Presupposing Novel Information: A Cross-Trigger Experiment in English

The Authors

October 31, 2023

Abstract

This paper brings new empirical evidence to the *novelty problem* for presuppositions: why can some presupposition triggers be used to introduce novel information into a discourse while others cannot? For example, "Avi is still reading *Crime and Punishment*" may sound infelicitous unless it has been established that Avi was reading it previously. However, "Avi misplaced her copy of *Crime and Punishment*" is an acceptable way to introduce the novel information that she owns a copy. We conduct an online acceptability study that assesses fourteen English triggers on whether or not their presuppositions can be used to convey novel information, which we refer to as a trigger's Contextual Felicity Constraint, or CFC following Tonhauser et al. (2013). We show that there is substantial trigger-by-trigger variation, that semantically-like triggers have similar CFCs, and that triggers fall (roughly) into three clusters—triggers that impose no CFCs (possessives, change-of-state predicates, factives), triggers that impose weak CFCs (non-focus iteratives and additives), and triggers that impose strong CFCs (focus-sensitive additives). We show that a trigger's CFC in our experiment is highly correlated ($\rho > 0.9$) with the proportion of times a trigger is used to convey novel information in a production corpus (Spenader, 2002). We then evaluate the results of our study against previous proposals for novelty effects and endorse a theory that treats presuppositions as fundamentally hybrid—some merely impose constraints on their local context, however others include an (additional) anaphoric element, which is less amenable to accommodation. We conclude by linking trigger anaphoricity to obligatoriness as articulated in Bade (2016) and argue that this property can explain a wide variety of trigger behavior.

Word Count: 12906

1 Introduction

Presuppositions are the parts of meanings of utterances that are backgrounded, seemingly non-novel, and often survive various entailment-canceling operations.[1] Presuppositions can be introduced by individual lexical items called *presupposition triggers*. To give a brief example, in the sentence "Alex spilled coffee again" the trigger *again* introduces the presupposition that Alex has spilled coffee previously. Typically, triggers are allergic to environments where the context does not support (i.e. imply, though not necessarily logically entail) their presuppositions. In the example, it would be infelicitous for a speaker to utter the sentence above unless the context has put into the common ground among the conversational participants that Alex has spilled coffee previously. However, sometimes presuppositions *can* be used to bring new information into a discourse. Example (1) gives the basic variation: both (1-a) and (1-b) presuppose the same content (that someone spilled coffee on the floor), but the first is a more natural continuation given the minimal context.

(1)   Alex is having a bad morning at work...

      a.   ... she is annoyed that someone spilled coffee on the floor.
      b. #... it was Amos who spilled coffee on the floor.

This puzzle—why is (1-a) good whereas (1-b) is not?—goes by a number of different names in the literature. In this paper we use the following terminology: We'll refer to the empirical puzzle outlined in (1) as the *novelty problem* for presuppositions (or as *novelty effects*). When triggers occur in cases where their presuppositions are not supported by the local context, we'll call this an *informative usage* of the trigger or a case of *informative presupposition* (von Fintel, 2008; Beaver and Zeevat, 2007). When talking about the variation in frequency of informative usage between triggers, or the variation in acceptability for the same trigger between different contexts we will discuss things in terms of variation in *contextual felicity constraint* strength, or CFC strength following Tonhauser et al. (2013).[2]

---

[1]Following Göbel (2020), we treat presuppositions as a natural class of meanings. Formally, they are the parts of meanings of an utterance that cannot be targeted by clause-level negation (i.e. they are backgrounded and not at-issue); cannot be canceled by the speaker (they are strongly committed-to); and have an obligatorily local effect (they are necessarily not speaker-oriented). For a more detailed discussion and examples of diagnostic tests for presuppositions see Tonhauser et al. (2013).

[2]Although to the term *missing accommodation* is also used in the literature (roughly in the same way as we we use *informative presupposition*) we will generally not use this term, as it takes a stronger theoretical stance about what is going on under the hood, i.e. that the CFC variation is due to accommodation failure.

One major block in resolving the novelty problem is the relative scarcity of robust empirical evidence about when triggers can be used informatively. We gain traction on the problem by conducting an acceptability study that quantifies the strength of a trigger's CFC. In our study, participants rate sentences with triggers whose presuppositions are both supported and unsupported by the context (i.e. they are used informatively). Adopting terminology from Tonhauser et al. (2013), if a trigger is infelicitous in contexts where its presuppositions are not supported by the context we say that it is subject to a Contextual Felicity Constraint, or CFC. This approach builds on previous studies testing contextual felicity (Wilcox et al., 2021; Göbel, 2020) but it is novel insofar as we test triggers in a variety of semantic environments, and we test fourteen different types of English triggers, making this the largest cross-trigger comparison reported in the literature to-date. The major questions we aim to resolve with these data are: (a) Is there trigger-by-trigger variation in terms of CFC strength? (b) Does the variation follow known semantic properties of the triggers? (c) How does CFC strength vary with semantic environment? and (d) Do any current theoretical accounts of the novelty problem explain the observed variation? We find that the majority of triggers tested are subject to a Contextual Felicity Constraint, but that there is significant variation between the triggers. Crucially, semantically similar triggers tend to cluster together; for example, additive particles (e.g. *too*, *even*, *still*) are all associated with the strongest CFCs, whereas presuppositions associated with verbs (e.g. *stop*, *find*, *know*) are all associated with weak-to-no CFCs. This supports the hypothesis that semantic features of the trigger affect its relative contextual felicity. To validate our methods, we compare our comprehension-side results to the proportion of times triggers are used informatively in production data (Spenader, 2002), finding that that there is a (remarkably) strong correlation between the two. This suggests that the same underlying mechanism determines informative usage from both the production side and the comprehension side.

Comparing our data to the predictions of previous theoretical accounts of the novelty problem, we find that many identify some semantic features which are clearly implicated in CFC variation, but that some have better overall empirical coverage than others. We find that the two theories with best fit to our data are those which root CFC effects in a trigger's anaphoricity and whether or not it requires optional discourse repair (i.e. its *weak/strong* status). We conclude by suggesting that both of these properties are actually determined by a trigger's *obligatoriness*, as articulated in Bade (2016). Specifically, Bade argues that some triggers are obligatory in order to cancel inferences that lead to semantic contradictions. We suggest that the same features that make a trigger obligatory also result in its strong CFCs, and that this in-

terpretation unites two related phenomena in the presupposition literature. We see this suggestion is essentially in the same spirit as Abrusán (2011) and Abrusán (2016)—as an attempt to resolve theoretical debates through a close analysis of a trigger's discourse structural and logical properties. Before we outline the paper in greater detail, we first turn to a more formal introduction of presupposition and accommodation.

1.1 Presupposition and Accommodation

We will approach presuppositions as a semantic phenomenon (*c.f.* Stalnaker 1973; Schlenker 2008); they are requirements imposed by a particular utterance on the context in which it is uttered. Within this overall framework, presuppositions can be formally modeled in one of two ways—either as anaphoric elements that, like pronouns, must be bound by a discourse referent (van der Sandt, 1992), or as imposing satisfaction conditions, or requirements, which must be entailed by the context for the utterance to be successfully interpreted (Heim, 1988). Following Stalnaker (2002), contexts are taken to be unstructured sets of information which have been mutually assented-to previously during discourse.

While the presuppositions-as-requirements approach is good at explaining some things, such as presuppositions' projective behavior, it under predicts the distribution of triggers in natural language. Specifically, if presuppositions are requirements on the context, then sentences with presupposition triggers should never be able to be uttered in cases where their presuppositions are not met. However, (1-a) demonstrates, there are cases where presuppositions *can* be used informatively. To account for this behavior, a mechanism called *accommodation* is invoked (Lewis, 1979), which pre-updates the context prior to utterance interpretation so that the requirements imposed by the presupposition are met. Accommodation was first introduced in the following way:

> If at time t something is said that requires presupposition P to be acceptable, and if P is not presupposed just before t, then—*ceteris paribus* and within certain limits—presupposition P comes into existence at t. (Lewis, 1979)

But merely postulating the existence of an accommodation mechanism doesn't solve the problem. Without spelling out the certain limits, the theory of presuppositions as requirements-plus-accommodation is neither explanatory nor predictive. Attempting to fill the current theoretical gap, there have been three types of factors that have been discussed as potentially influencing whether the mechanism is successful, including (i)

4

social aspects of a situation, such as the amount of trust between speakers or the local goals of the conversation (von Fintel, 2008); (ii) information theoretic properties of the presupposition, such as how much information it communicates and how likely that information is (Lassiter, 2012; van der Sandt and Geurts, 2001); and (iii) formal aspects of the trigger itself, such as the syntax of the sentence in which it is uttered, the semantic relationship between the presupposed and the asserted content, as well as local information-structural considerations (Blutner, 2000; Göbel, 2020). This paper is primarily concerned with (iii), but we review each of these, briefly, below.

Discussing informative presuppositions, von Fintel (2008) suggests multiple reasons why accommodation may or may not be successful. In unsuccessful cases, instead of incorporating presupposed material into the common ground, participants may actively challenge it, or ask for clarifying questions about it. One example of a case where participants are likely to challenge a presupposition instead of simply accommodating it is when they know it to be false. In cases of successful accommodation, there may be a number of different reasons why participants choose not to challenge the presupposed material:

> Informative use of presupposition may be successful in two particular kinds of circumstances: (i) the listeners may be genuinely agnostic as to the truth of the relevant proposition, assume that the speaker knows about its truth and trust the speaker not to speak inappropriately or falsely; (ii) the listeners may not want to challenge the speaker about the presupposed proposition because it is irrelevant to their concerns and because the smoothness of the conversation is important enough to them to warrant little leeway. (von Fintel, 2008)

The two pieces that we want to draw out here are the suggestion that trust between interlocutors, as well as knowledge of each interlocutor's local goal may influence a presupposition's contextual felicity. That is, extra-linguistic social factors can influence whether a presupposition-bearing utterance is challenged or accepted during conversation. At an even more abstract level, certain social settings may influence whether or not accommodation is successful; for example, accommodation may be eschewed in a court of law.

Turning now to information-theoretic influences, there is a longstanding hypothesis in the literature that when presuppositions *can* felicitously introduce novel information into the discourse, the information must be unsurprising, uncontroversial or high probability (Singh et al., 2016; van der Sandt and Geurts, 2001). For example, the response in (2-b) is typically judged as infelicitous in a context where the question-asker does

not know that Isaiah has a civet (a type of nocturnal animal indigenous to Southeast Asia that would not make a good pet), whereas (2-a) would be fine.

