Proceedings of SALT 31: 345–364, 2021

Which presuppositions are subject to contextual felicity constraints?*

Ethan Gotlieb Wilcox
Harvard

Roger P. Levy
MIT

Kathryn Davidson
Harvard

Abstract

Some sentences with presupposition triggers can be felicitously uttered when their presuppositions are not entailed by the context, whereas others are infelicitous in such environments, a phenomenon known as Missing Accommodation / Informative Presupposition or varying Contextual Felicity Constraints (CFCs). Despite an abundance of recent quantitative work on presuppositions, this aspect of their behavior has received less attention via experimentation. Here, we present the results from a semantic rating study testing the relative CFC strength of thirteen presupposition triggers, making this the largest cross-trigger comparison reported in the literature to date. The results support a three-way categorical analysis of presupposition triggers, based on imposing strong, weak, or no CFCs. We observe that strong CFC triggers are all focus-associating, suggesting that (at least some of the) variation in behavior arises due to naturally-occurring semantic classes. We compare our results to three previous proposals for CFC variation and argue that none yet account for the full empirical picture.

Keywords: presupposition, accommodation, contextual felicity constraints

1 Introduction

Presuppositions are the parts of meanings of utterances that are seemingly non-novel and backgrounded, and survive various entailment-canceling operations. Individual lexical items that introduce presuppositions are typically called *presupposition triggers*, and they constitute a heterogeneous functional class, including semantic operators, additive particles, determiners, and embedding verbs, to name a few. In this paper, we focus on the observation that some sentences with presupposition

* We thank the reviewers and audience at SALT 31, the Meaning and Modality Lab at Harvard and the Computational Psycholinguistics Lab at MIT, and Gennaro Chierchia for valuable feedback during this project.

triggers are felicitous when their presuppositions are not entailed by the preceding context, whereas others are allergic to such environments; that is they vary in the strength of a Contextual Felicity Constraint (CFC) (Tonhauser, Beaver, Roberts & Simons 2013), or are cases of Missing Accommodation / Informative Presupposition (Beaver & Zeevat 2007). To give a basic example, suppose that Xavi arrives at work one day to notice that his boss, Ari, is acting strangely. He ask a coworker about it, following the model dialog in (1).

(1) Xavi: What's up with Ari?

a. #She spilled coffee on herself, too.

b. She is embarrassed that she spilled coffee on herself.

Here, the presupposition trigger *embarrassed that* presupposes the truth of its complement, while the presupposition trigger *too* presupposes that someone else spilled coffee, that Ari spilled something else, or that Ari did something else that is contextually relevant (depending on the placement of focus). Speakers of English typically report that (1-a) is a less acceptable answer to Xavi's question than (1-b), given the context.

There are two families of approaches to explain the variation observed in (1). The first approach, which we take to be the majority view, treats presupposition triggers as introducing meanings that have a special status; they are either constraints on the context in which they are uttered (Heim 1983) or anaphoric elements which must be bound by referent in the discourse context (van der Sandt 1992). When these requirements are not met, an *accommodation mechanism* can kick in, and adjust the context on-the-fly to meet the requirements imposed on it by the presupposition trigger (Lewis 1979). Under this view, the variation observed in (1) is a case of success or failure in the accommodation mechanism, giving rise to the term 'Missing Accommodation' to describe (1-a). Under the second approach, what is inferred in (1-b) is not actually a presupposition, insofar as it is not associated with a constraint on the common ground, or a binding requirement (Tonhauser 2015). Rather, lexical items like *embarrassed that* pattern with other projective content like Conventional Implicatures (Potts et al. 2005).[1] Throughout, we will refer to the first approach as *constraint + accommodation* and the second approach as *not-presuppositions* approach.

When adjudicating between these two approaches, we must keep in mind both empirical coverage and explanatory power. The *not-presuppositions* approach has potentially perfect empirical coverage—if CFC variation is controlled by a lexical

1 Under this view, the mechanism that gives rise to a CFC is different from the mechanism that gives rise to projectivity. Tonhauser (2015) highlights that this proposal is compatible with approaches that derive projectivity from local information structure (Simons, Tonhauser, Beaver & Roberts 2010).

feature on the trigger, then we could just stipulate the number of features needed to derive the observed differences in CFCs. However, by the same token, it potentially lacks explanatory power, especially if it is the case that the CFCs within presuppositions pattern with other semantic properties (such as focus, as argued in Göbel (2020)). The *constraints + accommodation* approaches have the potential to be more explanatory, but it is unclear how well they predict the data, due to the paucity of systematic cross-trigger evaluations of CFC strength. Thus, our goal is to collect a broad set of cross-trigger data (in English), with the aim of evaluating the various candidate proposals for when and why the accommodation fails.

The rest of the paper will proceed as follows: In Section 2 we will discuss three candidate proposals for missing accommodation, the Information Content Hypothesis (van der Sandt & Geurts 2001), the Non-Presupposing Alternatives Hpyothesis (Blutner 2000), and the Focus Presupposition Antecedent Hypothesis (Göbel 2020). We will keep track of the predictions that each theory makes. In section 3 we outline the methods we deploy to assess CFC variation between presupposition triggers. In Section 4 we present the results of our study, which indicate that all presuppositions except factive embedding verbs and possessive pronouns are subject to some Contextual Felicity Constraints. We then present the results from a follow-up analysis suggesting that triggers can be further divided into two classes: Triggers that bear a strong CFC (*too*, *even*, *only* and clefts) and triggers that bear a weak CFC, (iteratives, change-of-state and accomplishment verbs, and the definite article). We validate our methods by comparing our results to a corpus study (Spenader 2002), and demonstrate a strong correlation between CFC strength and the proportion of times a trigger is used informatively in production. In Section 5, we assess the empirical coverage for each of the three candidate theories, arguing that none predicts the full range of observed variation, although our data provide support for the importance of focus-association, as hypothesized by Göbel (2020). In light of these results we conclude that it is worth pursuing a theory of CFC variation within the *constraints + accommodation* approach that grounds the behavior in naturally-occurring properties of the triggers, combined with the local context.

