Wagers, Matthew & Lau, Ellen F. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In Runner, Jeffrey (ed.), *Experiments at the interfaces*, 147–180. Bingley: Emerald Group.

Pickering, Martin J. & McElree, Brian & Frisson, Steven & Chen, Lillian & Traxler, Matthew J. 2006. Underspecification and aspectual coercion. *Discourse Processes* 42. 131–155.

Piñango, Maria M. & Zurif, Edgar & Jackendoff, Ray. 1999. Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistic Research* 28. 395–414.

Polinsky, Maria & Clemens, Lauren Eby & Morgan, Adam Milton & Xiang, Ming & Heestand, Dustin. 2013. Resumption in English. In Sprouse, Jon & Hornstein, Norbert (eds.), *Experimental syntax and island effects*, 341–360. Cambridge: Cambridge University Press.

Pollard, Carl & Sag, Ivan A. 1987. *Information-based syntax and semantics*. (CSLI Lecture Notes 13). Chicago, IL: Center for the Study of Language & Information.

— 1994. *Head driven phrase structure grammar*. Chicago: CSLI/Chicago University Press.

Prince, Ellen F. 1990. Syntax and discourse: A look at resumptive pronouns. In Hall, Kira & Koenig, Jean-Pierre & Meacham, Michael & Reinman, Sondra & Sutton, Laurel A. (eds.), *Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society*, 482–497. Berkeley, CA: Berkeley Linguistics Society. DOI: 10.3765/bls.v16i0.1719.

Pustejovsky, James. 1991. The syntax of event structure. *Cognition* 41. 47–81.

— 1995. *The generative lexicon*. Cambridge, MA: MIT Press.

Pylkkänen, Liina & Oliveri, Bridget & Smart, Andrew J. 2009. Semantics vs. world knowledge in prefrontal cortex. *Language and Cognitive Processes* 24(9). 1313–1334.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. (https://www.R-project.org/). (Accessed 2021-03-31).

Radford, Andrew. 2018. *Colloquial English: Structure and variation*. Cambridge: Cambridge University Press.

— 2019. *Relative clauses: Structure and variation in everyday English*. Cambridge: Cambridge University Press.

Ramchand, Gillian. 2008. *Verb meaning and the lexicon*. Cambridge: Cambridge University Press.

— 2017. The event domain. In D'Alessandro, Roberta & Franco, Irene & Gallego, Ángel J. (eds.), *The verbal domain*, 233–254. Oxford: Oxford University Press.

Ritter, Elizabeth & Rosen, Sara Thomas. 1998. Delimiting events in syntax. In Butt, Miriam & Geuder, Wilhelm (eds.), *The projection of arguments: Lexical and syntactic constraints*, 135–164. Stanford, CA: Center for the Study of Language & Information.

Rothstein, Susan. 2007. Two puzzles for a theory of lexical aspect: The case of semelfactives and degree achievements. In Dölling, Johannes & Heyde-Zybatow, Tatjana & Schäfer, Martin (eds.), *Event structures in linguistic form and interpretation*, 175–197. Berlin: Mouton de Gruyter.

Schindler, Samuel & Drożdżowicz, Anna & Brøker, Karen. 2020. *Linguistic intuitions: Evidence and method*. Oxford: Oxford University Press.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press. (Reprint available: https://langsci-press.org/catalog/book/89).

Shieber, Stuart M. 2014. Bimorphisms and synchronous grammars. *Journal of Language Modelling* 2(1). 51–104.

Sprouse, Jon. 2008. The differential sensitivity of acceptability to processing effects. *Linguistic Inquiry* 39(4). 686–694.

Steedman, Mark. 1996. *Surface structure and interpretation*. Cambridge, MA: MIT Press.

— 2000. *The syntactic process*. Cambridge, MA: The MIT Press.

— 2019. Combinatory categorial grammar. In Kertész, András & Moravcsik, Edith & Rákosi, Csilla (eds.), *Current approaches to syntax: A comparative handbook*, 389–420. Berlin: Mouton de Gruyter.

Steedman, Mark & Baldridge, Jason. 2011. Combinatory categorial grammar. In Borsley, Robert & Borjars, Kersti (eds.), *Non-transformational syntax: Formal and explicit models of grammar*, 181–224. New York: Blackwell.

Todorova, Marina & Straub, Kathy & Badecker, William & Frank, Robert. 2000. Aspectual coercion and the online computation of sentential aspect. In Gleitman, Lila R. & Joshi, Aravind K. (eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, 523–528. Mahwah, NJ: Lawrence Erlbaum.

Townsend, David J. 2013. Aspectual coercion in eye movements. *Journal of Psycholinguistic Research* 42. 281–306.

Travis, Lisa. 1994. Event phrase and a theory of functional categories. In Koskinen, Päivi (ed.), *Proceedings of the 1994 Annual Conference of the Canadian Linguistic Association*, 559–570. Toronto: University of Toronto.

— 2000. Event structure in syntax. In Pustejovsky, James & Tenny, Carol (eds.), *Events as grammatical objects*, 145–186. Stanford, CA: Center for the Study of Language & Information.

Vasishth, Shravan & Suckow, Katja & Lewis, Richard L. & Kern, Sabine. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes* 25(4). 533–567.

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2). 143–160.

Warren, Tessa & McConnell, Kerry. 2007. Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review* 14(4). 770–775.

Warren, Tessa & Milburn, Evelyn & Patson, Nikole D. & Dickey, Michael Walsh. 2015. Comprehending the impossible: What role do selectional restriction violations play? *Language, Cognition and Neuroscience* 30(8). 932–939.

Contact Information:

Jana Häussler
Faculty of Linguistics and Literary Studies
Bielefeld University
Universitätsstraße 25, 33615 Bielefeld
Germany
e-mail: jana(dot)haeussler(at)uni-bielefeld(dot)de

Tom S. Juzek
Department of Linguistics
Ruhr University Bochum
Universitätsstraße 150, 44801 Bochum
Germany
e-mail: tom(dot)juzek(at)posteo(dot)net