(2)   Q: What's new with Isaiah?

    a.   A: He had to take his cat to the vet this weekend.

    b.   A: #He had to take his civet to the vet this weekend.

Caching this out in technical terms, Lassiter (2012) suggests that informative usages may be related to the information content of the presupposed proposition, specifically to its *surprisal* (or its negative log probability given the conversational context).[3] This interaction between plausibility and presupposition has been demonstrated experimentally by Singh et al. (2016), who show that participants have a harder time accepting implausible material if it is presupposed, rather than asserted.

Finally, there are numerous proposals suggesting that formal aspects of a presupposition trigger can influence its ability to introduce novel information. By formal aspects we mean the semantic denotation of the trigger itself, as well as local information-structural considerations (Kripke, 2009; Göbel, 2020). As evaluating these various theoretical proposals is one of the main contributions of the present work, we refrain from discussing them here, and instead introduce them in greater detail in the next section.[4]

The rest of this paper will proceed as follows: In Section 2 we introduce previous structural proposals for CFC variation and outline the predictions each makes for the fourteen triggers we test. In Section 3 we introduce the methods used in our semantic acceptability study, which features a 2x2 interaction design between whether a trigger is present and whether it is supported (i.e. whether its presuppositions are entailed by the context). In Section 4 presents our results. In section 5 we compare our results to the various proposals for novelty effects, and endorse a theory that roots novelty in trigger anaphoricity, which is determined by its obligatoriness. We conclude by discussing two challenges that this approach faces, going forward.

---

[3]One point that seems to be overlooked when examples like this are given is that it's not necessarily the case that high surprisal content can be introduced felicitously into a discourse, even if it's not presupposed. Simply asserting that one has a pet civet and expecting this information to be accepted without further elaboration would be very odd.

[4]We want to briefly mention a third alternative for informative presupposition: Tonhauser (2015) argues that the reason why some presuppositions can be accommodated is because they are not presuppositions but rather backgrounded, projective material that is not associated with a contextual requirement. While we acknowledge that this hypothesis is entirely possible, within the context of our experiments it would also be circular. That is, we cannot explain a lack of contextual requirement by stipulating that some triggers merely lack a contextual requirement.

2   Background

In this section, we introduce five structural hypotheses for CFC variation. The first three were formulated explicitly in response to the novelty problem, attempting to explain why some presuppositions can be used informatively and others cannot. The last two were developed to explain other empirical differences between triggers, such as their ability to be cancelled by the speaker, however we include them here as additional structural considerations that may affect contextual felicity.

Table 1 identifies the fourteen types of triggers that we will test in the rating study, along with an example and a brief description of their associated presupposition.[5] The two right-hand columns give simple non-presupposing alternatives for each trigger, where they exist, which will be crucial for evaluating the competition model of Blutner (2000), discussed below.

2.1   Previous Proposals for CFC Variation

The first two hypotheses we discuss posit that presupposition triggers are a fundamentally heterogeneous class. At their core is the intuition that lexical items can trigger presuppositions in multiple ways. CFC variation comes about because one of these two types of triggers is easier to accommodate than the other.

2.1.1   Anaporic vs. Satisfaction Hypothesis:

Under the strong version of this approach *all* presuppositions are anaphors, and when a presupposition is used without local support an antecedent is built on-the-fly, which is then used to bind the trigger. There is, in addition, a *weak* form of this approach, which postulates that some triggers impose constraints while others are anaphoric (Zeevat, 1992; Kripke, 2009), with particular attention having historically been given to the additives *too* and *again*. The idea, here, is that while interlocutors can adjust their representation of the context, they cannot easily admit new discourse referents into the conversation and thus anaphoric presupposition triggers should be harder to accommodate than satisfaction-based triggers. One challenge for that approach is to come up with independent mechanisms for determining when a trigger is anaphoric and when it is satisfaction-based, otherwise the explanation becomes circular. For the purposes of evaluating the predictions of this theory, we will treat additive particles as

---

[5]For the purposes of this study, we focus on the existence presupposition for *the* following (Strawson, 1950), and not on the ways that the definite article encodes uniqueness or familiarity, although these may also be taken as presuppositional (Roberts, 2003; Heim, 1982).

Trigger	Example	Presupposition	Non-Presupposing Alternative	Example Alt.
Accomplishment Verbs	x found y	x was looking for y (preparatory activity)	None	N/A
Again	x danced again	x has danced previously	Remove trigger	Avi danced again → Avi danced
Back	x went back to y	x has been in/to y previously	Remove trigger	Avi went back to Canada → Avi went to Canada
It-Clefts	It was x who danced	Someone danced	Active voice variant	It was Avi who danced → Avi danced
Cognitive Factives	x knows that p	p is true	Believe	Avi knows p → Avi believes p
Definite Determiner	the x	x exists	Indefinite Article	Avi saw the play → Avi saw a play
Emotive Factives	Angry that p	p is true	Emotive predicate	Avi is angry that p → Avi is angry
Even	Even x p	p is unusual or notable & x did something other than p	Remove trigger	Avi even danced → Avi danced
Only	Only x danced	x danced	Remove trigger	Only Avi danced → Avi danced
Possessive Pronouns	Their x	They have x (possession)	Indefinite article	Avi saw her cat → Avi saw a cat
Wh-QUestions	Who danced?	Someone danced	Polar Questions	Who danced? → Did someone dance?
State-Change Verbs	x stopped dancing	x was dancing (previous state)	Matrix clause from presupposed content	Avi continued dancing → Avi is dancing
Still	x is still dancing	x had been dancing previously (previous state)	Remove trigger	Avi still dances → Avi still dances
Too	x danced too	Someone other than x danced	Remove trigger	Avi danced too → Avi danced

Table 1: **Presupposition Triggers:** The triggers discussed in this article and their associated presuppositions. The two right-hand columns show simple non-presupposing alternatives and examples.

anaphoric. Thus, the following triggers are predicted to impose CFCs: *again*, *again*, *back*, *even*, *still* and *too*. All other triggers are non-anaphoric and predicted to not impose CFCs.

2.1.2 Focus Presupposition Antecedent Hypothesis:

In two sets of studies, Göbel (2020) notes that focus-associating triggers behave differently than non-focus associating triggers. First, he argues that focus-sensitive triggers are harder to accommodate globally, and second observes that they are sensitive to the salience of the Question Under Discussion (or QUD, Roberts 2012) to which their alternative set provides a (partial) answer. In line with a suggestion from Beaver and Zeevat (2007), he proposes the Focus Presupposition Antecedent Hypothesis (FoPAH) based on these data: "Focus-sensitive presupposition triggers require a linguistic antecedent in the discourse model, whereas triggers lacking Focus-sensitivity merely require their presupposition to be entailed by the [context]." Here, the Discourse Model is a structured representation that keeps track of previous referents and QUDs. It is more difficult to update than the common ground, which is why focus sensitive presuppositions are more difficult to accommodate than their non-focus alternatives. As above, presuppositions are of two different species: non-focus associating presuppositions are essentially satisfaction-based presuppositions, whereas focus-associating presuppositions are (or could easily be) modeled as anaphoric. This theory predicts that the following, focus-based triggers are expected to produce strong CFCs: it-clefts, *even*, *only* and *too*. Other triggers are predicted to produce weak CFCs, or none at all.

2.1.3 Non-Presupposition Alternative Competition Hypothesis:

This approach, developed in Blutner (2000), posits a competition mechanism, where trigger-bearing sentences compete with non-presupposing alternatives. Blutner works within the presuppositions-as-anaphors approach, and formulates his theory within a special variant of Optimality Theory (Prince and Smolensky, 2004) called Bidirectional Optimality Theory, where form/meaning pairs must be optimal both from a comprehension and production perspective. The proposal includes two OT constraints that are implicated in presupposition processing: The first is AVOID ACCOMMODATION (AVOIDA), which assigns a penalty for each time an anaphor is associated with a discourse marker via accommodation. The second is BESTRONG, which evaluates pairs with stronger meanings higher than weaker ones (where strength is determined by entailment). Crucially, AVOIDA is ranked higher than BESTRONG.

To get a sense for how the competition works, consider the variation in acceptability

between the dialogues in (3), and, following Blutner (2000) let's assume for the moment that the two have identical meanings and that the existence of a single salient exhibit is not already in common ground.

(3)   A: What did Avi do yesterday?

    a.   B: She went to an exhibit.

    b. #B: She went to the exhibit.

In this case, the two forms convey the same asserted content and only differ in terms of presupposed content. Thus, both are equally strong as far as assertions go, but (3-b) violates AvoidA, and is therefore less optimal than (3-a). Starting from this simple example, we can derive the fact that, under this approach, when two utterances convey the same message, if they differ only in that one presupposes and the other doesn't, then the presupposing alternative will always be blocked. This leads to *Blutner's Theorem* (Beaver and Zeevat, 2007): If a presupposition trigger has simple expression alternatives that do not presuppose, the trigger does not accommodate.

What predictions does Blutner's Theorem make? As with all competition-based approaches, the predictions of the theory lie in which alternatives we allow to enter the competition. Zeevat (2002) states that the alternatives must be "simple non-triggering expression alternatives with the same meaning" but no formal algorithm for determining alternatives is given. In order to formalize alternative selection, we will adopt a simple approach that constructs alternatives from lexical substitution and deletion. We will also treat negation as a single substitution.[6] Using this simple alternative generating procedure, we can derive simple non-presupposing alternatives for every presupposition except for accomplishment verbs. Simple descriptions of non-presupposing alternatives, as well as examples, are given in the two right-hand columns of Table 1. So while this theory is now more precise, it is now relatively strong. Specifically, all triggers are predicted to impose strong CFCs, except for accomplishment verbs.

2.2 Other Structural Categories

We now turn to two previous categorizations of presupposition triggers which were formulated to explain different behaviors of presupposition triggers. Because neither of these two makes explicit predictions about CFC strength, our data does not provide

---

[6]Otherwise change-of-state verbs, which are traditionally thought to be a single class, would be split: *Continue* would have a simple non-presupposing alternative (*Alex continued to sing*/*Alex sang*) but *stop* would not (*Alex stopped singing*/*Alex **did not** sing*). This is fixed by counting negation as a single substitution.

direct evidence for or against them. Rather, we inspect the way they cut up the presupposition triggers and ask whether their categorization aligns with our empirical results. If so, then the study may provide additional evidence in favor of these theories, and give us a clue as to what causes CFC strength variation.