2 Background

Theoretical approaches to presuppositions fall into three broad categories: Semantics, Pragmatic and Hybrid. Semantic approaches to presuppositions treat them as pieces of meaning that have a distinct semantic status compared to the rest of the entailments associated with an utterance. In early work, triggers could impute a special truth-conditional value onto a sentence in cases where their presuppositions were not met (Strawson 1950). In contemporary frameworks that treat utterances as instructions for updating a shared conversational context, presuppositions are modeled as either

constraints placed on the context that must be met prior to utterance interpretation (Heim 1983), or else anaphoric elements that must be bound by a referent in the context (van der Sandt 1992). Under pragmatic approaches, presuppositions are treated as regular-old entailments of an utterance whose special properties are derived from the way that they relate to the local context. The source of this derivation can be speaker attitude (Stalnaker 1973), whether the entailment is *at issue* for answering a local Question Under Discussion (Simons et al. 2010), or whether the entailment is necessarily about the same event-time as the utterance's matrix predicate (Abrusán 2011, 2016). Hybrid approaches posit that presuppositions are a heterogeneous class, and recruit semantic, pragmatic and other proposals to explain cases of variation in intra-trigger behavior. Two important hybrid approaches are the *soft vs. hard* distinction of Abusch (2002, 2010) and the *strong vs. weak* distinction of Glanzberg (2005); Domaneschi, Carrea, Penco & Greco (2014). The former treats presupposition behavior as a blend of semantic constraints plus alternative-based reasoning. The latter divides presuppositions into two categories based on whether they can cause interpretation failure.

The majority of work attempting to explain Contextual Felicity Constraint variation has been conducted within semantic approaches, and so it is within these paradigms that we will remain for the rest of this section. One problem that arises is that modern semantic approaches by themselves are too brittle, and become too flexible with added mechanisms. If presuppositions impose constraints that are necessary for utterance interpretation, then any trigger-bearing utterance whose presuppositions are not met should result in catastrophic interpretation failure. However, sentences like (1-b) can be used informatively in normal discourse. To solve this problem, semantic approaches invoke an accommodation mechanism: a process through which comprehenders can update the context quietly and without fuss prior to utterance interpretation in order to satisfy an utterances' presuppositions (Von Fintel 2008). Here is the original proposal from Lewis (1979): "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." Of course, without spelling out these *certain limits*, accommodation is not a *theory* in the sense that it does not make testable predictions about what can be accommodated and what can't be. In the sections below, we review three proposals that develop concrete proposals for Lewis's certain limits. (We focus on simple, matrix sentences here, and for the rest of the paper.)

Information Content The first proposal for CFC variation we will discuss is the 'Information Content' approach, advocated in Geurts & van der Sandt (2004). These authors, who were working within the presuppositions-as-anaphors approach, postulate that the only difference between presuppositions and pronouns is the

amount of information content presuppositions contain, and the fact that they can have an internal structure. Working from this observation, their proposal is that presuppositions cannot be accommodated if they are semantically impoverished, which makes it difficult to build discourse referents on the fly. They complicated the picture, however, by noting that many presuppositions which are difficult to accommodate are not semantically impoverished: Take *too*, whose presupposition varies with the placement of focus and will have as much information content as the focused constituent. To solve this problem, Geurts & van der Sandt propose that triggers like *too* actually encode two presuppositions: "The first [presupposition] resembles a pronoun in the sense that it has no descriptive content to speak of, and therefore should be hard to accommodate. The second [presupposition] is richer in descriptive content" (p. 48). As noted in Beaver & Zeevat (2007) this proposal has a number of technical challenges, including one of overgeneration. In addition to the technical challenges, Geurts & van der Sandt do not include a set of general criteria for determining when a trigger has just one or two presuppositions. As such, this theory does not make clear predictions about which triggers should be difficult or easy to accommodate. So while we acknowledge that the proposal is theoretically well-motivated, we set it aside here for lack of predictive clarity.

Non-Presupposing Alternatives The second proposal, first introduced in Blutner (2000), treats accommodation as the result of a competition mechanism, in which non-presupposing alternatives compete with and potentially block presuppositional utterances. While this proposal was developed within the framework of Bidirectional Optimality Theory, we simplify things here by approximating the OT results with *Blutner's Theorem* from Beaver & 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 details lie in which alternatives we allow to enter the competition. Developing the proposal, 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 operationalize the notion of *simple* alternative expressions, we adopt the grammatical alternatives approach from Katzir (2007), with the addition of negation as a single substitution.[2] But defining a set of structural alternatives is only half the challenge, for the non-presupposing alternatives have to have *the same meaning* as the presuppositional sentence. There

2 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.

are two possible ways to construe this requirement. For presupposing sentence $p + p'$ with asserted content p' and presuppositions p we can say that non-presupposing alternative q has the same meaning with respect to the whole content (that is $q \models p' \wedge p$) or just the asserted content ($q \models p$). Furthermore, we will assume that candidates alternatives must convey only the same asserted content, so it is this later requirement that we will use to construct inputs into our competition mechanism. The triggers discussed in this paper, along with their simple non-presupposing alternatives and the predictions of this proposal, are given in Table 1. All presuppositions are predicted to not accommodate, except for accomplishment verbs, for which there are no simple non-presupposing alternatives.

