Context, conflict, and the time-course of interpreting irony

Rachel Adler,[1,2] Lauren K. Salig,[1,2] Jared M. Novick,[1,2,3] & Yi Ting Huang[1,2,3]

1. University of Maryland College Park, Department of Hearing and Speech Sciences

2. University of Maryland College Park, Program in Neuroscience and Cognitive Science

3. University of Maryland College Park, Maryland Language Science Center

Word count: 11,257 words

Corresponding Author:

Yi Ting Huang

Department of Hearing and Speech Sciences

University of Maryland College Park

0100 Samuel J. LeFrak Hall

College Park, MD 20742

Email: ythuang1@umd.edu

Abstract

Irony requires listeners to infer that speakers mean the opposite of what they say (e.g., “What a fabulous chef he is” to imply that he cooks poorly), but it remains unclear what interpretative algorithms enable listeners to arrive at stable yet flexible meanings during comprehension. Across four experiments, listeners heard various speakers describe two characters while their eye-movements were measured to these referents. Afterwards, listeners provided judgments about the speaker’s intent. Eye-movements revealed frequency effects for literal but not ironic meanings, systematic reference restriction for ironic but not “opposite” speakers, and late-emerging conflicts between literal and ironic meanings. Judgments revealed intuitions about the pragmatic function of irony and relations to truth conditions. These findings demonstrate that ironic interpretations are not directly retrieved from conventionalized representations in the lexicon. Instead, they involve real-time reasoning about utterance semantics and speaker intentions, and that these dual processes yield distinct signatures during comprehension.

Keywords: *Irony, pragmatics, social communication, eye-tracking, cognitive control*

1. Introduction

During communication, speakers sometimes say the opposite of what they mean. For example, if you heard someone proclaim, “What a fabulous chef she is!” alongside a pile of half-baked batter, you might realize that they were being ironic. Speakers use verbal irony for wide-ranging pragmatic functions such as expressing humor, sharing perspectives, establishing mutual knowledge, and increasing social affiliation (Averbeck & Hample, 2008; Gibbs, 2000; Jorgensen, 1996). Since irony conveys information that goes beyond what is said, listeners must recruit extralinguistic context to guide interpretations, and draw from cues relating to politeness (Colston, 1997; Ivanko & Pexman, 2003), speaker knowledge (Ivanko & Pexman, 2003; Katz et al., 2004; Pexman & Olineck, 2002; Pexman et al., 2000), discourse context (Filik et al., 2014; Giora et al., 1998; Giora et al., 2015), and predicted meanings (Olkoniemi et al., 2016; Olkoniemi et al., 2019; Spotorno & Noveck, 2014). As a well-studied phenomenon, irony has inspired many accounts that explain elements of interpretation, including requisite context cues, processes for accessing interpretations, social and emotional reactions, inter alia (e.g., Echoic mention theory - Gibbs, 2000; Jorgensen, 1996; Pretense theory - Clark & Gerrig, 1984; Tinge hypothesis - Dews et al., 1995; Graded salience hypothesis - Giora & Fein, 1999, Filik et al., 2014; Literal-first - Dews & Winner, 1999; Schwoebel et al., 2000; Direct access view - Gibbs, 1986; Retention/suppression hypothesis - Giora & Fein, 1999; Defaulted hypothesis - Giora et al., 2015; Giora et al., 2018).

Importantly, the goal of language comprehension is to access a speaker’s intent, not to diagnose whether an utterance is ironic. Hence, if irony arises as a *product* of a mental architecture for generalized language comprehension – rather than a modular system with specialized inputs – it becomes essential to describe a range of general-purpose, interpretive algorithms that generate stable yet flexible meanings during communication. Drawing inspiration from psycholinguistic theories, the current study tests the hypothesis that irony emerges from real-time reasoning about utterance semantics and speaker intentions, and these dual processes yield distinct signatures during comprehension. We test our hypothesis by assessing listeners’ reasoning about utterance meanings and speaker properties in the following ways. Experiment 1 compares the timing of reference resolution for utterances produced by Literal Lucy (who speaks literally) vs. Ironic Ike (who speaks ironically) to evaluate whether frequency effects for irony are akin to those found for literal meanings. Experiment 2 replaces Ike with Opposite Ollie (who speaks "in opposites") and compares access to meanings with the same truth conditions but without clear pragmatic implications. Experiment 3 manipulates the cross-task adaptation of cognitive control to assess the timing of competition between literal and ironic meanings. Finally, Experiment 4 collects off-line judgments about how inferences about utterance meanings relate to speaker attributions.