2.2.1 Soft vs. Hard Triggers

Abusch (2002) distinguishes between soft and hard triggers. She notes that the presuppositions of some triggers (the *soft* ones) can be suspended, whereas the presuppositions of other triggers (*hard* triggers) cannot. (4) gives an example, with the soft trigger *win* in (4-a) and the hard trigger *her* in (4-b).

(4)   a.   I don't know if Ari participated in the race, but if she won it then she will have more titles than anybody else.
      b.   #I don't know if Ari has dog, but if her dog is sick, she'll pay a lot in vet bills.

This distinction has been previously explained as a difference in whether the presupposition is carried by the whole sentence, as in (4-b), or merely by a smaller clausal constituent, in (4-a) (Heim, 1988). The problem for this approach is that it is not clear why triggers like *win* should be associated with the smaller constituent, whereas *her* should not. By way of a solution, Abusch (2002, 2010) proposes that the two presuppositions are derived differently: The presuppositions of hard triggers are *bona fide* presuppositions, which impose semantic constraints on their context (and are interpreted globally), but the presuppositions of soft triggers are the result of pragmatic reasoning from lexical alternatives. For example, *win* exists in a lexically-based alternative set with the verb *lose*. When a comprehender hears the utterance "Ari won the race", its alternative "Ari lost the race" is activated. Because, "[t]ypically, some alternative in a topical alternative set is assumed to be true" and in this case both alternatives convey that Ari participated in the race, the whole sentence conveys that *Ari participated in the race* is true. (A similar approach using *scalar* alternatives is offered in Romoli and Schwarz (2015).) The reason why such a conclusion can be suspended is that it is a merely a pragmatic inference, and not a hard-and-fast semantic constraint.[7] Looking at the list in Table 1 our soft triggers include aspectual verbs and

---

[7]There are a number of challenges to the soft/hard distinction, however. Empirically, it has been found that presuppositions and implicatures are processed at different speeds (Bott and Noveck, 2004; Bott et al., 2012; Schwarz, 2014; Chemla and Bott, 2013; Romoli and Schwarz, 2015; Bill et al., 2015) leading some experimentalists to endorse the "perspective of seeing the two phenomena as distinct in nature" (Bill et al. (2015), p.1). Theoretically, Abrusán (2016) argues that the cancellation differences

achievement verbs and wh-questions.[8]

2.2.2 Weak vs. Strong Triggers

Glanzberg (2005) argues that a single theoretical mechanism can account for all presuppositional phenomena, but that the particulars of how the presuppositional content relates to the asserted content of a trigger results in two distinct categories of presuppositions. Consider the following pair of sentences:

(5)   Alex read *Crime and Punishment.*
    a.   In fact, she read *Daemons*, too.
    b.   In fact, she couldn't put the novel down.

Glanzberg argues that in (5-a), that *too* requires a comprehender to check that some alternative to *Daemons* is in the context, but to do nothing more with it. On the other hand, when interpreting (5-b), the comprehender must not only check that a referent of *the novel* is in the context, but must additionally involve that referent in predication. Primarily interested in how expressions can fail to make meaning, Glanzberg argues that presuppositions like *the* induce obligatory discourse repair whereas presuppositions like *too* only induce optional repair. That is, if someone uttered (5-b) in a context where no referent existed, conversation would not continue as normal, and would enter a special state where participants try to determine what went wrong and why. Not so for (5-a).

Building on this insight, Domaneschi et al. (2014) make the distinction between *weak* vs. *strong* triggers, where strong triggers require discourse repair and weak triggers do not. Tiemann et al. (2015) further suggest that the relevant facts for strong vs. weak are determined by the semantic role of the presupposition trigger. Triggers that change semantic type of their arguments cannot be ignored and are strong, whereas semantic adjuncts can be ignored and are weak. The intuition is that if a trigger can simply be removed from a sentence without rendering the rest of the sentence meaningless, then it is a weak trigger, whereas triggers whose deletion would render the whole sentence ungrammatical are strong triggers. To go back to our earlier example, because the sole role of *too* is to introduce a presupposition, most of the time it can be removed without causing the rest of the sentence to loose its non-presuppositional meaning. Thus, *too* is

---

can be explained by local information structural factors, without resorting to two distinct species of presuppositions.

[8]Although not traditionally categorized as such, emotive factives could also be thought of as soft triggers. For example, "angry that p" could exist with alternatives "sad that p", "happy that p" etc. all of which presuppose p.

a weak trigger. Tiemann et al. present a self-paced reading time study for the German presupposition *wieder* ('again') with offline comprehension questions supporting this hypothesis. However, the weak vs. strong distinction has been critiqued on empirical grounds, most notably by Bacovcin et al. (2018). Looking at the items in Table 1, weak triggers include *too*, *even still*, *back*, *again* and *only*. All the other triggers are strong.

2.3 Summarization of Predictions

Table 2 summarizes the predictions of the various theories discussed in this section. For the hard/soft and weak/strong distinctions, we simply note which triggers fall into which categories. In the next section, we introduce an experiment designed to test these predictions.

3 Experimental Methods

Numerous experimental methodologies have been used to investigate presuppositions and their contextual felicity. Previous studies have used online measures, such as eye-tracking and self-paced reading (Schwarz, 2007; Schwarz and Tiemann, 2017; Tiemann et al., 2011; Schwarz and Tiemann, 2017; Singh et al., 2016), speeded true/false judgements (Chemla and Bott, 2013; Romoli and Schwarz, 2015), inference judgements (Chemla, 2009; Chemla and Schlenker, 2012), judgements of speaker certainty (Tonhauser et al., 2018), comprehension or recall questions (Domaneschi et al., 2014; Tiemann et al., 2011; Bacovcin et al., 2018), as well as acceptability judgements (Cummins et al., 2012, 2013; Amaral and Cummins, 2015; Göbel, 2020; Bade and Renans, 2021).

For this study, we elect to use acceptability judgements, which are best suited to our main theoretical objectives. First, acceptability is the measure most directly linked to felicity/infelicity, which is our theoretical property of interest. Second, acceptability judgements are relatively intuitive and require little training for naive participants to produce. Third, acceptability judgements have been used successfully to make comparisons across a wide range of triggers: For example both Cummins et al. (2013) and Göbel (2020) use acceptability to compare 8 different presupposition triggers. Finally, because acceptability judgements are a relatively simple offline measure they have been previously deployed to study contexts where presupposition triggers are embedded in complex semantic environments (Bade, 2016), as we do here. While issues of contextual felicity and accommodation could be approached through multiple

Trigger	Hard/Soft Distinction (Abusch, 2002)	Weak/Strong Distinction (Glanzberg, 2005)	Non-PSP Competition (Blutner, 2000)	FoPAH (Göbel, 2020)	Anaphor / Constraint (Kripke, 2009)
Accomplishment Verbs	Soft	Strong	Accom	Potential Accom	Accom
Again	Hard	Weak	No Accom	Potential Accom	No Accom
Back	Hard	Weak	No Accom	Potential Accom	No Accom
It-Clefts	Hard	Strong	No Accom	No Accom	Accom
Cognitive Factives	Hard	Strong	No Accom	Potential Accom	Accom
Definite Det.	Hard	Strong	No Accom	Potential Accom	No Accom
Emotive Factives	Hard	Strong	No Accom	Potential Accom	Accom
Even	Hard	Weak	No Accom	No Accom	No Accom
Only	Hard	Weak	No Accom	No Accom	Accom
Possessive Pronouns	Hard	Strong	No Accom	Potential Accom	Accom
Wh-Questions	Soft	Strong	No Accom	Potential Accom	Accom
State Change Verbs	Soft	Strong	No Accom	Potential Accom	Accom
Still	Hard	Weak	No Accom	Potential Accom	No Accom
Too	Hard	Weak	No Accom	No Accom	No Accom

Table 2: Predictions of various theoretical proposals.

experimental paradigms, acceptability ratings are simple, and provide room for flexible item creation that can target different syntactic structures and semantic environments.

3.1 Design

To assess the strength of Contextual Felicity Constraints, a 2x2 experimental design was employed testing acceptability of a sentence that either contained a presupposition trigger or not (+TRIGGER vs. −TRIGGER) and in which the immediate preceding context either supports the presupposition or not (+SUPPORTING vs. −SUPPORTING).[9] A context is taken to "support" a presupposition if it either entails the content of the presupposition or if it provides the trigger with a possible discourse referent. Example (6) gives a sample for the trigger *even* in each of the four possible conditions, with the context sentence on the left and the target sentence underlined. More information about construction of the materials is given in the paragraphs below.

(6) a. What did Josh do today? He went to the grocery store.
        [−SUPPORTING, −TRIGGER]
     b. What did Josh do today? He even went to the grocery store.
        [−SUPPORTING, +TRIGGER]
     c. Josh went all over town today. He went to the grocery store.
        [+SUPPORTING, −TRIGGER]
     d. Josh went all over town today. He even went to the grocery store.
        [+SUPPORTING, +TRIGGER]

The logic of the design is as follows: If a trigger imposes a Contextual Felicity Constraint, then by definition a trigger-bearing sentence should be more acceptable in a context where its presupposition is supported than in a neutral context where it is not supported. Thus, (d) should be rated as more acceptable than (b). In addition, if a trigger imposes a CFC, then in a non−supporting context, a trigger-bearing sentence should be less acceptable than a minimal-pair sentence that does not contain a presupposition trigger. Thus, we expect (a) to be rated as more acceptable than (b). Each of these two contrasts has been deployed in previous experimental setups for testing CFC strength: Tonhauser et al. (2013) investigates the (d) vs. (b) contrast, which will be referred to as the *+trigger* contrast. Additionally, Göbel (2020) investigates the (a) vs. (b) contrast, which will be referred to as the *-supporting* contrast.

One problem with each of these previous studies is that looking at binary contrasts may lead to incorrect conclusions about CFC presence or CFC strength. For exam-

---

[9]These are the same as what Tonhauser et al. (2013) call NEUTRAL (here, our SUPPORTING variable).

What did Josh do today?