Content vs. Discourse Presuppositions Göbel (2020) proposes the Focus Presupposition Antecedent Hypothesis (FoPAH): "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 Common Ground." In this case, common ground is the one defined in Stalnaker (2002)—an unordered set of propositions which are mutually-assented to for the purposes of conversation. The Discourse Model, on the other hand, is a structured representation that keeps track of previous referents and questions under discussion. If we assume that the discourse model is harder to amend on-the-fly than the common ground, then we can derive the variation in CFC strength between focus-sensitive and non focus-sensitive triggers.

Although this approach must postulate two categories of presupposition trigger and thus introduce more complexity into the semantic theory, it derives CFC behavior from independent facts about the triggers (i.e. their focus-sensitivity) so it is arguably less stipulative than the *not-presuppositions* approach. However, there are two areas where the theory may require additional development: The first has to do with why the Discourse Model is easier to amend on the fly than the Common Ground. Göbel (2020) makes it clear that this is an assumption of the proposal, and postulates that it is due to the fact that the discourse record is not subject to the same Gricean principles that govern the Common Ground. A different approach suggested by Göbel may be to link the discourse model more closely with the discourse record. Grounding accommodation difficulty in the discourse record has been advocated previously, e.g. by Beaver & Zeevat (2007) and Von Fintel (2008), who states "[T]here cannot be accommodation with presuppositions that do not just target what is in the common ground 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)..."

Another area where the theory may require additional development would be to account for a richer range of CFCs. As proposed in Göbel (2020), the FoPAH is

Trigger	Alternative	Blutner (2000)	Göbel (2020)
It-Clefts	Bare utterance	No Accom	No Accom
Even	Bare utterance	No Accom	No Accom
Too	Bare utterance	No Accom	No Accom
Only	Bare utterance	No Accom	No Accom
Wh-Questions	If-Questions	No Accom	Potential Accom
The X	An X	No Accom	Potential Accom
State Change Verbs (e.g. stop, continue)	Isn't Xing	No Accom	Potential Accom
Back	Bare utterance	No Accom	Potential Accom
Again	Bare utterance	No Accom	Potential Accom
Still	Bare utterance	No Accom	Potential Accom
Accomplishment Verbs (e.g. win)	**none**	Accom	Potential Accom
His/Her X	An X	No Accom	Potential Accom
Factive Verbs (e.g. know that / annoyed that)	Believe that X	No Accom	Potential Accom

Table 1 Triggers investigated, with alternatives and the predictions of previous theories. *Bare utterance* indicates that the alternative is created by removing the presupposition trigger (e.g. *Alex sang, too→Alex sang*)

intended to explain a dichotomous distinction in CFC strength, between triggers that are focus-sensitive and those that are not. Additional mechanisms may be needed in order to explain richer variation, or to explain why some CFCs disappear altogether, as is the case with informative presuppositions. The predictions of this approach are presented in the right column in Table 1.

3 Methods

3.1 Design

To assess the strength of Contextual Felicity Constraints, we employed a 2x2 experimental design 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).[3] By "supports" we mean that a presupposition is either entailed or

3 These are the same as what Tonhauser et al. (2013) call NEUTRAL (our -SUPPORTING) and POSITIVE (our +SUPPORTING)

that the trigger is provided with a discourse referent or linguistic antecedent in the immediate context. Example (2) 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. For more information on the construction of the materials see Section 3.2, below.

(2) 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, we expect (d) to be rated as more acceptable than (b). In addition, if a trigger imposes a CFC, then in a non-supporting context, we expect a trigger-bearing sentence to be less acceptable than a sentence without a presupposition trigger, provided there are no other differences in meaning between the two. 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 we will refer to as the ±*Supporting within +Trigger* contrast. Additionally, Göbel (2020) investigates the (a) vs. (b) contrast, which we refer to this as the ±*Trigger within +Supporting* contrast.

In addition to these two-way contrasts, we propose to measure CFCs by looking at the interaction between trigger presence and trigger support. That is, if a trigger imposes a CFC then we expect lower semantic acceptability when the trigger is present and its presuppositions are supported, compared to contexts when the trigger is not present or when it is present but not supported. Measuring CFC strength with the interaction term provides controls that may be lacking in the simpler two-way contrasts. In addition to providing controls, this fully-crossed design allows us to compute both the previously-used contrasts, giving this study a secondary purpose of answering methodological questions about inter-reliability between these three proposed metrics.

For the study, we employed the presentational design advocated in Marty, Chemla & Sprouse (2020), who report that joint presentation of conditions with a continuous

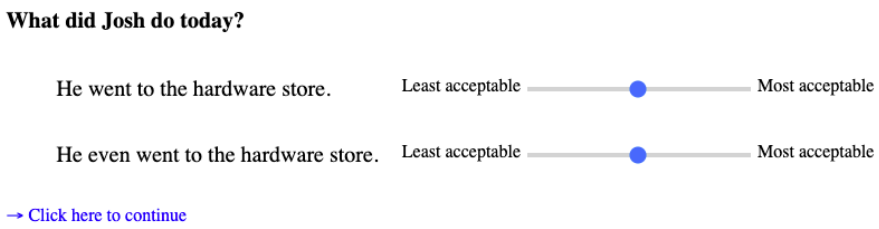

Figure 1 Sample item for the *even* trigger in the −*supporting* condition.

scale and labeled endpoints draw out robust contrasts between conditions in a rating task of this type. There are two advantages we would like to highlight about this experimental paradigm: First, it draws out robust contrasts because it allows for direct comparison between conditions 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. Cast in a positive light, this can be seen as an invitation to the participant to join linguists in reporting a range of linguistic data.