To preview our results, we demonstrate that (1) listeners exhibit frequency effects for literal but not ironic meanings; (2) interpretive delays are highly systematic for ironic but not opposite speakers; (3) conflict between literal and ironic meanings arises late in interpretation; and (4) listeners generate specific inferences about speakers' intent based on irony usage. Critically, these findings contrast with prominent characterizations of irony as conventionalized in the lexicon (cf. indirect request, idioms) and directly retrievable during comprehension (Gibbs, 1986; Giora & Fein, 1999). Instead, they suggest that irony arises from listeners' top-down reasoning about known characteristics of speakers, which rapidly update in zero-shot learning contexts (Arnold et al., 2007; Fairchild & Papafragou, 2018; Fairchild et al., 2020; Grodner & Sedivy, 2011; Van Berkum et al., 2008). Nevertheless, listeners' ability to leverage knowledge about speakers to predict real-time utterance meanings is strikingly slow and insensitive to usage statistics (Huang & Snedeker, 2009, 2011, 2018; Gardner et al., 2021; Pogue et al., 2016; Ryskin et al., 2019). These dual algorithms – fast reasoning about speakers, slow predictions about utterances – may provide the basis for stable and flexible interpretations in the case of irony and beyond.

2. Experiment 1: Does irony exhibit frequency effects?

To describe the interpretive algorithms that give rise to irony, Experiment 1 examines how usage frequency interacts with current context to influence access to meanings during comprehension. Our logic is based on research on homophones, which has found that the relative frequency of meanings associated with a phonological form (e.g., how often "pitcher" relates to drink vs. baseball in corpora) influences reading times during sentence comprehension (Duffy et al., 1988; Rayner et al., 1994). Homophones provide a useful analogy for understanding access to irony. Theories of irony often distinguish between the use of positive statements to *critique* negative situations (e.g., “What a fabulous chef he is” referring to Fred’s failure in the kitchen) versus negative statements to *compliment* positive events (e.g., “What an awful chef she is” refers to Sally’s beautiful cakes). Compared to ironic compliments, ironic critiques are more frequent, remembered better, and require less context to interpret (Colston, 1997; Ivanko & Pexman, 2003; Bruntsch & Ruch, 2017; Caffarra et al., 2019; Climie & Pexman, 2008; Dews & Winner, 1999; Gibbs, 2000; Kowatch et al., 2013; Pexman & Olineck, 2002; Schwoebel et al., 2000).

Based on the analogy to homophones, Figure 1 illustrates two hypotheses about how irony might be accessed during comprehension. One possibility is that similar to homophones (Direct access), the lexicon routinely links multiple meanings to the same phonological form for irony (Gibbs, 1986; Giora & Fein, 1999). For example, adjectives like “fabulous” may be associated with a positive literal meaning (blue lines) and a negative ironic meaning (red lines). Moreover, since ironic critiques are more frequent in communication than ironic compliments, the relative activation of this meaning may be greater than their less conventional counterpart. However, another possibility is that unlike homophones, the lexicon regularly stores usage statistics for literal meanings, but *not* for ironic ones (Literal first). To comprehend irony, listeners must implement separable algorithms for generating utterance meanings and modify these on the basis of speaker properties (blue line to green line).

INSERT FIGURE 1 ABOUT HERE

To distinguish these accounts, Experiment 1 sets up a visual-world eye-tracking experiment to measure the availability of literal and ironic meanings for reference restriction. At the start of the study, participants are introduced to two speakers’ communication styles. Literal Lucy is always literal when she speaks, and Ironic Ike is always ironic when he speaks. In each trial, a narrator first tells a vignette about the contrasting fortunes of Sally and Fred (Figure 2). Next, Lucy or Ike chimes in with opinions on the situation like in (1). As the sentence unfolds, we measure participants’ eye movements to Sally and Fred. Here, the grammatical gender of the pronoun disambiguates referents: “she” refers to Sally, “he” refers to Fred (Arnold et al., 2000). Critically, participants can anticipate the referent based on earlier information from the adjective and speaker identity. When Literal Lucy speaks, participants are expected to look at

Sally after they hear "fabulous" and Fred when they hear "terrible." In contrast, when Ironic Ike speaks, they are expected to look at Fred after they hear "fabulous" and Sally when they hear "terrible."

(1) What a fabulous/terrible chef s/he is!

INSERT FIGURE 2 ABOUT HERE

If ironic meanings are stored in the lexicon along with usage statistics (Direct access), listeners' eye-movements should reveal frequency effects for literal *and* ironic interpretations. For literal meanings, reference restriction should be faster after more frequent positive adjectives compared to less frequent negative adjectives (Huang & Snedeker, 2013; van Tiel & Pankratz, 2021). For ironic meanings, it should be faster for more frequent ironic critiques compared to less frequent ironic compliments. Critically, since ironic compliments are highly infrequent relative to all other meanings, its interpretive delay should yield an interaction between adjective valence (negative vs. positive adjectives) and sentence type (literal vs. ironic sentences). In contrast, if irony is computed in real time based on access to literal meanings (Literal first), we would expect frequency effects for literal meanings, but not for irony. This will delay reference restriction for both literal and ironic interpretations, and yield both main effects of adjective valence and sentence type but no interaction between the two.