He went to the hardware store.   Least acceptable ———•——— Most acceptable

He even went to the hardware store.   Least acceptable ———•——— Most acceptable

→ Click here to continue

Figure 1: Sample item for the *even* trigger in the −*supporting* condition.

ple, it may be the case that participants rate *+trigger* sentences worse than −*trigger* sentences in −*supporting* contexts, producing a large *-supporting* contrast. However, without checking how *+trigger/+supporting* sentences are rated, we cannot rule out the possibility that participants merely give low ratings to sentences with the trigger across the board (which we do find in our data). Thus, in order to rule out potential confounds, we use a *conjunctive criterion* when assessing CFC presence. That is, in order to be confident that a trigger is imposing a CFC we look for a significant *+trigger* contrast and a significant *-supporting* contrast. If both are present, then we conclude the trigger is subject to a Contextual Felicity Constraint.[10]

The study employed the presentational design advocated in Marty et al. (2020), who report that joint presentation of conditions with a continuous scale and labeled endpoints draw out robust contrasts between conditions in a rating task of this type. There are two advantages worth highlighting about this experimental paradigm: First, it draws out robust contrasts because it allows for direct comparison between conditions

---

[10]Given our experimental design, one other option is possible, and that would be to look for a significant interaction between trigger and support. The problem with this measurement for CFC strength is that it runs into issues with trigger obligatoriness in *+supporting* conditions. Trigger obligatoriness is the phenomena where sentences sound less acceptable when they *don't* include a presupposition trigger, for example "In the sky, a sun is shining" (Heim, 1991). If a trigger is obligatory, then *+supporting/−trigger* sentences like (c), above, could be rated lower, resulting in deviant interaction effects. For example, if participants rate sentences (a) and (b) equivalently, but (c) as lower than (d), an interaction analysis would indicate that the trigger is subject to a Contextual Felicity Constraint. However, this would be entirely due to pressures imposed on the trigger when its presuppositions are supported. The question of trigger obligatoriness is undoubtedly related to contextual felicity, however such pressures are moot if participants find no differences between *+/−trigger* sentences in −*supporting* contexts. Thus, in order to avoid these pitfalls, we avoid interaction tests and stick to conjunctive criteria for *-supporting* and *+trigger* contrasts.

One potential concern with our paradigm is that *+supporting, -trigger* condition is not used in the computation of our conjunctive criteria. Originally, we were interested in running an interaction analysis similar to the one offered in (Wilcox et al., 2021), however we eventually decided against it for the reasons mentioned above. We believe that the *+supporting, -trigger* condition is still useful for our side-by-side presentational paradigm.

on a single screen, enabling participants to report small judgement differences even if judgments might cluster together amid a wider context of possible ratings. Second, it highlights the aspect of the judgement which the experimenter intends the participant to focus on. These advantages come at the expense of participant naivety—by situating both conditions on a single screen the experimenter draws back the curtain to reveal which aspects of the sentence should be most important to the judgement.

For each trial participants were shown the context, in bold, at the top of the screen, and asked to rate the two possible continuations ($+trigger$ and $-trigger$), which were presented below in a random order with continuous response bars at right. The slider bar responses were stored as an integer from 0-100, with 0 being "least acceptable" and 100 being "most acceptable". Figure 1 gives an example for the trigger *even*, in a $-supporting$ context. At the beginning of the experiment participants were instructed to think about acceptability as how well the sentence fits with the preceding context, following the instructions given in Göbel (2020). After the instructions, participants were given three warm-up trials, two of which involved a grammatical number mismatch between the context and one of the target sentences, to give them an example of a sentence that clashes with its context. Additionally, eight attention check items were dispersed randomly throughout the experiment, which contained a mismatch between a grammatically gendered noun in the context and a pronoun in the target (e.g. Context: "Yesterday a nun visited our school." Target: "We heard {her/him} speak in our class on world religion").

249 participants were recruited on Prolific, who were all self-identified native English speakers with IP addresses inside the United States. Because this experiment involves a large number of items, it was divided it into six sub-experiments, which took about 20 minutes to complete. Subjects were not allowed to participate in more than one sub-experiment and were excluded if their responses for the attention check items in the *match* condition were not in the top quartile, or *mismatch* not in the bottom quartile of the slider, on average. In total, we filtered 40 participants, or ∼16% of the total, indicating that participants were generally using the slider as intended.

3.2 Materials

Items were created for fourteen English presupposition triggers given in Table 1. The following standards were used when creating experimental items: Each context sentence introduced a character, and the target sentence provided further information about the character's recent activities. Neutral contexts were constructed using wh-questions, which are associated with speaker ignorance. Positive contexts were constructed with

simple past-tense statements that satisfied the target trigger's presuppositions. Characters were introduced using first names familiar to English readers. When noun phrases were repeated between the context and target sentence they were turned into pronouns, if the change was judged to increase semantic felicity. *+Trigger* target items consisted of simple past-tense statements that included the presupposition; *−trigger* items were created using simple non-presupposing alternatives.

As the goal of this experiment is to test CFC robustness in a variety of contexts, items were created in five semantic environments, with six items per environment for a total of thirty items per trigger. While our analysis will generally focus on the average behavior of the trigger across all environments, we will discuss the effect of semantic environment on CFC strength in Section 4.4. Below, (7)-(11) give examples for presupposition trigger *again* in the critical *−trigger/−supporting* condition for each of the five environments tested. These are intended to just give an overview for how items changed by environment.

The first environment tested presuppositions in matrix clauses:

(7)   What did Alex do over the weekend?
        She went to the beach again.   [MATRIX]

The second environment tested presuppositions under the scope of negation. Negated sentences are often judged to be degraded answers to simple wh-questions because they do not provide exhaustive or maximal information, which may be necessary under some semantic theories of questions (Dayal, 2016). To avoid these sorts of question/answer pairs, we modified the *−supporting* context to better set up expectations for negation by using *why not* questions. An example is given in (8) below.

(8)   Why didn't Alex get sunburned this summer?
        She didn't go to the beach again.   [NEGATION]

The third environment tested triggers embedded under possibility operators. A mix of possibility modals including *might have*, *maybe* and *it's possible* were used. (9) gives an example.

(9)   What did Alex do this summer?
        She might have gone to the beach again.   [POSSIBILITY]

The fourth environment tested presupposition triggers embedded in the antecedents of conditionals. Two important considerations constrained item creation: First, the consequent of the conditional had to answer the question in the *−supporting* context,

and second it had to follow from the antecedent. To accomplish this, questions and consequents were framed around a target character's mood. (10) gives an example.

(10)   How is Alex feeling?
        If she went to the beach again this weekend, she'll be in a good mood. [CON-DITIONALS]

The last environment tested presuppositions in polar questions. Here, we changed the −*supporting* context from a wh-question to a simple stative sentence that did not entail the presupposition. (11) gives an example.

(11)   Alex had a great summer.
        Did she go to the beach again?   [QUESTIONS]

Examples of each trigger, condition, environment pair can be found in the supplementary materials.

4   Results

4.1   Overall Results

The results from the study are shown in Figure 2. For now, we will discuss trigger behavior averaging across semantic environments, however we will turn to the question of trigger/environment interaction in Section 4.4, below. Results broken down for each trigger/environment pair can be found in Appendix A. Comparing across triggers, we find three types of patterns. The first are cases where there is (visually) no interaction between the two experimental conditions: +*trigger* conditions are rated more highly than or equal to −*trigger* conditions, regardless of support. This category includes possessive pronouns and factive predicates. The second are cases of spreading interactions, where there appears to be a main effect of +/− *trigger* that is enhanced in the −*supporting* context. Triggers with spreading interactions include clefts, *only* and *even*. More common are cases of cross-over interactions, where the relative felicity of the +/− *trigger* targets are reversed between the −*supporting* and +*supporting* contexts.

In order to provide statistical assessment for which triggers are subject to a Contextual Felicity Constraint, we fit pairs of linear mixed effects models to test the -*supporting* and +*trigger* contrasts discussed previously. Models were fit using the following methods: All models had participant rating as the response variable and used
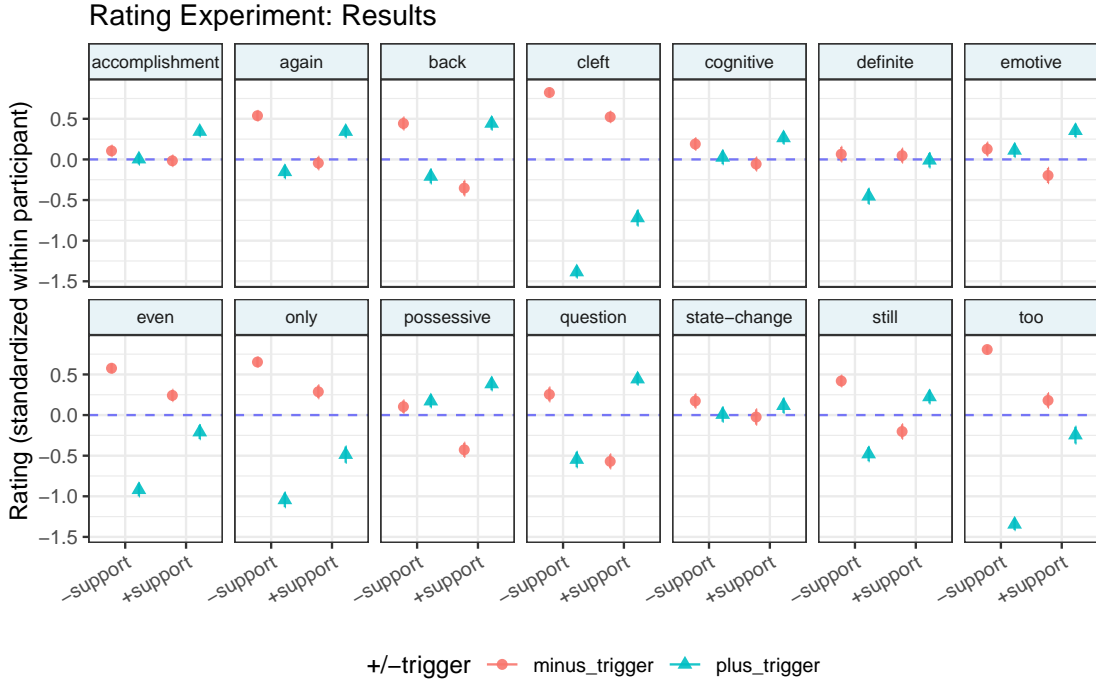
Figure 2: **Results from the rating study:** Triggers are arranged in alphabetical order starting from the top left. The x-axis indicates whether or not the context supports the presupposition or not and the y-axis indicates participants' ratings, which have been standardized (i.e. z-scored) for each participant to control for cross-subject variation. Error bars show 95% confidence intervals; red plots are $-trigger$ ratings and blue plots are $+trigger$ ratings.

the single crucial contrast as the sole predictor after filtering out data that was not relevant for the metric. (For example, for the $+trigger$ contrast data from the $-trigger$ condition was filtered out.) We included by-participant and by-item random slopes for these models, and report the significance value of the sole predictor variable.[11] If both contrasts are found to be significant, then we conclude that the trigger is subject to a CFC.