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.

3.2 Materials and Participants

We created items for the 14 triggers in Table 1, clumping Emotive and Cognitive Factives into a single category. For each trigger we created 5 items. 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 by the authors. +*Trigger* target items consisted of simple past-tense statements that included the presupposition; -*trigger* items were created using the non-presupposing alternatives from Table 1, with two differences: For accomplishment verbs the non-presupposing alternative was a verb describing the participatory action (e.g. *win*/*participate*, *pass the test*/*take the test*), and for factive predicates the non-presupposing alternative consisted in a mix of non-veritical predicates (e.g. *suspect*, *believe*, *think*).[4]

We recruited 32 participants on Amazon Mechanical Turk. Participants were all located within the US, were US High School graduates and had a lifetime MTurk completion rate of above 90%. They were instructed that they could only participate in the survey if they were native English speakers. The survey took about 20 minutes to complete and participants were paid for their participation. To make sure that participants were using the scale bar correctly, we filtered participants if they did not rate the number mismatched warm-up sentences in the bottom quartile of the response bar, which resulted in filtering out 6/32 participants.

4 Results and Analysis

4.1 Results

The results from the study can be seen in Figure 2, with the context on the x-axis; on the y-axis are ratings, which have been standardized (i.e. z-scored) for each participant to control for cross-subject variation. Red points are -*trigger* ratings and blue points are +*trigger* ratings. Error bars are 95% confidence intervals pooled by subject. In order to provide an initial assessment for which triggers were subject to a Contextual Felicity Constraint we use the interaction estimate discussed in section 3.1, and turn to comparison between the metrics in the next section. We fit a linear mixed-effects regression model, with experimental conditions as predictors (using 1/0 treatment coding) and random by-participant and by-item slopes for experimental conditions. We found a significant positive effect of the interaction term for Clefts, *even*, *too*, *only*, Wh-Questions, *the*, *back*, *again*, *still* (all $p < 0.001$), state change verbs ($p < 0.01$) and accomplishment verbs ($p < 0.05$). We found no effect for factive verbs, and a significant negative effect for possessive pronouns

4 All materials and results are hosted online at https://osf.io/cdm4h/.

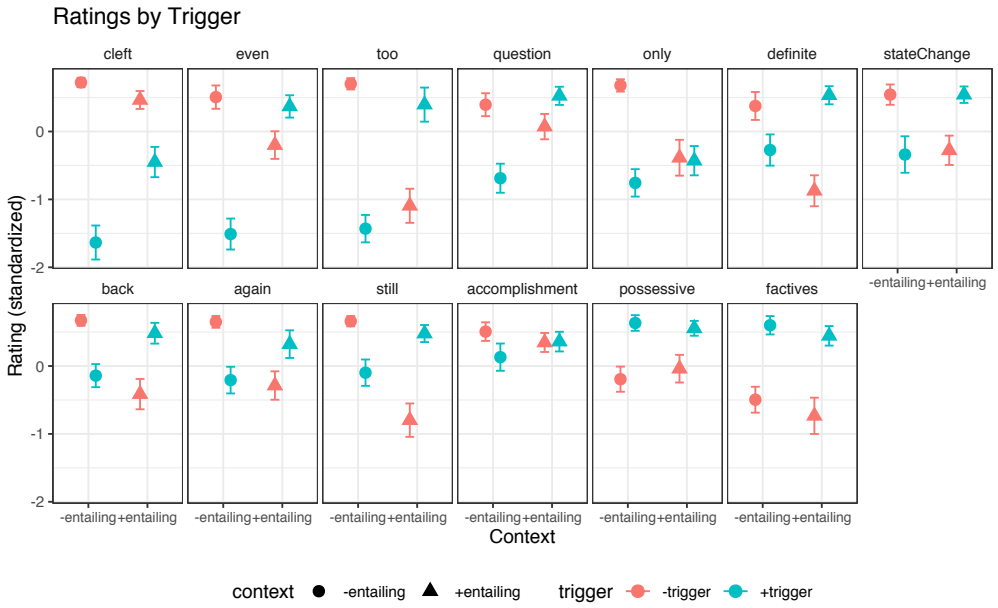

Figure 2 Results from the rating study, standardized within participant. Points are the means of each condition, error bars are 95% confidence intervals. The triggers are ordered from top left to bottom right by the mean rating for the trigger's *-supporting/+trigger* condition.

($p < 0.05$). This negative effect may be due to familiar pressures such as Maximize Presupposition (Heim 2008), but we set this point aside for future exploration. Looking at the triggers that do impose CFCs we find two types of interactions: The first are cases of spreading interactions, where we find a main effect of *+/- trigger* that is enhanced in the *-supporting* context. Triggers with spreading interactions include clefts, *only* and accomplishment verbs. More common are cases of cross-over interactions, where the relative felicity of the *+/- trigger* targets are reversed between the *-supporting* and *+supporting* contexts.

