Production Expectations Modulate Contrastive Inference

Anonymous CogSci submission

Abstract

Include no author information in the initial submission, to facilitate blind review. The abstract should be one paragraph, indented 1/8 inch on both sides, in 9 point font with single spacing. The heading "**Abstract**" should be 10 point, bold, centered, with one line of space below it. This one-paragraph abstract section is required only for standard six page proceedings papers. Following the abstract should be a blank line, followed by the header "**Keywords:**" and a list of descriptive keywords separated by semicolons, all in 9 point font, as shown below.

Keywords: add your choice of indexing terms or keywords; kindly use a semicolon; between each term

Introduction

One of the most interesting features of language is its flexibility. To refer to one single object a speaker can choose an utterance out of an indefinite set of possible referring expressions. *The banana*, *the yellow banana*, *the yellow, curvy fruit-thingy* for example are all possible utterances that can refer to the same object in Figure 1. At the same time, the same utterance – e.g., *banana* – can be used to refer to different kinds of objects. But this flexibility poses a challenge for for the listener, who needs to pragmatically infer the speaker's intention.

One of the most fundamental findings is that listeners process utterances incrementally, i.e., new information is incorporated into the interpretation of the utterance as soon as it becomes available [ek: cite]. For instance, eye-tracking experiments have shown that if a listener hears the incomplete utterance *the yellow* in a display like Figure 1, they fixate the yellow objects in the display even before they hear the disambiguating noun *banana* and [ek: arrive at] the final referent [ek: cite].

But listeners go beyond the information contained in the signal itself; they also take into account contextual information and specifically the nature of other possible referents to draw rapid pragmatic inferences about the speaker's intention [ek: cite]. One of those inferences is called the "contrastive inference" [ek: (sedivy 1999)]. Consider the context in Figure 1a that shows a yellow and brown banana, a yellow notebook and some other distractor item. When a listener is asked to *pick out the yellow...*, there are two eligible objects to choose from: the yellow banana and the yellow book. [ek: Sedivy 1999, 2003] shows that when there is a contrast to one of the objects (here, another banana), listeners rather fixate the yellow banana than the notebook, suggesting that there is a preference for the banana interpretation.

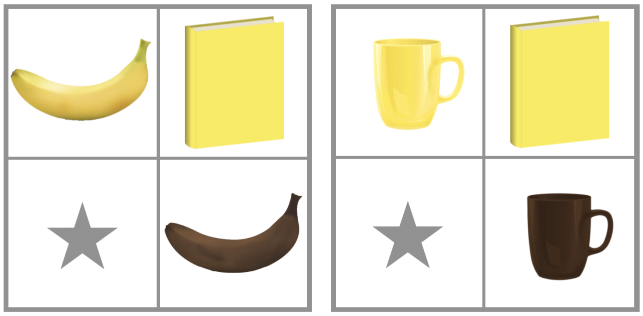


Figure 1: This is a figure. [ek: include subtitles]

Originally, this effect has been shown with size adjectives in eyetracking experiments [ek: Sedivy 1999] and has since been replicated reliably, especially in the scalar adjective domain [ek: cite]. However the effect seems to be less stable with color adjectives [ek: cite]. [ek: Sedivy 2003] reports that the contrastive inference arises in contexts where the target object has a "predictable" color (Figure 1a) but not when it has an "unpredictable" color (Figure 1b). [ek: jd: I'm careful here with the term "color diagnostic" here because that's not really how they define it. They just talk about predictability and proportion of modifier use in isolation.] They suggest that those objects differ in how likely a speaker is to produce the color modifier for the object in isolation, i.e., in the absence of a contrast, a yellow banana is usually just called a *banana* while a blue cup is still sometimes called a *blue cup*.

Since then inferences have been shown to be modulated by multiple factors, including adjective semantics [ek: cite], property salience [ek: cite], speaker reliability [ek: cite], and expectations of informativity [ek: cite Sedivy 2003].[jd: say a sentence about each of these, to introduce what you mean by all these different terms, but also to give a reader a sense for how much of an unsatisfying laundry list of features this comprises.]