Significant *-supporting* contrasts were found for *again*, *back*, clefts, the definite determiner, *even*, *only*, questions, *still* and *too* (all $p < 0.001$), but not for accomplishment verbs, cognitive or emotive factives, possessive pronouns and state-change verbs. Significant $+trigger$ contrasts were found for all triggers, except state-change verbs ($p < 0.001$ for accomplishment verbs, *again*, *back*, clefts, the definite determiner, *even*, *only*, questions, *still* and *too*; $p < 0.01$ for emotive factives and possessive pro-

---

[11]An example lmer formula for the $+trigger$ contrast: `rating ∼ supporting + (supporting | participant) + (supporting | item)`
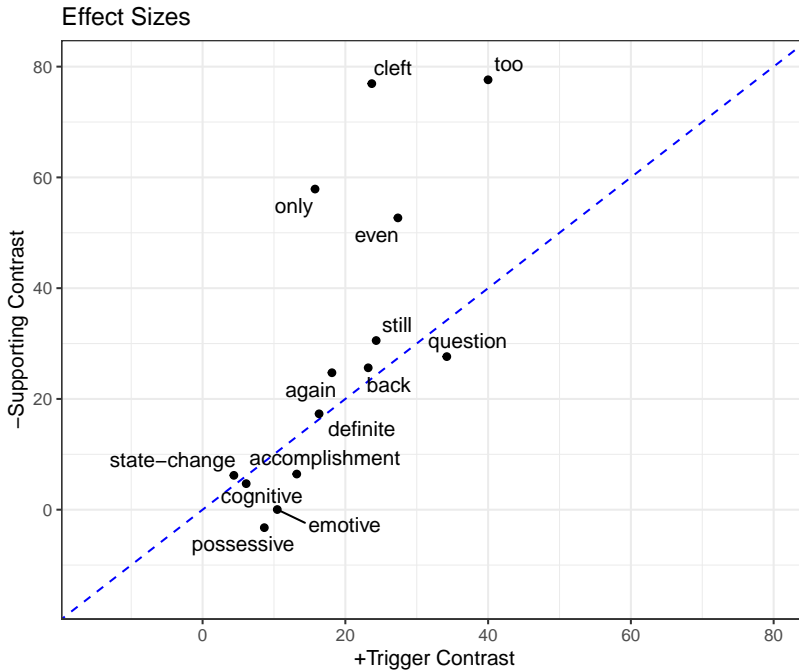
Figure 3: **CFC Effect Sizes:** Points represent mean contrasts, averaged across items and semantic environments.

nouns; and $p < 0.05$ for cognitive factives). These results indicate that the majority of triggers tested are subject to a CFC, however factive predicates, possessive pronouns, state-change and accomplishment verbs are not.

4.2 Effect Sizes

In addition to knowing whether a trigger is subject to a CFC, we might want to know the relative strength that the CFC imposes. CFC strength could be estimated by looking at the size of differences between conditions. In order to get a sense of agreement between the *+trigger* and *-supporting* contrasts, Figure 3 shows effect sizes, with the *-supporting* contrast on the y-axis and *+trigger* contrast on the x-axis. Effect sizes were computed by taking the relevant differences between conditions after averaging across trials, participants and semantic environments for each trigger. Based on their proximity to the $y = x$ line (which is shown in dotted blue), its clear that both metrics agree for the majority of triggers, with the exception of the focus-sensitive operators—clefts, *too*, *even* and *only*—which were found to have larger *-supporting* contrasts than *+trigger* contrasts.

We compare these effect sizes against those from Wilcox et al. (2021), who used
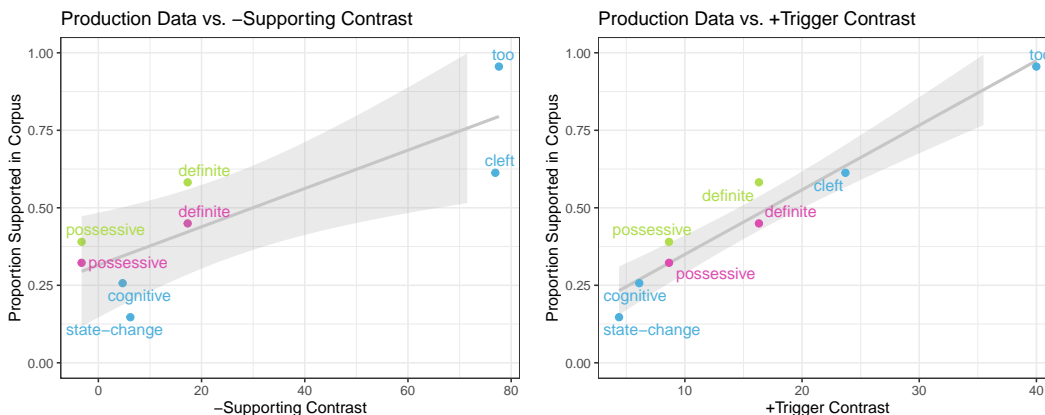
Figure 4: **Comparison to Production Data:** Production data is from Spenader (2002). Color-coding of triggers in the figure corresponds to three different annotators: One annotator for *too*, factives, clefts and change of state verbs; and two annotators each for possessive pronouns and the definite determiner.

the same experimental setup and tested similar triggers as as we do here[12], but only in the *matrix* environment. Our results are highly correlated to those from this previous study by a Pearson correlation test for both *-supporting* contrasts ($\rho = 0.86, p < 0.001$) and *+trigger* contrasts ($\rho = 0.75, p < 0.01$).

Based on these data, three clusters of triggers emerge. The first cluster contains all the triggers which were found not to be associated with CFCs in our first analysis: state-change verbs, accomplishment verbs, factive predicates and possessive pronouns. These are associated with small, or in some cases a total lack of, CFC effects. The second cluster contains triggers associated with moderate CFCs, including *again*, *still*, *back*, questions and the definite determiner. The final cluster contains the aforementioned focus-sensitive particles. Looking at the *-supporting* contrasts for these items, we might conclude that they are associated with very strong CFCs. However, *+trigger* contrasts tell a different, and more complicated story: Here, *too* and *even* are associated with the strongest effect sizes, but cleft structures and *only* pattern more closely with the intermediate cluster of triggers. We will return to this point in the discussion.

4.3 Comparison to Production Data

As with any online comprehension study, there may be questions about the ecological validity of the experimental paradigm for capturing naturalistic behavior. In this section, we validate the methods against production data from Spenader (2002), who

---

[12]Wilcox et al. (2021) combine cognitive and emotive factives into a single category. For the sake of comparison, we average between these two triggers when computing correlations.
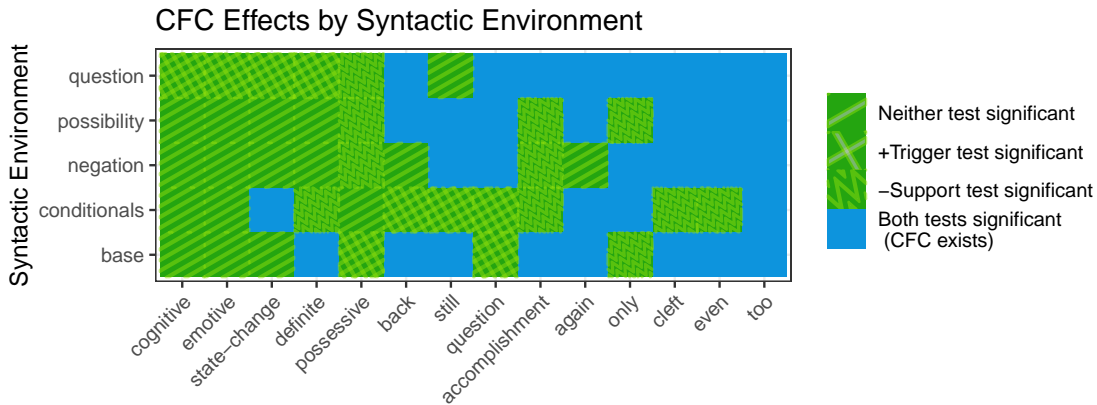
Figure 5: **CFCs by semantic environment:** Blue indicates that statistical tests for the conjunctive criteria were significant (evidence of a CFC). Green indicates that the conjunctive criteria was not met (no evidence for a CFC). Fill pattern details which tests were significant. Triggers are ordered on the x-axis based on the number of environments for which CFC effects were found. Trigger behavior is relatively stable across environments, with the exception of conditionals.

collected data from the London-Lund Corpus of Spoken English, and hand coded them as to whether each trigger's presuppositions were supported in the preceding context.[13] Data was collected for only a subset of the triggers tested in our study: possessive pronouns, factive predicates, the definite determiner, change of state verbs, clefts, and *too*. The assumption is that if a trigger imposes strong Contextual Felicity Constraints, then it will be costly for speakers to use and listeners to interpret in cases where its presuppositions are not supported by the context. Speakers would be expected to avoid such costly uses and thus we predict a correlation between the proportion of supported use in the production data and the strength of the CFC, as measured in our study.

The comparison between production data and the effect-sizes of the two metrics can be seen in Figure 4, with the proportion of support on the y-axis, and the results of our study on the x-axis. Overall, the results show strong correlation between the strength of the CFC, as measured in our experiment, and the proportion of times a presupposition is used with contextual support in production as measured in Spenader (2002) (by a Pearson correlation test). For the *+trigger* contrast we find $\rho = 0.82$ ($p < 0.05$); for the *-supporting* contrast we find (an astonishing) $\rho = 0.97$ ($p < 0.001$).

	Base	Conditionals	Negation	Possibility
Base	1			
Conditionals	0.58	1		
Negation	0.60	0.25	1	
Possibility	0.82	0.44	0.73	1

Table 3: **Between-Environment Correlation:** Correlation of trigger behavior between different semantic environments, computed using the matrix described by Figure 5 where green corresponds to 1 and blue to 0.