4.2 Comparing Metrics

Apart from assessing the CFCs of a large number of triggers, this study has a further methodological aim of assessing different metrics for measuring CFC strength. There are two previous proposals: Tonhauser et al. (2013) advocates that CFCs should be assessed by measuring the difference in acceptability for sentences that do have a presupposition between the *+supporting* and *-supporting* contexts (This is the $\pm$*Supporting within +Trigger* contrast; it corresponds to the difference between

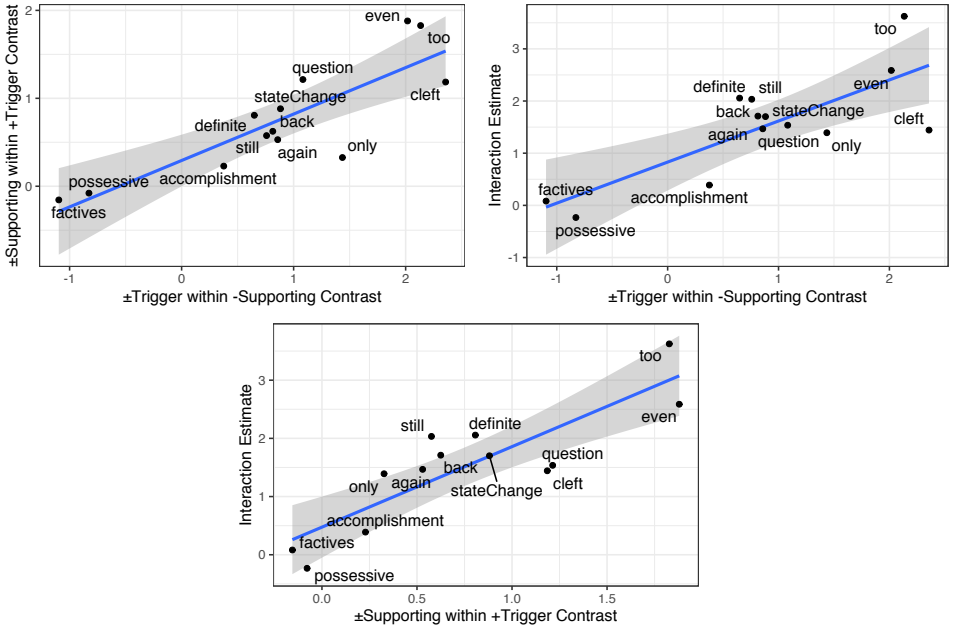

Figure 3 Three two-way comparison between three methods for assessing CFCs: ±*Trigger within +Supporting* Contrast (Göbel 2020), ±*Supporting within +Trigger* Contrast (Tonhauser et al. 2013) and *Interaction Term* (this work). We find strong correlations between all three approaches.

the blue circle and the blue triangles in panels in Figure 2.) The second metric, proposed by Göbel (2020), uses the contrast between *+trigger* and *-trigger* sentences in *-supporting* contexts. (This is the ±*Trigger within +Supporting* contrast; it corresponds to the difference between the blue circle and red circle at the left of each panel if Figure 2.) We compute each of these contrasts by taking the relevant differences after averaging across participants and trials for each trigger. To compare with interaction size we take the difference of differences between conditions for each trigger.

The three two-way comparisons between our metrics can be seen in Figure 3. Each panel shows a correlation between two metrics; for the two contrasts, the axes are differences in standardized (i.e. z-scored) ratings, and are relatively small. Regardless of comparison, we find strong correlations between each of the three metrics, indicating that they produce similar results for assessing the relative felicity of a presupposition trigger given its context. For the *interaction*/±*Supporting within +Trigger* comparison we find $corr = 0.86$ ($p < 0.001$); for the *interaction*/±*Trigger within +Supporting* comparison we find $corr = 0.78$ ($p < 0.01$); and for the ±*Supporting within +Trigger* /±*Trigger within +Supporting* comparison

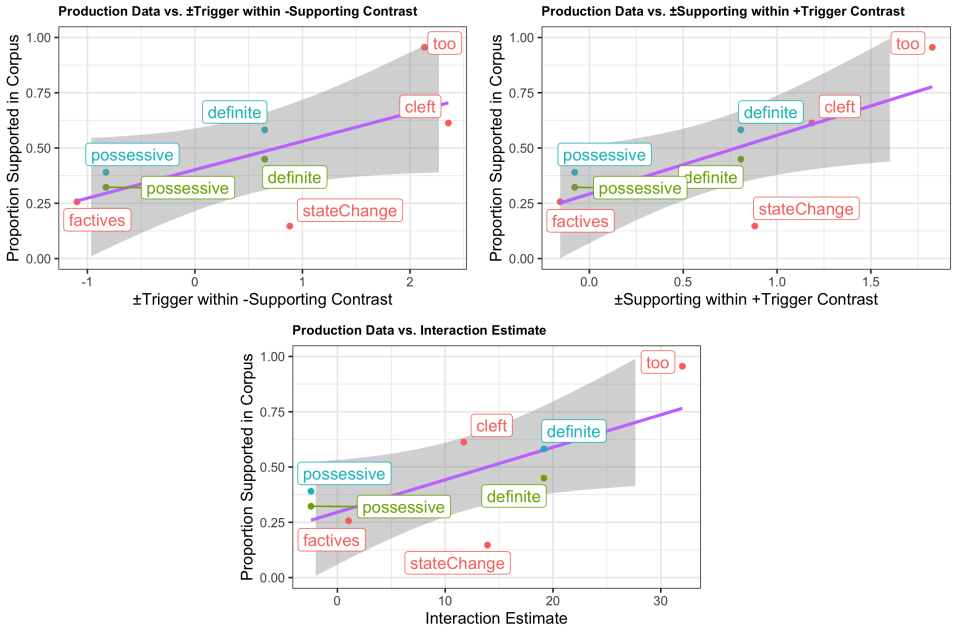

Figure 4 Comparison between methods for assessing CFCs to production data from Spenader (2002). All methods show strong correlations. Colors correspond to annotator identity (some triggers were annotated twice).

we find $corr = 0.84$ ($p < 0.001$). Given that each of these metrics seem to produce similar participant responses, which one should be used? Although our 2x2 interaction design was necessary for validating this methodological equivalence, it involves creating twice as many stimuli as either the *±Supporting within +Trigger* or *±Trigger within +Supporting* contrasts. Therefore, depending on the research question, one of the two simpler methods may be preferred.