[jd: "We provide a novel account of contrastive inferences that has the potential to unify all the above properties by reducing them to listeners' expectations about the speaker's contextual probability of producing the pre-nominal adjective. We couch this account within the Rational Speech Act framework... etc, leading into the next paragraph"]

Following recent research highlighting the importance of the listener's generative model of the speaker in generating pragmatic inferences [ek: cite], we propose the Rational-Speech Act (RSA) framework [ek: cite] as a new way to think about contrastive inference incrementally. In this framework, the listener reasons about a speaker's possible utterances, therefore giving the speaker model a central role in the predictions. It provides a way to quantitatively assess which predictions a listener with prior beliefs and expectations about the speaker *should* make. This shifts the focus away from specific cognitive and linguistic factors that influence contrastive inference onto listener's production expectations (and their prior beliefs). [jd: nice]

In this paper we will first show on qualitative examples how the RSA account can make the same predictions about the basic contrast effect as for instance the default description ([ek: Sedivy]) or the contrastive-adjective-function account [ek: cite]. We will then show new predictions about factors affecting contrastive inferences. To test the model, we report a production study to get modifier probability estimates.we then collect the data we want to model in a comprehension experiment using an incremental decision task (which will also serve as a proof of concept that contrastive inferences can be elicited in offline tasks); then model evaluation.

Model

[ek: RSA model here, formula; how we define the prior,... how is it incremental? NEW example predictions; talk about Westerbeek study]

[jd: introduce the two example contexts ttp and tap first and what different accounts predict for them (all except for RSA make the same prediction, namely the inference should arise). postulate production values for the RSA speaker model (take them from the empirically elicited values so people already have the right numbers in their head as they're reading)]

To investigate when a listener with a generative speaker model should draw a contrastive inference, we need to elicit how likely a listener can expect a modified over an unmodified referring expression for each object in the display. To evaluate the performance of the model and to gain information about the prior of the objects, we need comprehension data that informs us which object is considered the most likely target referent.

Experiment 1: Modifier Production in an Interactive Reference Game

To investigate whether production probabilities affect listeners' contrastive inferences, we need to manipulate how likely a listener is to observe a modified utterance for each object in the display.[jd: this first sentence should be in the previous paragraph. here just say "Exp. 1 was aimed at obtaining modifier production probabilities for all the displays ultimately used in the contrastive inference experiment (Exp. 2). We elicited these probabilities in a free production interactive reference game."] Our main assumption is that the typicality of a color for an object will affect these modifier production probabilities. When the object is in isolation for instance, a listener should expect the speaker to call a yellow banana simply *the banana*, but a blue banana *the blue banana*. In this production experiment, we tested this hypothesis and empirically elicited the proportion of color modifier mention for the target and competitor. The results are taken as the basis of what production probabilities a listener can expect to observe.

Participants

[jd: just flagging this section as one that can be radically shortened if you run out of space] We recruited 282 participants over Amazon Mechanical Turk, who were randomly matched to form one listener-speaker chat pair (i.e., 141 pairs in total). The estimated time for completion was 10 to 12 minutes and each participant was paid $2.30. We restricted participation to workers with IP addresses in the US and an approval rate of previous work above 97%.

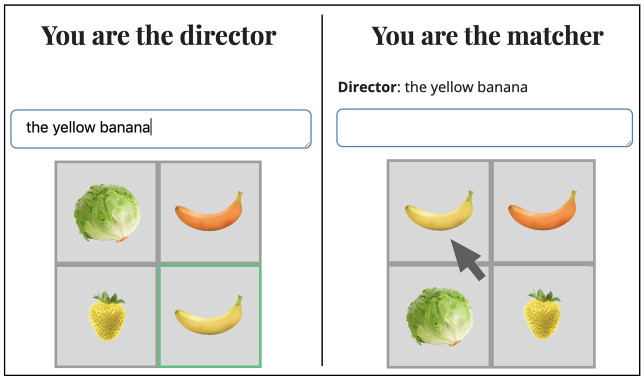
Exclusions were performed on the 141 speakers, since they provided the utterances. Participants were excluded when they participated multiple times in the experiment (1 participant; 139 pairs remaining) and when they did not use a noun from the display in at least half of the cases (27 participants; 112 pairs remaining). These participants clearly misunderstood the task, using expressions such as *yellow monkey* instead of *yellow banana*, or *should be yellow, must have teeth to eat* for *corn*. All speakers indicated that their native language was English.