2.1 Method

2.1.1 Participants

Thirty-five undergraduates from the University of Maryland (UMD) participated for $5 pay or course credit. One participant's data was not analyzed due to equipment malfunction. We excluded two participants with low task performance, as defined by less than 80% accuracy in one or more conditions. Data analysis was conducted over the remaining 32 participants. Across all experiments, participants reported high proficiency with English. The UMD Institutional Review Board approved all procedures.

2.1.2 Procedures

Participants sat in front of an EyeLink 1000 desktop eye tracker (SR Research) and were told that they would encounter several trials, each unfolding over two parts. First, they would hear a narrator tell brief vignettes about two characters: Fred and Sally. Next, they would hear one of two speakers (distinct from the narrator) describe Fred and Sally. One speaker, Literal Lucy, always means what she says. Every time she says something, she means it literally. For example, if it is raining outside, she might say, “Wow, what a terrible day!” The other speaker, Ironic Ike, has a quirky sense of humor. Every time he says something, he always means it ironically. For example, if it is raining outside, he might say, “Wow, what a beautiful day!” Since the speakers differed by gender, their voice quality reliably cued literal and ironic interpretations. Participants’ task was to select the character (Fred or Sally) that the speaker was talking about. They completed three practice trials (two literal, one ironic) to ensure understanding.

After the practice block, participants moved on to the critical trials. Each trial was divided into two parts. During the familiarization phase, participants heard brief vignettes about two characters who attempt the same task but with varying outcomes. The vignettes co-occurred with related images in the display (Figure 1). During the test phase, participants heard a target sentence produced by either Ironic Ike or Literal Lucy describing Fred or Sally. Participants used a computer mouse to select the character that was referenced. Once they did so, the trial ended, and the next trial began with a new vignette. Across trials, looks to display locations were sampled continuously (every 2ms) from the start of the test sentence uttered by Ironic Ike or Literal Lucy until character selection. Off-line sentence interpretation was measured by mouse clicks to a character after the target sentence.

2.1.3 Materials

The vignettes, displays, and sentences for critical items were based on a 2 (adjective valence: positive vs. negative) x 2 (character valence: positive vs. negative) x 2 (speaker gender: man vs. woman) design. Vignettes always featured one character with a successful, positive outcome and another with a failing, negative outcome (e.g., Sally baked a beautiful cake while Fred made a mess in the kitchen). Half the stories attributed success to Sally and half to Fred, which varied randomly across trials. For each story, target sentences were created like those in (1). Each sentence was spoken by a male or female speaker, which established whether the intended meaning was literal or ironic. This relation was also confirmed through the adjective valence (e.g., *fabulous* vs. *terrible*) and grammatical gender of the pronoun (e.g., *he* vs. *she*). When literal sentences were associated with a female speaker (e.g., Lucy), referents were always predicted by adjective semantics (e.g., *fabulous-she* to describe successful Sally; *terrible-he* to describe failing Fred). Conversely, when ironic sentences were associated with a male speaker (e.g., Ike), referents were always the opposite of adjective semantics (e.g., *terrible-she* to describe successful Sally; *fabulous- he* to describe failing Fred). Across all sentences, the identity of the target character was always resolved by pronoun gender (i.e., *he* or *she*), and this matched the sentence interpretation (i.e., literal or ironic).

Adjectives expressed extreme valences in meaning to fit the story contexts and ranged from 2- to 5-syllables in length to increase the period of ambiguity before pronoun onset. This yielded a set of 16 positive adjectives (i.e., excellent, responsible, wonderful, outstanding, amazing, gifted, extraordinary, magnificent, fantastic, incredible, remarkable, exemplary, terrific, impressive, upstanding, phenomenal) and 13 negative adjectives (i.e., terrible, horrible, awful, dreadful, incompetent, hopeless, worthless, abominable, self-centered, horrendous, atrocious, bumbling, coldhearted). We analyzed single-word occurrence counts per 1,000 words in written corpora of American English using Google n-gram statistics (Michel et al., 2011), and confirmed that positive adjectives (M = 8.8, SD = 7.6) were 5.2 times more frequent than negative adjectives (M = 1.6, SD = 2.7) (t = 3.45, p < .001). Adjectives were yoked together in opposite-valence pairs and combined with role terms that related to events in the vignettes (e.g., remarkable/hopeless golfer, magnificent/incompetent speaker).

Materials for each item (e.g., the cake-baking item) were generated by manipulating sentence interpretation and adjective valence. These were distributed across four presentation lists. In each list, 20 critical trials featured five items in each adjective valence x sentence type combination, and each base item appeared once in every list. They were randomized with 24 filler trials, where both characters were similarly successful or unsuccessful at a given task. Since target characters could not be unambiguously identified until pronoun onset, filler trials provided transparent reinforcement of speaker tendencies (i.e., Ironic Ike or Literal Lucy), independent of adjective interpretation. To ensure that prior expectations about irony and speaker gender did not affect the current study (e.g., males tend to be more ironic than females; Colston & Lee, 2004), four presentation lists were created to counterbalance speaker gender.