4.3 Comparison to Production Data

Although each of the three metrics in question captures similar categorizations of presupposition triggers, there may be questions about the ecological validity of the experimental paradigm in capturing naturalistic uses. In this section, we validate our methods against production data from Spenader (2002), who collects 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. Following Spenader, for each trigger, we report the proportion of times it was supported. 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. Our 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 results from each of the three metrics discussed above 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. 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. Overall, we find a strong relationship 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). For the $\pm$*Supporting within +Trigger* we find $corr = 0.74$ ($p < 0.05$); for the $\pm$*Trigger within +Supporting* we find $cor = 0.67$ ($p = 0.06$); and for the interaction estimate we find $cor = 0.7$ ($p < 0.05$). The one point of difference between our results and the production data are change of state verbs, which were used with explicit support only about 15% of the time, but which we found to be associated with moderate contextual felicity constraints. Despite this, we take these strong correlations to provide validation for our results, and provide further evidence that each of the three metrics can be deployed to measure the strength of a trigger's CFC.

4.4 Cluster Analysis

Now that the main results of our study have been validated, we turn to a follow-up analysis that asks whether the data we have obtained support a gradient or categorical analysis of CFCs. There seems to be quite a bit of variation in the by-trigger ratings, especially in the crucial *-supporting/+trigger* condition, which is the left-hand blue dot in the panels in Figure 2. As most of our semantic theories make categorical predictions about the relative success of the accommodation mechanism, and thus the relative strength of the corresponding CFC, we want to ask whether the data support a categorical analysis of CFC strength. In order to answer this question, we run a hierarchical clustering algorithm on the results, treating the conditions as dimensions, and the mean response in each condition as a triggers value in that dimension. We use euclidean distance between points and select the `Ward.D` clustering algorithm from *R*'s `hclust` function. We run two clustering procedures: In the first, we filter out responses in the *+supporting/-trigger* condition, which served largely as a control condition; the second includes data from all conditions.

The results of our clustering algorithm can be seen in Figure 5, with the fil-

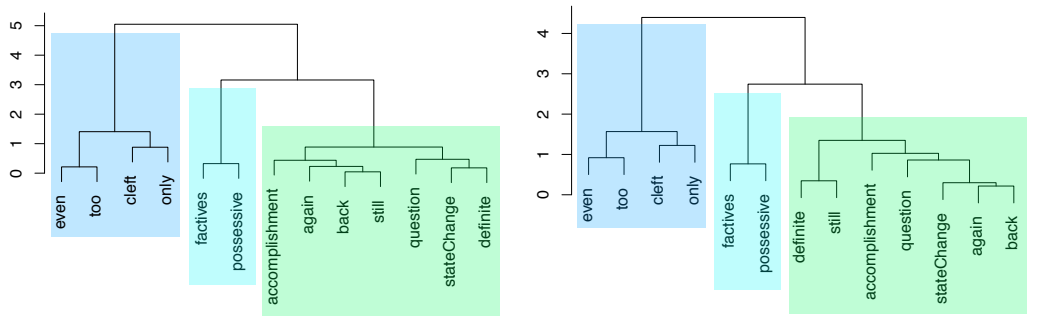

Figure 5 Clustering of the 13 triggers tested based on average standardized response in each condition. Colors highlight top two distinctions among clusters. Left image shows clusters with responses in the *+supporting/-trigger* condition filtered out, right image shows clusters with all data. Filtering does not effect overall cluster shape.

tered clusters on the left and the clusters with data from all conditions on the right. The filtering does not affect the overall outcome of the results. The shape of the clusters supports a categorical three-way division among the triggers. In the first cluster (highlighted in blue), we have all the focus-sensitive particles; these are associated with large ±*Trigger within +Supporting* contrasts, as well as relatively large ±*Supporting within +Trigger* contrasts, except for *only*. In the middle cluster (highlighted in teal) we have triggers that do not display traces of CFCs. These are associated with a negative ±*Trigger within +Supporting* contrast and a very small ±*Supporting within +Trigger* contrast. Finally, in the last cluster (highlighted in green), we have triggers which are associated with a weaker CFC. They have moderate ±*Trigger within +Supporting* and ±*Supporting within +Trigger* contrasts. For the rest of the paper, we refer to these three top-level groups as Strong-CFC triggers, Weak-CFC triggers and Non-CFC triggers. These clusters provide some additional support for the ±*Trigger within +Supporting* contrast as the best measurement of CFC strength. The three clusters are linearly separable using this metric, but not for either the interaction estimate or the ±*Supporting within +Trigger* contrast.

Before we continue, we want to emphasize that the clustering procedure is a method for picking up similarities in *surface level* behavior. Just because we find support for a categorical distribution in our results, that does not mean that these three categories are intrinsic aspects of the presupposition triggers. They may very well result from exogenous semantic and pragmatic factors. Furthermore, we do not predict that CFC behavior can or will ever be fully categorical. In any experimental paradigm and in any discursive context there will be some CFC variation that can be