4.4 Environment-by-environment Breakdown

So far, we have been looking at the behavior of each trigger averaged across five semantic environments. Now, we explore the effect of environment on the triggers' ratings, to get a sense of the robustness of CFC effects. Rating results for each condition/environment pair can be seen in Appendix A, with examples items for each trigger/environment/condition in the supplementary material. However, here we will focus on the results of our statistical tests, shown in Table 5, in order draw out patterns of trigger behavior.[14] If both *-supporting* and *+trigger* contrasts are significant (i.e. the conjunctive criteria is met), we conclude that the trigger imposes a CFC and the cell is colored blue. If one or both tests is not significant then the cell is colored a shade of green, with the fill pattern detailing which tests were found to be significant. This summary table is intended to give a sense of CFC stability across environments.

The most important takeaway from these summary visualizations is that there is a good agreement between trigger stability and effect size of the CFC as reported in Figure 3. Possessive pronouns, factive predicates and state change verbs are associated with the fewest CFCs across environments, whereas triggers which were associated with large effect sizes are found to produce significant CFCs across multiple environments (i.e. *too*, *even* and clefts).

But does trigger behavior change based on environments? In order to quantify this, we compute the correlation matrix for the rows of Figure 5. To do this, we assign a cell 0 if it is associated with a significant CFC effect (i.e. it is colored blue), and 1 otherwise. The correlations between the rows of this matrix are shown in Table 3. These correlations suggest robust homogeneity across semantic environments of CFCs for each trigger: All correlations are positive, and the average correlation between se-

---

[13]Data is take from Table 5 and Table 6 of Spenader (2002), section 5.1

[14]Statistical methods are the same as those described in Section 4.1, except statistics are computed on trigger/environment subgroups, instead of a trigger's ratings across all environments.

mantic environments is 0.57. However, these numbers do reveal one difference between environments, which is that CFCs tend to pattern differently when triggers are placed in antecedents of conditionals. The correlations between Base, Negation and Possibility environments range from 0.6-0.82, whereas the correlations between Conditionals and other environments are much lower, ranging from 0.25 - 0.58. The unique behavior of triggers in conditionals is visually evident in Figure 5, where the conditionals row has more green-shaded cells, indicating a lack of CFC effects.

4.5   Discussion

There are two takeaways that we believe it is important to recapitulate at the outset: First, we observe a great deal of variation between the triggers, suggesting that different triggers are subject to different types of CFCs, and some are not subject to CFCs at all. Second, these results indicate that the behavior of triggers can be thought of as clustered into focus sensitive operators (which we will discuss in the next paragraphs), triggers that do not impose CFCs (possessive pronouns, state-change verbs, accomplishment verbs and factive predicates), and triggers that impose moderate CFCs (*again*, *back*, *still*, definite determiners and questions). Crucially, semantically-similar triggers pattern together. So, for example, all of our non-focus additive particles are in the middle cluster, and all of our lexical verbs are in the weak-CFC cluster. This clustering is another reason to conclude that formal semantic properties are implicated in CFC strength.

Turning now to some of the outstanding questions, what should we make about the proportionally large *-supporting* contrasts for focus-sensitive items, that drive them above the $x = y$ line in Figure 3? First, this phenomena is somewhat less of an issue for *too* and *even*. Even though their *-supporting* contrasts are much larger than their *+trigger* contrasts, both metrics produce similar relative rankings (that is, they are both associated with some of the strongest CFCs). The over-estimation of *-supporting* contrast, however is more problematic for *only* and clefts. Going off *-supporting* contrasts, one would be tempted to conclude that they pattern with *too* and *even*. However, looking at *+trigger* contrasts, they pattern closer with the middle-ground triggers, like *again* and the definite determiner. Given that, in production data, clefts were found to be used with support only about 60% of the time, whereas *too* was found to be used with support more than 95% of the time, it seems inappropriate to clump these two triggers together in the same category. Furthermore, Tonhauser et al. (2013) argue that *only* is not associated with a CFC, and present data from both English and Paraguayan Guarani that suggests it can be used to felicitously introduce new information. Taking

this larger suite of empirical evidence into account, we believe that it is appropriate to conclude that *too* and quite likely *even* produce strong CFCs, but that *only* and cleft structures are associated with mild or weak CFCs.[15]

We turn, now, to a second perplexing pattern in the data, which is the relative lack of CFC effects in conditionals. We propose two possible explanations: First, it may be the case that sentences with conditionals are more difficult to process. It is well established, for example, that certain semantic operators produce processing difficulties when presented with only minimal context like the items in this experiment (Carpenter and Just, 1975; Kaup and Dudschig, 2020). Because of this, it may be the case that when assessing semantic acceptability for complex sentences, participants devote resources to the relationship between the asserted content and the truth-canceling operators, and fail to factor in contextual felicity when making their judgement. However, under this type of account, it would need to be explained why conditionals cause more processing difficulty than, say, negation, which is more strongly associated with processing costs in the literature.

Alternatively, the reduction in effects could be due to discourse-structural properties, which stem from the questions used to set up $-supporting$ environments. For matrix and possibility environments $-supporting$ contexts were introduced with "What did X do?" questions. For conditionals, the most common question used was "How is X feeling?", and the structure of the target sentence was something like "If p then q" where p is sentence describing an activity that bears a presupposition and q is a sentence about their emotional state or mood. It may be the case that participants perceive the antecedent as less related to the topic of the discourse (X's mood), and are therefore more willing to accommodate or ignore presuppositions in this structural location. This hypothesis assumes that CFCs are related to information structure of the sentence relative to a local question or question under discussion, which is one leading

---

[15]One reason why *-supporting* contrasts may over-estimate the CFCs for the focus-sensitive exclusives, *only* and clefts, is that these triggers may be associated with strict Question/Answer congruence conditions. For example, the Q/A pair: "Who did Amos talk to? He only talked with Zack" may appear infelicitous, but not necessarily because the presupposition trigger imposes a CFC. In this case, use of the form *only* is blocked by the presuppositionless alternative ("He talked to Zack"), which conveys the same content if it is exhaustified with respect to the question. One reason to think that it is Q/A congruence and not CFC strength that is driving lower ratings are cases like (12) below, in which the unsupported use of *only* is acceptable (author judgement). In this case, the presuppositionless alternative may not successfully convey the exhaustive meaning, and the variant with *only* becomes available, even though its presuppositions are not supported by the context.

(12)   *Two people are at a party with their friend Amos, who is a social butterfly. Person B knows that Amos talked to Zack and nobody else at the party.*
        A: Amos must have talked to a lot of people. Who did he talk to?
        B: He talked only with Zack.

hypothesis for their projective properties (Simons et al., 2010; Tonhauser et al., 2018). These data suggest that local informational structural properties may (also) be at play when it comes to contextual felicity.

5 General Discussion

5.1 Evaluating Predictions of the Various Theories

How well do the various theories discussed in Section 2 explain the experimental results? Table 4 gives an overview of the predictions of each theory for our triggers tested. Triggers are ordered roughly based on the strength of their *+trigger* effect (i.e. their placement on the x-axis in Figure 3), and lines separate the different clusters. We cluster clefts and *only* together, and place them under the medium-CFC triggers, based on the discussion above. Furthermore, we cluster wh-questions together with *too* and *even*—they were found to have some of the strongest CFC effects and, like the focus sensitive operators, their semantics is taken to involve operating over sets of focus-based alternatives.

For each theory, we'll inspect the way that it cuts up the presupposition triggers, and ask whether its categorization aligns with our empirical results. Starting with the soft/hard distinction, we find some good overlap between the relevant categories and our results: Of the three soft triggers tested, the two open-class verbal items impose weak or no CFCs. Wh-questions, however, impose strong CFCs. While this does split up soft triggers into different CFC categories, there might be a potential explanation for this: Presuppositions associated with state change and accomplishment verbs are said to be triggered by implicit alternative reasoning (i.e. over alternatives or scales Romoli 2015), whereas wh-questions work with focus-based alternatives and introduce them explicitly in their semantics. Thus, accommodation of these items could be affected by the way in which those alternatives are generated. Furthermore, because the notion of alternative-based reasoning was developed to explain the cancellation properties of verbs like *stop* and *win*, it makes sense that their CFCs would be weak and easy to suspend. Thus, the soft/hard distinction presents a theoretically-grounded story for CFC variation that is compatible with our data.

The one problem for adopting soft/hard distinctions as a main locus of CFC variation is that not all of the triggers that fail to impose CFCs are soft. Possessive pronouns, for example, are quintessential hard triggers, and yet they were found to impose no, or very minimal CFCs. Thus, while the soft/hard distinction may be able to explain why some triggers impose weaker CFCs, it leaves important questions about

	Trigger	Hard/Soft (Abusch, 2002)	Weak/Strong (Glanzberg, 2005)	Competition (Blutner, 2000)	FoPAH (Göbel, 2020)	Anaphor (Kripke, 2009)
STRONG CFC	Too	Hard	Weak	✗	✗	✗
	Even	Hard	Weak	✗	✗	✗
	Wh-Qs	Soft	Strong	✗	✓	✓
MEDIUM CFC	Still	Hard	Weak	✗	✓	✗
	Back	Hard	Weak	✗	✓	✗
	Again	Hard	Weak	✗	✓	✗
	Definite Det.	Hard	Strong	✗	✓	✗
	It-Clefts	Hard	Strong	✗	✗	✓
	Only	Hard	Weak	✗	✗	✓
WEAK CFC	Accomp. Verbs	Soft	Strong	✓	✓	✓
	Emotive Factives	Hard	Strong	✗	✓	✓
	Possessive Pronouns	Hard	Strong	✗	✓	✓
	Cognitive Factives	Hard	Strong	✗	✓	✓
	State Change Verbs	Soft	Strong	✗	✓	✓

Table 4: Predictions of various theoretical proposals with the results of our study, which are ordered from strongest CFC triggers (top) to weakest CFC triggers (bottom). ✗ means the theory predicts no accommodation (i.e. a strong CFC); ✓ means the theory predicts potential accommodation (i.e. a weaker CFC).

CFC variation unanswered.

Turning now to the weak/strong approach, we again find relatively good fit between our data and the way this theory carves up the triggers. All the triggers associated with minimal CFCs are all strong triggers whereas the more robust CFCs are all associated with weak triggers. The only exception to this trend are wh-questions, which are strong triggers à la Glanzberg (2005), but demonstrate more robust CFCs. However, as discussed above, questions are a bit of a special case. They are associated with complex syntactic operations and may trigger their presuppositions via explicit introduction of alternatives into the semantic derivation. So, granting questions as the possible exception, the weak/strong approach does seem to separate triggers in a way that is consistent with CFC effects.