Half the participants heard a literal female speaker and an ironic male speaker (i.e., Literal Lucy and Ironic Ike). The other half heard a literal male speaker and an ironic female speaker (i.e., Literal Luke and Ironic Iris). Stimuli for all experiments are available at https://osf.io/4298j/.

The vignettes and target sentences were recorded by different speakers in a sound-attenuated room using a Shure SM-51 microphone. For target sentences, one male and one female actor were instructed to adopt an enthusiastic tone that was consistent with an ironic interpretation, but did not preclude a literal one. The sound files were checked to ensure that two regions were similar in length across conditions: (1) 598ms (on average) region from sentence onset to the adjective cue (e.g., "What a") and (2) 1296ms (on average) region from adjective onset to the onset of the pronoun (e.g., "fabulous chef"). Since prosody for irony varies widely across speakers and contexts (Attardo et al., 2003; Bryant & Fox Tree, 2005; Kreuz & Roberts, 1995; Rockwell, 2000), half the participants heard sentences paired with literal speakers and the other half heard the same stimuli paired with ironic ones across presentation lists. This ensured that the timing of interpretations across conditions could not be attributed to low-level differences in sound quality.

2.2 Coding

Approximately 0.63% of critical trials involved incorrectly selecting Distractor characters and were removed from further analysis. Eye movements in the critical trials were coded as looks to the two characters or missing due to looks away from these areas (e.g., other parts of the displays, blinking). Missing looks accounted for 6.29% of the eye-movement data. Next, looks and actions were re-coded based on relations to test sentences. Targets refer to the character who was described, and Distractors refer to the other character. To compare fixations across conditions, our primary dependent measure was Target preference, which was calculated as the proportion of Target looks divided by Target plus Distractor looks. If participants looked exclusively to the Target, then Target preference was 1. If they looked exclusively to the Distractor, then Target preference was 0. We calculated Target preference in 2-ms time windows throughout the sentence and used those values to calculate Target preference for larger regions that were time-locked to informative word onsets. Each region was shifted by 200ms to account for the timing of saccadic eye movements (Matin et al., 1993). The regions we analyze are:

- Pre-adjective: This region began at sentence onset and ended before adjective onset (e.g., "*What a…*"), and served as a baseline of Target preference prior to hearing informative linguistic cues.
- Adjective-Noun: This region began at adjective onset and ended before pronoun onset (e.g., "*…fabulous chef…*"). Relationships between adjective valence and sentence type in this region will reveal the extent to which irony frequency impacts access to irony.
- Pronoun: This region began at pronoun onset and ended at sentence completion (e.g., *"…he is"*). Since referents are fully disambiguated by the grammatical gender of the pronoun, it was expected that Target preference would be high across conditions.

2.3 Results

Figure 3 illustrates that prior to the adjective, Target preference was equivalent across conditions. This demonstrates that there was no preference for either character before the onset of relevant linguistic information. After pronoun onset, Target preference across conditions approached ceiling, suggesting that participants used the pronoun gender to restrict reference to appropriate characters. To confirm these observations, we analyzed Target preference with linear mixed-effects models using the lme4 package (version 1.1.23; Bates et al., 2015) in R (version 4.0.2; R Core Team, 2020). Our analytical strategy was based on best practices outlined in Barr et al. (2013), and included fixed and random effects that were justified by the study design. We adopted deviation coding to compare condition means to the grand mean, which increased the ease of interpreting main effects and interactions. Random intercepts for subjects and items were included. Parameter-specific p-values were estimated through Satterthwaite approximation (Luke, 2017). Data and analysis code are available at https://osf.io/4298j/.

INSERT FIGURE 3 ABOUT HERE

2.3.1 Validating the paradigm

Our goal in Experiment 1 was to evaluate the impacts of real-world usage statistics on sentence processing, but one potential concern was that participants would rapidly assimilate to the statistics of the current experimental task and interpret sentences based on the frequency of occurrence in the study. If this task strategy was the basis of their performance, participants should be equally fast at restricting reference for positive and negative adjectives since they occurred in comparable proportions in the study. They should also be equally fast to restrict reference to positive and negative characters since this too was balanced across trials. If, however, our paradigm successfully probed prior experiences, we should find two types of frequency effects. First, participants should be faster to access literal interpretations for positive compared to negative adjectives. Second, they should be faster to look at Targets when adjective valence matches character valence (literal sentence) compared to mismatched (ironic sentence).