Material

Each context included four items, as displayed in [ek: Figure]. The pool of items consisted of 10 types (banana, broccoli, carrot, corn, egg, lettuce, pumpkin, strawberry, swan, tomato), each of which could occur in a typical or atypical color. For example, the broccoli could occur in its typical color green or in the atypical color red. The resulting pool contains 20 items, 10 of which are atypically colored. The number of colors is carefully counterbalanced such as each color occurs twice as a typical and twice as an atypical instance. All items were carefully normed for color-diagnosticity [ek: Tanaka Presnell], typicality and nameability.

Design

The contexts varied in the typicality of the target, the typicality of the competitor and the presence of a contrast, resulting in eight conditions. For the critical trials, each participant saw four randomly created contexts from each of these eight conditions. In each condition we are interested in the modifier production probabilities for the target and the competitor, which is why in half of the trials when a contrast was present, the target was marked as the item to be communicated and in the other half, the competitor was marked. However when there is no contrast present, this distinction is irrelevant. For example, when target and competitor are both (a)typical, it



You are the director

the yellow banana

You are the matcher

Director: the yellow banana

Figure 2: This is a figure. [ek: make text bigger in figure]

is irrelevant, which is underlyingly coded as the target. Similarly when the modifier production probability for a typical target in context with an atypical competitor is the same as the probability for a typical competitor in a context with an atypical target. The fillers were eight randomly created contexts where the contrast was the item to be communicated and 20 randomly created contexts where the distractor was the item to be communicated. Overall, each participant saw 60 different contexts (32 critical trials) in a completely randomized order.

Procedure

Participants were randomly paired up and each was randomly assigned either to the role of a speaker or listener. They could communicate freely through a real-time multi-player interface similar to [ek: Hawkins (2015)]. The speaker was instructed to communicate a target object out of a four-object context to the listener. The target could be identified by a green border surrounding it. The speaker and the listener saw the same set of objects but in a randomized order to avoid trivial position-based references such as “the left one”. After the listener clicked on the presumed target, both the speaker and listener received feedback about whether the right object had been selected.

Results

[ek: only selected items with correct selection] Figure 4 shows the probability of color modifier mention for the target and competitor in each condition[1].

When a contrast to the target is present (e.g., another banana), a speaker needs to include the color modifier to fully disambiguate the two items (see the upper row in Figure 4). When the typicality of the target is atypical, this is completely borne out in the data. However if the target is typical, participants sometimes also used the unmodified utterance. [ek: refer to speculations about the reason for this in discussion]

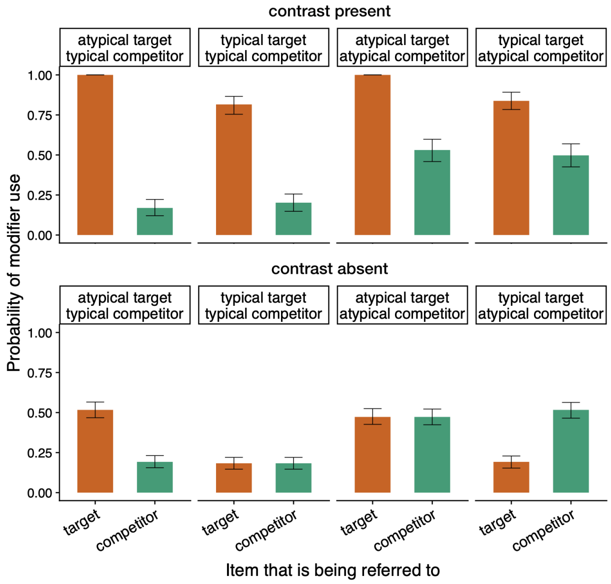


Figure 3: This is a figure. [ek: make text bigger in figure]

When the contrast is absent (see the lower row in Figure 4), speakers were more likely to include a color modifier when referring to an atypical target than a typical one [ek: stats?].

Independent of contrast, speakers were more likely to include the color modifier for an atypical color competitor over a typical one [ek: stats?].

The results of this production experiment show that the probability of a speaker’s modifier use is modulated by the color typicality of the item and the presence of a contrast. Our experiment therefore manipulates the modifier production probabilities a listener can expect in different contexts.

Comprehension Experiment: An Incremental Decision Task

To investigate which objects listeners consider to be the most likely target after observing the color adjective, we conducted an incremental decision task. [ek: ...]

Participants

We recruited 239 participants over Amazon’s Mechanical Turk, 121 of which saw atypical color competitors and 118 saw typical color competitors in the critical trials. The study took on average 7 minutes and each of them were paid $1.80 for their participation. We restricted participation to workers with IP addresses in the US and an approval rate of previous work above 97%.