Trigger	Abusch (2002) Hard/Soft	Glanzberg (2005) Weak/Strong	Blutner (2000)	Göbel (2020)	Our Results
It-Clefts	Hard	Strong	No Accom	No Accom	Strong CFC
Even	Hard	Weak	No Accom	No Accom	Strong CFC
Too	Hard	Weak	No Accom	No Accom	Strong CFC
Only	Hard	Weak	No Accom	No Accom	Strong CFC
Wh-Questions	Soft	Strong	No Accom	Potential Accom	Weak CFC
The X	Hard	Strong	No Accom	Potential Accom	Weak CFC
State Change Verbs	Soft	Strong	No Accom	Potential Accom	Weak CFC
Back	Hard	Weak	No Accom	Potential Accom	Weak CFC
Again	Hard	Weak	No Accom	Potential Accom	Weak CFC
Still	Hard	Weak	No Accom	Potential Accom	Weak CFC
Accomplishment Verbs	Soft	Strong	Accom	Potential Accom	Weak CFC
His/Her X	Hard	Strong	No Accom	Potential Accom	No CFC
Factive Verbs (know that/annoyed that)	Hard	Strong	No Accom	Potential Accom	No CFC

Table 2 Predictions of various theoretical proposals with the results of our study.

reduced to prior probability of the presupposed content, the goals of the speakers and listeners, the amount of mutual trust, etc. What we do take these results to mean is that a successful semantic theory may want to take seriously why certain triggers pattern together within the same category.

5 Discussion

The predictions of the various proposals discussed in Section 2 along with our results are shown in Table 2. In addition to the two theories that address the issue of accommodation failure directly, we include categorizations from two prominent hybrid proposals for presuppositions: the soft/hard distinction of Abusch (2002) and the weak/strong distinction of Domaneschi et al. (2014).

First, let's compare the distinctions made by the hybrid theories to our results. Because neither of these two makes explicit predictions about accommodation, our data does not provide 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 variation. We briefly introduce each proposal: First, the soft/hard distinction (Abusch 2002) was developed to explain why the presuppositions of some triggers can be canceled more

easily than others. It proposes that while some triggers (the *hard* ones) are *bona fide* semantic presuppositions, other presuppositional behavior results from alternative-based pragmatic reasoning. Second, the weak/strong distinction (Glanzberg 2005) was proposed to explain why some cases of missing accommodation result in interpretation failure and obligatory context repair (for *strong* triggers), whereas for other, *weak*, triggers, context repair is optional. For our purposes, these two categories can be cast in terms of type-based semantic frameworks, with semantic adjuncts as weak triggers and non-adjuncts as strong triggers. Semantic adjuncts are triggers modeled as two-place predicates with the same type in each place, in this case $\langle t,t \rangle$ for all of our additive and iterative presuppositions.

So what categorizations do these two approaches make? Starting with the weak/strong approach, there appears to be good overlap between strong triggers and triggers that don't impose CFCs. However triggers that do impose CFCs are split between the weak/strong categories, indicating that these divisions may be tangential to CFC strength. Furthermore, 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 no weak trigger is associated with a lack of CFCs. Thus, we conclude that while our data does not contradict the weak/strong hypothesis, it suggests that such a division is unlikely to explain CFC variation. Turning to the soft/hard distinction, we find more overlap between the relevant categories and our results: All of our Strong-CFC and Non-CFC triggers are hard, whereas all the soft triggers are associated with weak CFCs. But even if soft triggers are associated with weak-CFCs, this still leaves variation for 11/13 triggers unexplained. At best, grounding some CFC variation in the soft/hard distinction, still leaves important, unanswered questions about how hard triggers can impose both strong, weak, and no CFCs.

Now, we turn to theories that attempt to explain CFC variation by way of accommodation failure. We make the linking hypothesis that successful accommodation results in no CFCs, and accommodation failure results in weak and strong CFCs, which allows us to translate between the predictions of each theory and our data. Turning first to the Non-Presupposing Alternatives Proposal (Blutner 2000), we find that this theory has two types of problems: First, it overgenerates for accomplishment verbs, predicting that they should not be associated with CFCs, when we find that they are. Second, it undergenerates in cases of possessive pronouns, and cognitive/emotive factives, predicting that they should be associated with CFCs, when we find they are not. We contend that this latter shortcoming will be difficult for the theory to overcome, especially in the case of cognitive factives, which have high-frequency non-presupposing alternatives (*think*, *believe*, *suspect*). While it may be the case that accommodation variation is grounded in a competition mechanism between presupposing and non-presupposing sentence variants, simple

non-presupposing alternatives are associated with all types of CFC strength. Future competition-based theories will have to construct different alternative sets in order to capture the observed human behavior.

Turning to the Focus Presupposition Antecedent Hypothesis (Göbel 2020), we find better overlap between theory and data. This approach predicts that focus-sensitive triggers should be difficult to accommodate, and these are precisely the Hard-CFC triggers identified by our clustering analysis. But while the FoPAH correctly predicts which triggers impose Hard-CFCs, it does not make predictions about the rest of the triggers, nor can it explain why some triggers do not impose *any* CFCs. Thus, while our results provide support for the hypothesis that focus association is implicated in CFC strength, it highlights the fact that focus can only provide a partial answer. Furthermore, our data are agnostic about the content vs. discourse distinction, which Göbel hypothesizes as the representational mechanism driving Strong-CFC behavior. It may be the case that focus associating triggers require antecedents in a structured discourse record. But our data are equally compatible with an approach to CFC strength that views all presuppositions as constraints on an unstructured common ground, and derives strong-CFC behavior from outside pragmatic factors, such as the interference of focus association in question-answer congruence. For discussion of the importance of question-answer congruence in presuppositions, see Abrusán (2016).

Finally, we return to the question posed at the beginning: Does our data support the *constraints + accommodation* approach, or the *not-presuppositions* approach? First, we find that no proposal for accommodation failure manages to reach adequate empirical coverage. While this may seem like a reason to favor the *not-presuppositions* approach, we believe that this would be too-hasty a conclusion for two reasons: First, lack of an empirically adequate theory within the *constraints + accommodation* approach might have been due to a lack of good cross-trigger data as much as anything else. Second, this contribution, as well as Göbel (2020), clearly identify focus association as a mechanism that is implicated in CFC strength. With this initial theoretical step, combined with the cross-trigger data presented here we are optimistic about proposals that reaches better empirical coverage of the data.