To evaluate the extent to which usage frequency impacts current interpretations, we evaluated Target predictions based on adjective (positive vs. negative) and character valence (positive vs. negative). Before adjective onset, there were no significant main effects or interactions ($p > .60$). However, during the Adjective-Noun region, Target preference was greater for positive compared to negative adjectives, leading to a main effect of adjective valence ($t = 2.90$, $p < .01$). Likewise, in positive adjective sentences, Target preference was lower when referring to negative characters ironically compared to positive characters literally, and this appropriately reversed for negative adjectives. This led to a significant interaction between adjective and character valence ($t = 3.47$, $p < .001$). There was no additional effect of character valence ($p > .25$), suggesting that properties of the characters on their own were not driving fixations. During the Pronoun region, Target preference appeared to be greater for positive adjectives than negative adjectives, but this was not significant ($t = 1.68$, $p < .10$). There was also no effect of or interaction with character valence ($p > .40$).

2.3.2 Testing hypotheses

Next, to evaluate the extent to which usage frequency impacts access to irony, we evaluated Target predictions based on adjective valence (positive vs. negative) and sentence type (literal vs. ironic). Unlike analysis of character valence, evaluating Target predictions based on sentence type directly tests the extent to which the frequency advantage of ironic critiques leads to faster real-time access compared to ironic compliments. Figure 4 and Table 1 illustrate that before adjective onset, there were no significant main effects or interactions (p > .60). During the Adjective-Noun region, Target preference was greater for positive compared to negative adjectives, leading to a main effect of adjective valence. Target preference was greater for literal compared to ironic interpretations, leading to a main effect of sentence type. Importantly, there was no significant interaction between adjective valence and sentence type (p > .25), suggesting that real-world usage patterns for ironic interpretations did not impact current real-time reference predictions.[1] During the Pronoun region, Target preference continued to be greater for positive than negative adjectives, but this was not significant. There was no additional effect of, or interaction with, sentence type (p > .40).

INSERT FIGURE 4 AND TABLE 1 ABOUT HERE

Although we found irony delays were no different across irony type, the absence of evidence is not evidence of absence. One possibility is that participants came into the study with expectations about irony usage, but experience with usage statistics within the experiment taught them that ironic critiques and compliments were equally frequent in the current context. If so, then adjective valence and sentence type may interact at the start of the study (first 10 critical trials), but this interaction would disappear by the end of the study (last 10 critical trials). We did not find this in our analyses. First-half trials revealed a main effect of sentence type (t = 2.28, p < .05), but no additional effect of or interaction with adjective

[1] We conducted post-hoc power analyses to estimate the sample size required for detecting differences in Target preference in the Adjective-Noun region with 80% power. For adjective valence, we needed 28 participants to distinguish positive and negative adjectives ($\eta^2 = 0.24$). For sentence type, we needed 32 participants to distinguish literal and ironic sentences ($\eta^2 = 0.21$). For the interaction, however, we would need 191 participants to detect differences in sentence type by adjective valence ($\eta^2 = 0.04$). This suggests with 32 participants, the current study was adequately powered to detect effects of adjective valence and sentence type, but we would need roughly 6x more participants to detect the interaction between the two. While this does not rule out the possibility of directly accessing ironic interpretations, it suggests that this process (if it exists) differs in magnitude from accessing literal interpretations.

valence (p > .30). Thus, early in the study, participants already showed an advantage for processing literal sentences compared to ironic sentences. Second-half trials reveal main effects of both adjective valence (t = 3.23, p < .01) and sentence type (t = 2.50, p < .05), but no interaction between the two (p > .60). These patterns suggest that participants were not sensitive to irony frequency at any point during the experiment. If anything, greater experience with the current task led participants to become faster at predicting *literal* interpretations over the course of the study.

Another possibility is that effects of speaker gender obscured our ability to detect interactions between adjective valence and sentence type. Irony is more likely to be associated with male speakers compared to female speakers (Colston & Lee, 2004). Since we used speaker gender to cue interpretation type, it is possible that this introduced confounds that impacted the timing of interpretation. To assess this, we predicted Target preference with sentence type (literal vs. ironic) and speaker gender (male vs. female) as fixed-effects variables. No significant differences were found during the Pre-adjective region (p > .30). During the Adjective-Noun region, ironic sentences were accessed slower than literal ones, leading to a main effect of sentence type (t = 3.42, p < .001), but there was no additional effect of or interaction with speaker gender (p > .60). During the Pronoun region, Target preference was greater for literal compared to ironic sentences when ironic speakers were female, and this difference diminished when speakers were male. However, the interaction between sentence type and gender was not significant (t = 1.93, p > .05). To the extent that social knowledge influences irony interpretation, these effects appear to be small, and emerging after initial reference restriction.