We excluded participants who did the Hit multiple times (1) who indicated that they did the Hit incorrectly or were confused (13), who indicated that they had a native language other than English (6), and who gave more then 20% erroneous responses (7). An erroneous response is defined as a

[1] Note that some data is duplicated in the conditions where the contrast is absent ([ek: see section...for explanation]). In the conditions where target and competitor are both (a)typical, the modifier probabilities are created by the same data. The underlying data is also identical in the two conditions where one of the items is (a)typical.

click to a non-target object after observing the fully disambiguating noun, i.e., participants are excluded who selected the wrong final object more than 11 times. Overall, we excluded 27 people, which is 11% of the subjects. 211 participants remain, 108 of which were in the atypical competitor and 103 were in the typical competitor condition.

Material

The item pool is the same as described in the production study.

Design

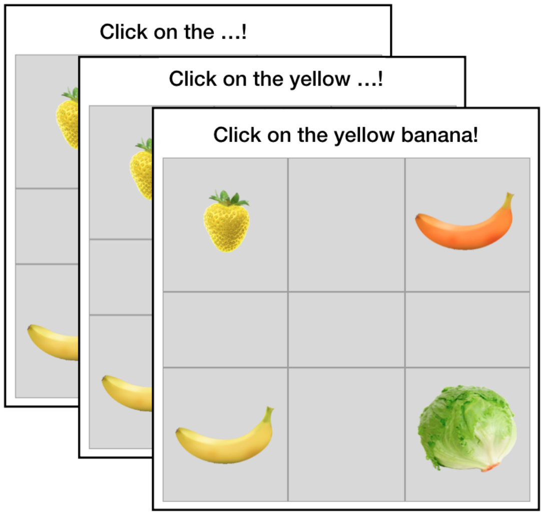
[ek: clarify what is within and between-subject manipulation with rationale]

Participants completed 55 trials in total, 20 of which were critical trials and 35 were fillers. The contexts varied for each participant with respect to the presence of a contrast and the target's color typicality (within-subject manipulation). Participants were randomly assigned to see either typical or atypical competitors on critical trials (between-subject manipulation). All critical trials included color modified utterances. To avoid that participants learn that the color modifier is always part of the referring expression, we need filler trials with unmodified referring expression. Another confound could be that participants can derive the target already from the context without seeing a referring expression, since the target is the only object that shares its color and type features with other objects in the display. The filler trials therefore needed to introduce primarily unmodified referring expressions that target other objects in the display. The exact trial structure is summarized in table 1.

Procedure

This experiment is a one-player adaptation of the production study explained above and follows the design of an incremental decision task [ek: cite Qing].

All participant were assigned the role of the listener, which means that they needed to identify which object was the target given a referring expression placed above the context. Crucially, the referring expression was only gradually revealed and participants had to choose an object each time. The choices had to be made prior any information about the target (i.e, after observing "Click on the"), after observing an adjective ("Click on the yellow") and after observing the full referring expression with the disambiguating noun ("Click on the yellow banana!"). The selections a listener makes before receiving any information about the target, inform us about potential priors they might have. The critical selections are the clicks after observing the adjective but before the disambiguating noun, since they will inform us about the inferences participants seem to draw from observing the adjective. We also collected the clicks after the fully disambiguating noun was presented, which functioned as attention checks in the analysis.


Click on the ...!

Click on the yellow ...!

Click on the yellow banana!

Figure 4: This is a figure.

Table 1: Overview of the trial structure for the comprehension study.

trial type	number	utterance	referent
critical	20	modified	target
filler	5	unmodified	competitor (typical)
filler	5	modified	competitor (atypical)
filler	5	modified	contrast
filler	20	unmodified	distractor (typical)

To center the position of the mouse after each selection, a button appeared in the center of the grid which had to be clicked to reveal the next word or to advance to the next trial.

Before participants proceeded to the main trials, they had to complete four practice trials constructed from the speaker perspective. In the speaker role, they saw contexts with four non-color diagnostic objects, one of which was marked as the target by a green border surrounding it. They were then asked to refer to the object such that a second player could identify it. The practice trials were introduced to familiarize the participants with the task.