That being said, it is not immediately clear, at least within the theory proposed by Glanzberg (2005), *why* this should be the case. The most likely link between the weak/strong hypothesis and CFC variation is to assume that weak triggers, which require only optional discourse repair, are easier to accommodate. But this is precisely the opposite of what we find. On the assumption that CFC strength measures ease of accommodation, we get that weak triggers are *difficult* to accommodate, and strong triggers are *easy* to accommodate. Thus, while this approach is predictive, its explanatory potential is difficult to interpret. We will return to this puzzle in the General Discussion, after examining the results of the three theories that make explicit predictions about accommodation.

The first of these two theories is the Focus Presupposition Antecedent Hypothesis (FoPAH), which predicts that focus-sensitive triggers should be difficult to accommodate. The FoPAH has found previous empirical support in Göbel (2020), who used an experimental setup similar to our *-supporting* contrast to test CFC strength.[16] As mentioned in the discussion of the previous section, if only *-supporting* contrasts were to be used, then the results would support the FoPAH, however the experiment deployed a broader set of criteria to determine CFC strength, including the *+trigger* contrast and a conjunctive criteria between the two. As discussed above, these results suggest that *-supporting* contrasts can over-estimate CFC effects for focus-associating triggers, which may be more sensitive to question/answer congruence (Abrusán, 2016). Looking at CFC effects using the *+trigger* contrast, as well as comparison with production data, we argued that focus sensitivity does not necessarily result in larger CFC effects, and group the exclusive triggers (*only* and clefts) in the middle of our CFC ranking.

---

[16]Although there are similarities, this study was not a replication of Göbel. He is primarily interested in investigating differences between paired focus/non-focus triggers, and discusses things in terms of accommodation, rather than our "CFC."

That being said, focus sensitivity is clearly important for presuppositional phenomena, and we do find that, among the additive particles, focus-sensitive triggers tend to produce larger and more stable CFC effects than non-focus triggers. One possible reason for this is that because of necessary focus/question congruence (Roberts, 2012), the presuppositions of focus sensitive triggers will always be at-issue with respect to the question that sets up -*supporting* conditions. As we suggested above when discussing the reduction of CFC effects in the antecedents of conditionals, participants may be sensitive to whether a presupposition is at-issue or not. Under the assumption that people are supposed to answer at-issue material with asserted content, and because the presuppositions of *too* and *even* will always be at issue, they might be predicted to impose CFCs in a high proportion of contexts.

Next, we turn to the anaphoric vs. satisfaction hypothesis. On the assumption that all of our additive and iterative particles are anaphoric, we observe good alignment between this theory and our results. The two places where this proposal makes a wrong predictions are with wh-questions (once again) and the definite determiner, which we categorized as imposing a moderate CFC. There is, however, a substantial body of literature that treats the definite determiner as anaphoric; indeed, it was problems with satisfaction-based analysis of *the*'s presuppositions that originally motivated the anaphoric approach (see, e.g. the discussion around examples (4) - (7) in van1992presupposition). So it seems possible to readjust the analysis by grouping *the* with other anaphoric triggers.

Finally, we turn to the Non-Presupposing Alternatives Proposal, which says that failure to accommodate is the result of a competition mechanism between presupposing and non-presupposing sentence variants. As it was formulated in Blutner (2000), this proposal is not compatible with our data, predicting only that accomplishment verbs should be accommodated and therefore impose no CFCs. That being said, the results of our study do not rule out that CFC variation is not the result of a competition mechanism between a sentence and its alternatives. For example, drawing on insight from the weak/strong distinction, it may be possible to re-formulate this proposal by hypothesizing that competitors are formed based on deletion alone, rather than substitution *or* deletion. Because the additive triggers can all be removed without changing the asserted content of an utterance, such a theory would predict that they are harder to accommodate, which is precisely what we observed in our data.

5.2 Towards a Theory of Novelty Effects

In this section we are going to take a stance and endorse a theory of novelty effects which we believe is both predictive, insofar as it attains good empirical coverage, as well as explanatory, insofar as it explains *why* some triggers are difficult to accommodate. Specifically, it does a good job of explaining why both the anaphoric approach and the weak/strong distinction do a good job of predicting the data. We will then introduce two potential challenges for this theory. We believe these challenges are possible to overcome.

We endorse a hybrid theory of novelty effects from within a restrictions-plus-accommodation framework where a trigger's contextual felicity is determined by its anaphoricity. Some triggers are anaphoric while others impose constraints on the context set.[17] It is more difficult for people to adjust their representation of the common ground (i.e. what discourse referents have been introduced) than of the context (i.e. what propositions have been assented to), so when these triggers are used without a local antecedent, accommodation fails. However, we believe that simply stipulating that some triggers are anaphoric is not very explanatory. Why are they anaphoric? In this section, we answer this question by linking anaphoricity to a different presupposition property—namely its obligatoriness.

Instead of asking, when do triggers impose CFCs? Or, when do triggers resist accommodation? We ask, instead, why presuppose at all? That is: (a) Why do languages have presupposition mechanisms in the first place? And, (b), given the mechanisms for presupposition provided by the language, why do producers use these mechanisms?

For a long time the dominant approach to these types of questions grounded presupposition use in universal pressures like Maximize Presupposition (Heim, 1991), which does a good job of explaining (b) but has less to say about (a). A more recent approach, introduced in Bade (2016) attempts to answer both questions simultaneously. Bade's proposal is that certain triggers may be necessary to cancel inferences of exhaustivity that would create logical contradictions if the trigger weren't present. To give a brief example of this approach—called the Obligatory Implicatures, or OI approach—consider the sentences in (13), where the subscript F indicates that the word bears focus.

(13)   Ari likes Puccini...

---

[17]It's also possible for triggers to both impose constraints *and* be anaphoric, as advocated by Aravind and Hackl (2017) in their analysis of *too*. In this case, if a trigger introduces an anaphor which is unbound, it is not accommodated.

a. #She likes Verdi$_F$.

b. She likes Verdi$_F$, too.

The basic insight of Bade (2016) is that, in the second sentence, because *Verdi* conveys novel information it must bear focus, and is therefore obligatorily exhaustified against local alternatives. Exhaustification is a strengthening phenomenon that negates alternatives (Chierchia et al., 2012), which in this case we assume to be other composers. Unfortunately, this leads to a meaning something along the lines of "Ari like Verdi and no other composers," which clashes with what has just been uttered. What *too* does is to anaphorically identify itself with *Puccini* and put both under the scope of exhaustification, turning (13-b) into something like "She likes Verdi *and Puccini* and no other composers", which no longer contradicts the first sentence.[18] (For a full explanation of this process see Aravind and Hackl 2017.) Thus, *too* is obligatory to avoid unwanted and potentially contradictory inferences of exhaustivity.

Crucially, for us, there is a one-to-one correspondence between the triggers that are proposed to be obligatory under the OI approach, the traditionally anaphoric ones, and the weak triggers. Thinking about the issue in terms of potential semantic contradictions can explain why. Let's take the case of *too*: In order to avoid the contradiction, this item must bring material under the scope of a local exhaustifier by identifying true alternatives in the common ground (i.e. that Ari likes *Puccini*). Thus, it should be anaphoric. Furthermore, this item should accomplish the task without changing the asserted meaning of the current utterance. Thus, it should be *weak*. Finally, this item will be used every time the current message is exhaustified with respect to a QUD that has been already partially addressed, as in Example (13). Thus, it will be obligatory in many cases. The point we wish to draw out here is that Bade's proposal can give us a unified account of not only trigger obligatoriness, but also weak/strong status, anaphoricity and, by extension, contextual felicity as well. Finally, identifying the obligatory triggers with the anaphoric ones can give us an independent diagnostic to pick out this class of presuppositions, alleviating the concern of circularity about the anaphoric triggers raised in Section 2.

5.3 Two Challenges

Having laid out what we think is the most attractive theoretical option based on our data, we want to briefly mention two possible concerns, one theoretical and the other

---

[18]Actually, this analysis was not the one originally proposed in Bade (2016), but is taken from Aravind and Hackl (2017), who update Bade's proposal to account for some outstanding empirical problems.

empirical.

The theoretical concern is the following: The anaphoric/obligatoriness proposal relies on the premise that representations of the discourse (i.e. what discourse referents have been introduced) are harder to amend than representation of the context (i.e. what propositions have been assented-to). That is, people are unwilling to accommodate unbound anaphors, but willing to change their representations of the context to accommodate un-entailed presuppositions. Some previous work has attempted to justify this assumption. For example, von Fintel (2008) suggests that "[T]here cannot be accommodation with presuppositions that do not just target what is in the [context] but concern facts in the world that no manner of mental adjustment can bring into being. A particular case of that is the actual history of the conversation (the conversational record)..." However, we believe it is still an open question whether facts about the conversation really are less amenable to adjustment than the context, and suggest it as a potential avenue for experimental work that can test the assumptions of this theory.

The second, empirical, problem is the following: The items which we have been referring to as 'anaphoric' presupposition triggers don't seem to pattern together with other anaphors, in particular pronouns. In order for anaphors to be licensed there needs to be a salient, often linguistic, antecedent in the local context. This feature of anaphoricity is elegantly captured by (14), which is attributed to Barbara Partee (Heim, 1982):

(14)   Nine of the ten marbles are in the bag. ??It is probably under the sofa.

Although the existence of the missing marble is entailed by the first sentence, it is not made salient enough for the pronoun, *it* to co-refer. Because of this, the pronoun in the second sentence is unbound, and the whole utterance sounds infelicitous as a result. However, consider a variant of this sentence where *too* takes the place of the pronoun.

(15)   *Two people are cleaning up after playing a board game.*
         A: Only nine out of ten marbles are back in the bag.
         B: The dice$_F$ are missing, too.

Here, focus marking is placed on *dice* to help draw out the low-scope reading of person B's utterance, which presupposes that some alternative to the the dice is missing.[19] Under this reading, if *too* were anaphoric, it would require a salient antecedent in the

---

[19]There is an alternate, high-scope reading, under which *too* associates with the whole sentence, presupposing that something other than the dice missing is wrong. Both sentences are acceptable (author's judgement), but it is the low scope reading that is relevant to the argument.

discourse. Furthermore, because of the congruence conditions imposed by focus, this antecedent would have to have the property of being missing. The nine marbles that are back in the bag couldn't be this antecedent, as they don't meet this requirement (they are not missing). The missing marble also couldn't be its antecedent, as it is not salient enough to license anaphoric co-reference, as we saw with (14). Thus, on the assumption that the same types of linguistic contexts that license pronominal anaphora also license propositional anaphora, then the sentence is predicted, incorrectly, to be infelicitous. The point here is that if presuppositions' CFC behavior is explained by their being anaphoric, then we're going to need a story about why their behavior differs from other anaphors in some important ways.