2.3 Discussion

Experiment 1 tested the extent to which frequency effects for irony are akin to ones found with literal meanings. As expected, frequency influenced access to literal interpretations: positive adjectives were processed earlier than negative ones, and literal sentences were interpreted earlier than ironic ones. This was true even when the acoustics of speaker gender offered an early and unambiguous cue to sentence interpretation. Importantly, the absence of interaction between adjective valence and sentence type suggests that ironic critiques were processed as slowly as ironic compliments. Hence unlike literal interpretations, real-world usage frequency may not influence access to ironic interpretations. Follow-up analyses demonstrated that delays for irony were found in first-half trials and persisted into second-half trials, suggesting that the absence of a frequency effect did not reflect rapid learning of experiment-wide statistics. We found weak evidence that listeners may draw upon an association between irony and male speakers. However, these effects emerged after listeners had already restricted reference, suggesting that sensitivity to social information may arise as a consequence rather than a cause of ironic interpretation.

Our findings suggest that while irony links multiple meanings to a single phonological form, it does so through a different process from homophones. For homophones, storing lexical semantics may be algorithmically efficient since the current context distinguishes between stable and unrelated meanings. In contrast, the absence of frequency effects suggests that irony may instead be generated anew, based on current conversational demands. This asymmetry may reflect the degree of meaning flexibility for irony compared to homophones. While both phenomena draw from context for interpretation, ironic meanings tend to be far more idiosyncratic. Hence, directly accessing precompiled meanings, *even with strong situational cues*, may be insufficient for capturing the nuance of speaker goals and context (e.g., differing meaning implication for utterances produced by a close friend, comedian, stranger on Twitter). Moreover, since speakers employ irony to satisfy both informational (e.g., I know that you know that Fred is a failure) and social constraints (e.g., it is rude to point out Fred's shortcomings; it is clever to add humor to an unfortunate situation), listeners only fully access speakers' intent to the extent that they can appreciate the inherent relations between the multiple meanings.

However, a skeptical reader may quibble about the odd task that we developed, which required listeners to compute the opposite of a literal interpretation. Rather than assessing irony, we may instead be measuring listeners' sluggishness at performing an unusual task. Experiment 2 tests the extent to which interpreting irony is similar to accessing opposites. We replace Ironic Ike with a new speaker, Opposite Ollie, who always speaks in opposites. Like irony, computing the opposite meaning of an utterance's truth condition depends on initial semantic analysis. But unlike opposites, ironic interpretations may depend on distinct operations for semantic and pragmatic analysis (i.e., what is the speaker communicating here by being ironic?). If understanding opposites and ironic utterances are different in this way, then listeners' predictions of likely referents in Experiment 2 should differ from those in Experiment 1. In the absence of a pragmatic function, we may find that interpreting opposites does not generate systematic delays despite possessing overlapping truth-conditions as irony.

3. Experiment 2: Is irony like opposites?

3.1 Method

3.1.1 Participants

Thirty-seven UMD undergraduates participated for $5 pay or course credit. We excluded five participants with low task performance, and conducted analysis over the remaining 32 participants.

3.1.2 Procedures and Materials

The procedures and materials were identical to Experiment 1, except that participants were told that the story characters (Fred and Sally) would be described by either Literal Lucy or by another speaker, Opposite Ollie, who would always say the opposite of what was meant. Similar to Experiment 1, the gender of the speakers was counterbalanced across presentation lists (i.e., half the participants heard descriptions by Literal Luke and Opposite Olive).

3.2 Results

The data were analyzed in the manner described in Experiment 1. Missing looks accounted for 8.93% of fixations. Approximately 1.10% of trials involved incorrectly selecting Distractors and were removed from further analysis. Figure 5 illustrates that Target preference appeared to be similar across conditions before adjective onset and approached ceiling after pronoun onset. However, there was no difference between Target preference for opposite vs. literal sentences in the Adjective-Noun region. Comparisons of adjective and character valence offered converging evidence that opposites were interpreted differently than irony. Unlike Experiment 1, we found no differences or interactions across all regions ($p > .10$). Since Experiment 2 adopted the same materials as Experiment 1, this suggests that manipulating the speaker's communicative tendencies (using irony vs. simply saying the opposite of what was meant) yielded strikingly different processing patterns.

INSERT FIGURE 5 ABOUT HERE

Next, we tested the extent to which Target predictions varied with adjective valence and sentence type (literal vs. opposites). Figure 6 and Table 2 illustrate that prior to adjective onset, there were no significant main effects or interactions (p > .10). During the Adjective-Noun region, Target preference did not significantly differ across conditions (p > .10), despite numerical advantages for positive compared to negative adjectives, and for literal compared to opposite statements. There were no significant main effects or interaction during the Pronoun region (p > .20). Follow-up analyses revealed no significant differences in first-half and second-half trials (p > .10). Finally, we predicted Target preference with sentence type (literal vs. opposite) and speaker gender (male vs. female) as fixed-effects variables. We found no effect of speaker gender or interaction with sentence type during the critical Adjective-Noun region (p > .10). However, in the Pre-adjective region, Target preference was unexpectedly greater for opposite compared to literal sentences when speakers were male, and this pattern reversed when speakers were female. This led to a significant interaction between sentence type and speaker gender (t = 2.28, p < .05). A similar pattern emerged in the Pronoun region, but did not approach significance (t = 1.72, p > .05). Together, this demonstrates that participants were sensitive to speaker differences, but these effects appeared to be unrelated to their adjective interpretation.