The main trials were randomized with one restriction: Trials in which a speaker's color modifier use to refer to the target is very unlikely only occurred after the 15th trial. These were contexts where there was no contrast and either both target and color competitor were typical objects, or the target was typical while the color competitor was atypical. We introduced this restriction to minimize the risk that participants perceive the "utterance generator" (speaker) as unnatural.

Results

Figure 5 shows the proportion of clicks on the objects in the display prior observing the adjective (lighter colors) and after the adjective (darker colors), grouped by context condition. Before an adjective is observed, all items should appear equally likely to be the target, which is supported by

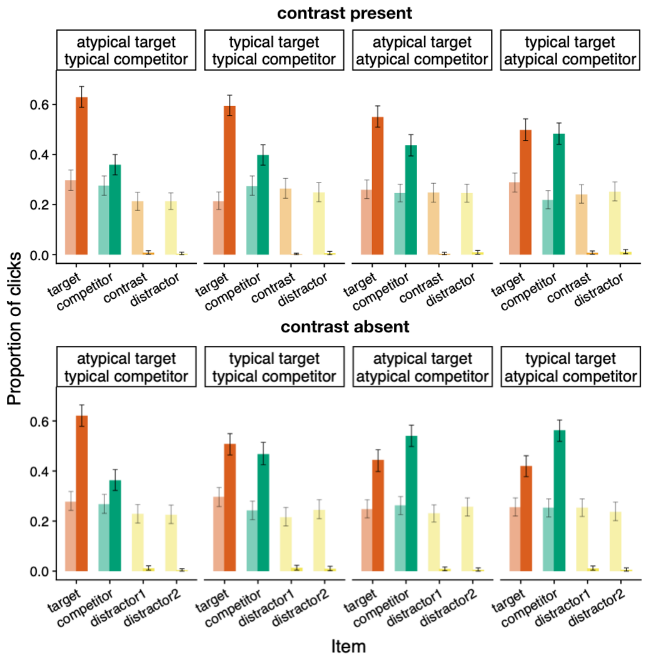


Figure 5: This is a figure. [ek: make bars and text bigger]

the generally uniform distribution in all conditions. After the adjective is observed (darker colors), only the target and competitor are legible options and we predict that the presence of the contrast and the typicality of the objects will affect the listeners' object choices.

When the contrast is present (upper row in Figure 5), there is a general preference for target selection over competitor selection. This preference is biggest for the case when the target is atypical and the competitor is typical and disappears for when the target is typical and the competitor is atypical.

When the contrast is absent (lower row in Figure 5) and target and competitor differ in typicality, there is a preference for the item with the atypical color. When the two items share their typicality, the selection are approximately at the same rate.

These results clearly show that the color typicality of the objects in the display affect the inference listeners draw about the target. We will turn to how these results correspond to the production data by considering the RSA model predictions.

Model evaluation

To assess the relationship between the modifier production probabilities and the comprehension data in the simplest way, we will assume a flat prior over all objects in the display. The probabilities to choose the target over the competitor are then simply the normalized modifier production probabilities as shown in Equation (1), where r is the possible referent, u the utterance and C the specific context.

$$P_{L_1}(r|u,C) = \frac{P_{S_0}(u|r,C)}{P_{S_0}(u|r_{target},C) + P_{S_0}(u|r_{comp},C)} \tag{1}$$

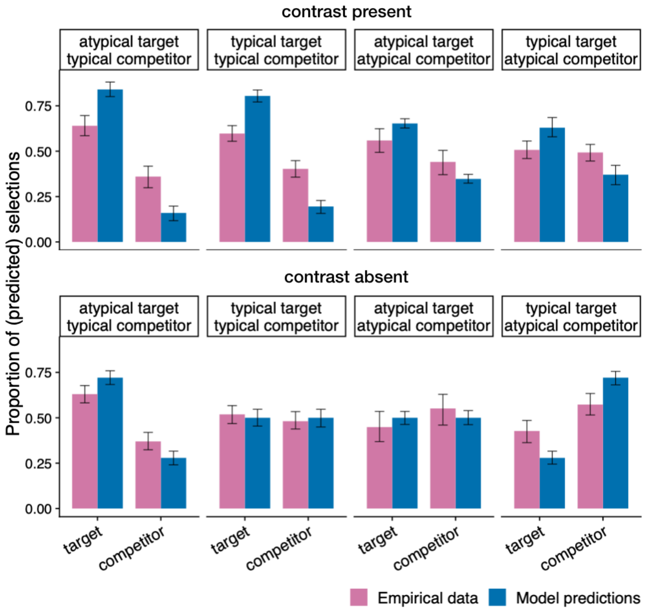


Figure 6: This is a figure.