6   Conclusion

The question of informative presupposition is a fruitful and challenging area of research that has received less attention than it should. At the highest level, we hope that this article serves to center these issues and show that they are connected in interesting ways to other issues in the literature on presupposition, semantics and pragmatics. At a more concrete level, this article has gained traction on the issue of informative presupposition by presenting the results of a large-scale rating study, which tests the contextual felicity of fourteen different English presupposition triggers. We show that there is substantial trigger-by-trigger variation, but that triggers fall (roughly) into three clusters. We compare these comprehension-side results to naturally occurring production data, and find strong correlations between CFC effects and the proportion of times triggers are used to introduce novel information. We endorsed a theory of these novelty effects that roots the observed behavior in presuppositions' hybrid status—some impose constraints on their context whereas others include an (additional) anaphoric element, which is less amenable to accommodation. We believe that this approach both captures the overall data, and also does a good job of explaining it, by connecting the triggers CFC strength, anaphoricity, weak/strong status, and obligatoriness. We argue that these properties of triggers do not just cluster together accidentally, but are a function of their role in discourse, which we can model as potentially canceling contradictory exhaustivity inferences.

Moving forward, this study raises a number of unanswered questions: What strength CFCs are associated with *only* and clefts, and are these due to accommodation difficulty or some other pragmatic infelicity? Are the CFCs associated with wh-questions the same as those for our other triggers, and why are they relatively strong? How does the role of information content relate to the theory we endorse? And do the findings

presented here hold up when tested in other languages? Ultimately, as Karttunen (2016) suggests for presuppositions themselves, novelty effects may be a heterogeneous phenomenon. We do hope, however, that the data and discussion presented here clear the way for a more satisfactory theory of the phenomena.

References

Abrusán, M. (2011). Predicting the presuppositions of soft triggers. *Linguistics and philosophy*, 34(6):491–535.

Abrusán, M. (2016). Presupposition cancellation: Explaining the 'soft–hard' trigger distinction. *Natural Language Semantics*, 24(2):165–202.

Abusch, D. (2002). Lexical alternatives as a source of pragmatic presuppositions. In *Semantics and linguistic theory*, volume 12, pages 1–19.

Abusch, D. (2010). Presupposition triggering from alternatives. *Journal of Semantics*, 27(1):37–80.

Amaral, P. and Cummins, C. (2015). A cross-linguistic study on information backgrounding and presupposition projection. In *Experimental perspectives on presuppositions*, pages 157–172. Springer.

Aravind, A. and Hackl, M. (2017). Against a unified treatment of obligatory presupposition trigger effects. In *Semantics and linguistic theory*, volume 27, pages 173–190.

Bacovcin, H. A., Zehr, J., and Schwarz, F. (2018). To accommodate or to ignore? the presuppositions of again and continue across contexts. *Glossa: A journal of general linguistics*, 3(1).

Bade, N. (2016). Obligatory presupposition triggers in discourse-empirical foundations of the theories maximize presupposition and obligatory implicatures. *Tübingen: Universität Tübingen PhD dissertation*.

Bade, N. and Renans, A. (2021). A cross-linguistic view on the obligatory insertion of additive particles—maximize presupposition vs. obligatory implicatures. *Glossa: A journal of general linguistics*, 6(1).

Beaver, D. and Zeevat, H. (2007). Accommodation. *The Oxford handbook of linguistic interfaces*, pages 503–536.

Bill, C., Romoli, J., and Schwarz, F. (2015). Are some presuppositions scalar implicatures? assessing evidence from reaction times.

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of semantics*, 17(3):189–216.

Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, 66(1):123–142.

Bott, L. and Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, 51(3):437–457.

Carpenter, P. A. and Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological review*, 82(1):45.

Chemla, E. (2009). Presuppositions of quantified sentences: Experimental data. *Natural language semantics*, 17(4):299–340.

Chemla, E. and Bott, L. (2013). Processing presuppositions: Dynamic semantics vs pragmatic enrichment. *Language and cognitive processes*, 28(3):241–260.

Chemla, E. and Schlenker, P. (2012). Incremental vs. symmetric accounts of presupposition projection: An experimental approach. *Natural language semantics*, 20(2):177–226.

Chierchia, G., Fox, D., and Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, 3:2297–2332.

Cummins, C., Amaral, P., and Katsos, N. (2012). Experimental investigations of the typology of presupposition triggers. *HUMANA. MENTE Journal of Philosophical Studies*, 5(23):1–15.

Cummins, C., Amaral, P., and Katsos, N. (2013). Backgrounding and accomodation of presuppositions: An experimental approach. In *Proceedings of Sinn und Bedeutung*, volume 17, pages 201–218.

Dayal, V. (2016). *Questions*. Oxford.

Domaneschi, F., Carrea, E., Penco, C., and Greco, A. (2014). The cognitive load of presupposition triggers: Mandatory and optional repairs in presupposition failure. *Language, Cognition and Neuroscience*, 29(1):136–146.

von Fintel, K. (2008). What is presupposition accommodation, again? *Philosophical perspectives*, 22:137–170.

Glanzberg, M. (2005). Presuppositions, truth values, and expressing propositions. *Contextualism in philosophy*, pages 349–396.

Göbel, A. (2020). *Representing Context: Presupposition Triggers and Focus-sensitivity.* PhD thesis, UMass Amherst.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases.* PhD thesis, University of Massachusetts Amherst.

Heim, I. (1988). On the projection problem for presuppositions. In *Proceedings of the Second West Coast Conference on Formal Linguistics*, pages 114–125. Stanford University Press.

Heim, I. (1991). Articles and definiteness. *Semantics: An international handbook of contemporary research*, pages 487–535.

Karttunen, L. (2016). Presupposition: What went wrong? In *Semantics and Linguistic Theory*, volume 26, pages 705–731.

Kaup, B. and Dudschig, C. (2020). Understanding negation: Issues in the processing of negation. *The Oxford Handbook of Negation.*

Kripke, S. A. (2009). Presupposition and anaphora: Remarks on the formulation of the projection problem. *Linguistic Inquiry*, 40(3):367–386.

Lassiter, D. (2012). Presuppositions, provisos, and probability. *Semantics and Pragmatics*, 5:2–1.

Lewis, D. (1979). Scorekeeping in a language game. In *Semantics from different points of view*, pages 172–187. Springer.

Marty, P., Chemla, E., and Sprouse, J. (2020). The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: A journal of general linguistics*, 5(1).

Prince, A. and Smolensky, P. (2004). *Optimality Theory: Constraint interaction in generative grammar.* John Wiley & Sons.

Roberts, C. (2003). Uniqueness in definite noun phrases. *Linguistics and philosophy*, 26:287–350.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:6–1.

Romoli, J. (2015). The presuppositions of soft triggers are obligatory scalar implicatures. *Journal of semantics*, 32(2):173–219.

Romoli, J. and Schwarz, F. (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In *Experimental perspectives on presuppositions*, pages 215–240. Springer.

van der Sandt, R. and Geurts, B. (2001). Too. In *Proceedings of the Thirteenth Amsterdam Colloquium*, pages 180–185. Amsterdam : ILLC.

van der Sandt, R. A. (1992). Presupposition projection as anaphora resolution. *Journal of semantics*, 9(4):333–377.

Schlenker, P. (2008). Be articulate: A pragmatic theory of presupposition projection. *Theoretical Linguistics*, 34(3):157–212.

Schwarz, F. (2007). Processing presupposed content. *Journal of Semantics*, 24(4):373–416.

Schwarz, F. (2014). Presuppositions are fast, whether hard or soft—evidence from the visual world. In *Semantics and Linguistic Theory*, volume 24, pages 1–22.

Schwarz, F. and Tiemann, S. (2017). Presupposition projection in online processing. *Journal of Semantics*, 34(1):61–106.

Simons, M., Tonhauser, J., Beaver, D., and Roberts, C. (2010). What projects and why. In *Semantics and linguistic theory*, volume 20, pages 309–327.

Singh, R., Fedorenko, E., Mahowald, K., and Gibson, E. (2016). Accommodating presuppositions is inappropriate in implausible contexts. *Cognitive Science*, 40(3):607–634.

Spenader, J. (2002). *Presuppositions in spoken discourse*. PhD thesis, Stockholm University.

Stalnaker, R. (1973). Presuppositions. *Journal of philosophical logic*, 2(4):447–457.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Strawson, P. F. (1950). On referring. *Mind*, 59(235):320–344.

Tiemann, S., Kirsten, M., Beck, S., Hertrich, I., and Rolke, B. (2015). Presupposition processing and accommodation: An experiment on wieder ('again') and consequences for other triggers. In *Experimental perspectives on presuppositions*, pages 39–65. Springer.

Tiemann, S., Schmid, M., Bade, N., Rolke, B., Hertrich, I., Ackermann, H., Knapp, J., and Beck, S. (2011). Psycholinguistic evidence for presuppositions: On-line and off-line data. In *Proceedings of Sinn und Bedeutung*, volume 15, pages 581–596.

Tonhauser, J. (2015). Are 'informative presuppositions' presuppositions? *Language and Linguistics Compass*, 9(2):77–101.

Tonhauser, J., Beaver, D., Roberts, C., and Simons, M. (2013). Toward a taxonomy of projective content. *Language*, pages 66–109.

Tonhauser, J., Beaver, D. I., and Degen, J. (2018). How projective is projective content? Gradience in projectivity and at-issueness. *Journal of Semantics*, 35(3):495–542.

Wilcox, E. G., Levy, R., and Davidson, K. (2021). Which presuppositions are subject to contextual felicity constraints? In *Semantics and Linguistic Theory*, volume 31, pages 345–364.

Zeevat, H. (1992). Presupposition and accommodation in update semantics. *Journal of semantics*, 9(4):379–412.

Zeevat, H. (2002). Explaining presupposition triggers. *Information sharing*, pages 61–87.

A   Experimental Results by Trigger/Environment

Figure 6 (next page) shows results for each trigger/condition pair. For summary visualizations of these data, and discussion see Section 3.2.

Figure 6: **Results by trigger/environment condition:** Triggers are arranged alphabetically, top-to bottom. The x-axis indicates whether the context supports the presupposition and the y-axis indicates participants' z-scored ratings. Error bars show 95% confidence intervals; red plots are $-trigger$ ratings and blue plots are $+trigger$ ratings.