INSERT FIGURE 6 AND TABLE 2 ABOUT HERE

Since interpreting opposites appeared to cause widespread interference in sentence interpretation, we evaluated whether participants in the current task at the very least, updated their Target preferences as they heard more linguistic information. For positive adjectives, we found that Target preference in the Adjective-Noun region was sensibly greater than the Pre-adjective region (t = 2.76, p < .01) and lower than in the Pronoun region (t = 10.89, p < .001). Likewise, for negative adjectives, Target preference in the Adjective-Noun region was greater than the Pre-adjective region (t = 3.32, p < .001) and was lower than in the Pronoun region (t = 11.65, p < .001). This confirms that participants incrementally updated interpretations with incoming words.

3.3 Discussion

Experiment 2 examined the time-course of interpreting opposites and found strikingly different profiles of reference restriction compared to Experiment 1. Recall that real-time processing in Experiment 1 was influenced by real-world usage patterns of literal meanings, leading to faster reference restriction for positive compared to negative adjectives and for literal compared to ironic interpretations. In contrast, in Experiment 2, we found no effects of adjective and sentence type across any time region. Instead, it appeared that listeners may be negotiating interpretive algorithms on the fly, leading to a disregard for relevant linguistic information (e.g., adjective meanings), slowness in updating literal interpretations, and overreliance on speaker gender in lieu of linguistic cues. The lack of systematicity in reference restriction for opposites contrasts with patterns of reference restriction for irony, and suggests that delays for irony go beyond calculating truth conditions. Together, these findings highlight a clear contrast: while opposites can be resolved without engaging additional inference, irony uniquely recruits pragmatic processes that are distinct from accessing literal meanings. This raises questions of what cognitive mechanisms support the generation of these inferences.

In Experiment 3, we explore the extra-linguistic, cognitive processes involved in interpreting ironic sentences. Theories of irony suggest that conveying pragmatic interpretations depends on the listener's ability to appreciate the mismatch between the literal meaning of an utterance and the current context in which the utterance is spoken (e.g., Clark & Gerrig, 1984; Gibbs, 2000). Doing so might require listeners to simultaneously consider multiple possible interpretations and select the most situationally appropriate one, which could lead to a momentary conflict as they navigate between literal and ironic meanings. Namely, the literal interpretation is licensed by the syntactic and semantic cues in the bottom-up input, but this meaning is incompatible with the top-down context cues that support ironic interpretations (e.g., broader discourse, identity of the speaker, speaker's likely intentions). If this is the case, Experiment 3 will directly test whether listeners experience measurable conflict between literal and ironic meanings during comprehension.

4. Experiment 3: Do literal and ironic interpretations conflict during comprehension?

One method of examining interpretative conflicts involves manipulating an individual's state of cognitive control during language processing. This is typically done by having participants perform trials of a non-language cognitive control task, such as the Stroop or flanker task, and observing the immediate effects on language processing when there is more than one strongly supported representation of the input (Hsu & Novick, 2016; Hsu et al., 2021; Ovans et al., 2022; Thothathiri et al., 2018; for a review, see Ness et al., 2025). The premise is that encountering conflict on an incongruent Stroop trial (e.g., the word "blue" displayed in red ink) upregulates cognitive control, which sustains for long enough to facilitate performance on a subsequent language trial that involves conflict. This phenomenon, referred to as *cross-task adaptation of cognitive control* (Ness et al., 2025), highlights how conflict in one task engages cognitive control, which modulates the effects of linguistic conflict in another.

In syntactic parsing, conflict can occur when early processing decisions must be revised due to conflicting evidence. Consider the instruction, "Put the apple on the napkin onto the box," in which the phrase "on the napkin" is initially ambiguous, potentially specifying the goal (where the apple should go) or a modifier (providing more information about the to-be-moved apple). Eye-tracking studies using the visual-world paradigm show that listeners initially misinterpret the phrase as the goal and then switch to the modifier interpretation when later-arriving, conflicting evidence like "onto the box" arrives (Novick et al., 2008; Spivey et al., 2002; Tanenhaus et al., 1995; Trueswell et al., 1999). Hsu and Novick (2016) found that manipulating cognitive control through the Stroop task affected listeners' ability to revise their early misanalysis, with increased control following incongruent Stroop trials leading to faster and more accurate resolution of syntactic ambiguity. This was indexed by earlier looks to the correct goal, and suggests that cognitive-control systems resolve conflicts between two incompatible interpretations of an ambiguous sentence (Kim et al., 2025; Ness et al., 2024; Ness et al., 2025; Novick et al., 2005; Novick et al., 2010; see also Hsu et al., 2021; Ovans et al., 2022; Thothathiri et al., 2018).