Figure 6 shows the model predictions (in blue) and the empirical results (in purple) for target and competitor selection after observing the adjective.

The model predicts...

The model qualitatively predicts the patterns for the different context conditions. However, it generally predicts a stronger inference than borne out in the empirical data. [ek: which is why the correlation is so low]

[ek: the rsa model is a significant predictor for the comprehension data; however there are also biases: position and non-switching (see figure); in fact if we add these predictors to the model, they also become significant. In the new plots, we can see that (when the position is equal), whatever was previously clicked highly affects the outcome. Now the model underpredicts the target clicks in most conditions. The reverse is true if the competitor was clicked previously. This is something where the eyetracking data will provide more information since we expect the switching cost to be lower.]

Discussion

[jd: somewhere make clear that it's not just color-diagnosticity of the target (as sedivy suggests) that matters – instead, it's that, and the color-diagnosticity of the other distractors, and their relative typicality, etc etc, which are all ultimately captured variable modifier production expectations, and *that's* the explanatory quantity]

References

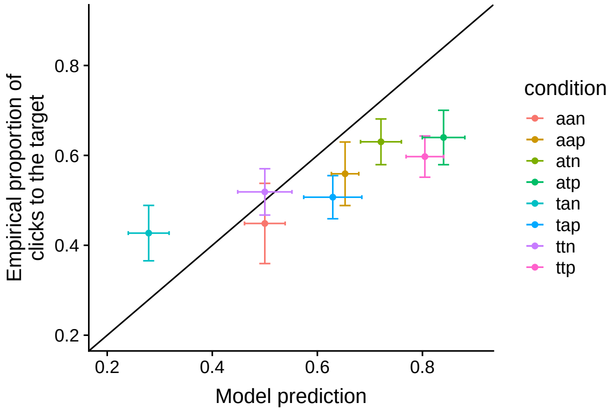


Figure 7: This is a figure.[ek: fix condition naming][jd: don't say "clicks to the target" but instead "target selection" on the y-axis and throughout the paper. clicks sounds clumsy. maybe also highlight the ttp and tap conditions visually if you use them as the motivating example at the beginning of the paper]

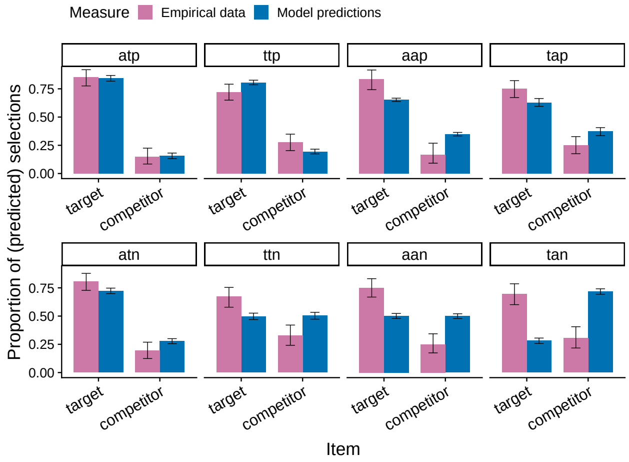


Figure 8: This is a figure.[ek: fix a lot]

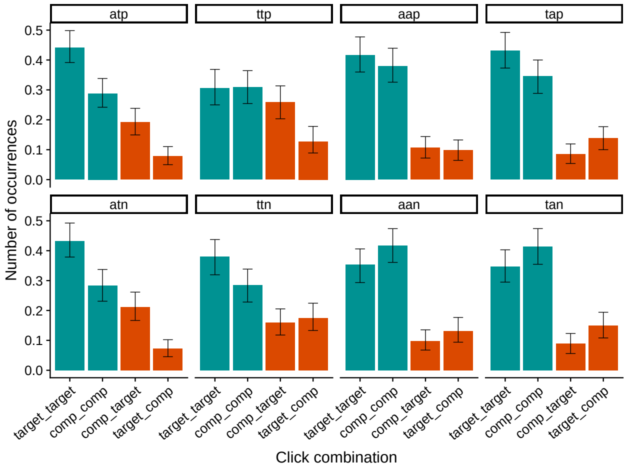


Figure 9: This is a figure.[ek: fix a lot]