6 Conclusion

We have investigated which presuppositions are subject to Contextual Felicity Constraints through a felicity judgement task for 13 different presupposition trigger categories, making our study the largest cross-trigger comparison reported in the literature to-date. Results showed a wide range of variation along the two dimensions studied (with and without trigger, with and without supporting context), but despite this variation, we found external validation for our methods by compar-

ing our resulting CFC strength to the proportion of times a presupposition trigger is supported in previously published production data (Spenader 2002). We ran a novel unsupervised clustering algorithm, advocating for a three-way split between Strong-CFC, Weak-CFC and Non-CFC triggers. Evaluating the resulting empirical picture against existing theoretical proposals for accommodation failure, we find that no theory reaches full empirical coverage, although our results broadly support focus-association being implicated in strong Contextual Felicity Constraints. Looking forward, we see promise broadly in theories that take into account the way that triggers interact with local context and information structure.

References

Abrusán, Márta. 2011. Predicting the presuppositions of soft triggers. *Linguistics and Philosophy* 34(6). 491–535. doi:10.1007/s10988-012-9108-y.

Abrusán, Márta. 2016. Presupposition cancellation: explaining the 'soft–hard'trigger distinction. *Natural Language Semantics* 24(2). 165–202. doi:10.1007/s11050-016-9122-7.

Abusch, Dorit. 2002. Lexical alternatives as a source of pragmatic presuppositions. In *Semantics and Linguistic Theory*, vol. 12, 1–19. doi:10.3765/salt.v0i0.2867.

Abusch, Dorit. 2010. Presupposition triggering from alternatives. *Journal of Semantics* 27(1). 37–80. doi:10.1093/jos/ffp009.

Beaver, David & Henk Zeevat. 2007. Accommodation. *The Oxford handbook of Linguistic Interfaces* 503–536.

Blutner, Reinhard. 2000. Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17(3). 189–216. doi:10.1093/jos/17.3.189.

Domaneschi, Filippo, Elena Carrea, Carlo Penco & Alberto Greco. 2014. The cognitive load of presupposition triggers: mandatory and optional repairs in presupposition failure. *Language, Cognition and Neuroscience* 29(1). 136–146. doi:10.1080/01690965.2013.830185.

Geurts, Bart & Rob A van der Sandt. 2004. Interpreting focus doi:10.1515/thli.2004.005.

Glanzberg, Michael. 2005. Presuppositions, truth values, and expressing propositions. *Contextualism in Philosophy* 349–396.

Göbel, Alex. 2020. *Representing context: presupposition triggers and focus-sensitivity*: UMass Amherst PhD dissertation.

Heim, Irene. 1983. On the projection problem for presuppositions. *Formal Semantics: The Essential Readings* 249–260. doi:10.1002/9780470758335.ch10.

Heim, Irene. 2008. Artikel und definitheit. In *Semantik/Semantics*, 487–535. De Gruyter Mouton. doi:10.1515/9783110126969.7.487.

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30(6). 669–690. doi:10.1007/s10988-008-9029-y.

Lewis, David. 1979. Scorekeeping in a language game. In *Semantics from Different Points of View*, 172–187. Springer. doi:10.1093/0195032047.003.0013.

Marty, Paul, Emmanuel Chemla & Jon Sprouse. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: a Journal of General Linguistics* 5(1). doi:10.5334/gjgl.980.

Potts, Christopher et al. 2005. *The Logic of Conventional Implicatures* 7. Oxford University Press on Demand. doi:10.1093/acprof:oso/9780199273829.003.0003.

van der Sandt, Rob A. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics* 9(4). 333–377. doi:10.1093/jos/9.4.333.

van der Sandt, Rob A & Bart Geurts. 2001. Too .

Simons, Mandy, Judith Tonhauser, David Beaver & Craige Roberts. 2010. What projects and why. In *Semantics and Linguistic Theory*, vol. 20, 309–327. doi:10.3765/salt.v0i20.2584.

Spenader, Jennifer. 2002. Presuppositions in spoken discourse .

Stalnaker, Robert. 1973. Presuppositions. *Journal of Philosophical Logic* 2(4). 447–457. doi:10.1007/bf00262951.

Stalnaker, Robert. 2002. Common ground. *Linguistics and Philosophy* 25(5/6). 701–721. doi:10.7748/ns.16.44.3.s1.

Strawson, Peter F. 1950. On referring. *Mind* 59(235). 320–344. doi:10.1093/mind/lix.235.320.

Tonhauser, Judith. 2015. Are 'informative presuppositions' presuppositions? *Language and Linguistics Compass* 9(2). 77–101. doi:10.1111/lnc3.12119.

Tonhauser, Judith, David Beaver, Craige Roberts & Mandy Simons. 2013. Toward a taxonomy of projective content. *Language* 66–109. doi:10.1353/lan.2013.0001.

Von Fintel, Kai. 2008. What is presupposition accommodation, again? *Philosophical Perspectives* 22. 137–170. doi:10.1111/j.1520-8583.2008.00144.x.

Zeevat, Henk. 2002. Explaining presupposition triggers. *Information Sharing* 61–87.

Ethan Gotlieb Wilcox
Boylston Hall
Harvard University
wilcoxeg@g.harvard.edu

Roger Levy
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
rplevy@mit.edu

Kathryn Davidson
Boylston Hall
Harvard University
kathryndavidson@fas.harvard.edu