Building on our understanding of how cognitive control impacts syntactic ambiguity, we examine whether comprehending irony involves conflict between literal and ironic interpretations. Experiment 3a evaluates if there is early conflict between the literal and ironic interpretations by manipulating the status of listeners’ cognitive control through Stroop trials *before* engaging in a language task involving ironic sentences. If both meanings are immediately available during comprehension, then upregulating cognitive control through prior Stroop-incongruent trials should facilitate resolution, much like it does during syntactic-ambiguity resolution. Experiment 3b reverses the direction of trial sequences: listeners process language first and then perform the Stroop task. This design tests the alternative hypothesis that conflict associated with irony processing arises later, after listeners first pursue the literal interpretation. If conflict between literal and ironic interpretations emerges late, then cognitive control should engage as listeners switch from processing literal to ironic meanings, which should affect performance on the subsequent Stroop task. We will report effects that are consistent with this notion. Because this conflict arises late, we will suggest that this pattern provides converging evidence that the pragmatic inferences for irony are generated in real time and are not pre-compiled.

4.1 Experiment 3a

4.1.1 Participants

Forty-five UMD undergraduates participated for $5 pay or course credit. Data from one participant was excluded due to excessive track loss, and four more due to experimenter error. We excluded seven participants based on their low performance in sentence trials and five participants based on their performance on Stroop trials. We analyzed data from the remaining 28 participants.

4.1.2 Procedures

Participants were told that they would complete the Stroop task and a sentence-interpretation task presented in random order. The sentence trials involved the same instructions about stories and speakers as Experiment 1. To preserve the window within which cross-task adaptation of cognitive control can be observed, we adopted the same timing scheme as Hsu and Novick (2016) (see also Hsu et al., 2021; Kan et al., 2013). The Stroop trials (trial n-1) asked participants to indicate the font color of color names on the screen using a 3-button mouse that corresponded to each color, and to do so as quickly and accurately as possible. Each trial began with a 500ms fixation cross, which was then replaced with a Stroop or sentence trial. All trials were followed by a 1000ms inter-trial interval. The color names remained on the screen until participants responded or 1000ms had passed, whichever occurred first. On sentence trials (trial n), participants heard ironic or literal sentences as we recorded their eye movements to the displays depicting Experiment 1 events. To minimize time between the Stroop trials and the sentence trials, participants did not hear the vignettes in Experiment 3, but character valence was disambiguated based on the display picture (Figure 1). Before starting the critical sequences, participants practiced three sentence trials (two literal, one ironic) and 144 Stroop trials.

4.1.3 Materials

The experiment employed a 2 x 2 x 2 design, with previous Stroop type (congruent vs. incongruent) and current sentence interpretation (literal vs. ironic) as within-subjects factors, and speaker gender counterbalanced as a between-subjects factor. Stroop trials were pseudo-randomly interleaved with sentence trials, leading to four critical sequences: $\text{Stroop}_{\text{congruent}}$-to-$\text{Sentence}_{\text{literal}}$; $\text{Stroop}_{\text{congruent}}$-to-$\text{Sentence}_{\text{ironic}}$; $\text{Stroop}_{\text{incongruent}}$-to-$\text{Sentence}_{\text{literal}}$; $\text{Stroop}_{\text{incongruent}}$-to-$\text{Sentence}_{\text{ironic}}$.

For the Stroop trials, congruent trials involved color names that matched the font color (i.e., blue printed in blue ink, yellow in yellow ink, green in green ink), and incongruent trials involved color names that mismatched the font color (i.e., the words *orange*, *brown*, and *red* printed in blue, yellow, or green ink). For the sentence trials, two versions of each critical item were created by manipulating the interpretation of positive adjectives only (i.e., ironic vs. literal). In total, participants performed 192 experimental trials: 96 Stroop trials and 96 sentence-interpretation trials. To prevent predictions of subsequent trial type, the 48 critical Stroop-sentence sequences were pseudo-randomly mixed with 48 filler Stroop and 48 filler sentence trials. Stroop trials were as likely to precede Stroop trials as sentence trials, and sentence trials were as likely to precede Stroop trials as sentence trials. Filler sentence trials recruited negative adjectives and mentioned both characters to prevent prediction of adjective valence across trials. Half of the trials corresponded to 12 tokens of the four critical-sequence conditions, and the other half corresponded to filler trials that decreased the predictability of sequences.

4.1.3 Results and Discussion

To validate that Stroop trials engaged cognitive control (trial n-1), we compared reaction times (RTs) in the congruent vs. incongruent trials. Approximately 4.6% of trials were excluded from analyses due to no response. To minimize the impact of outliers, we log-transformed raw values and excluded values that were two standard deviations greater than the mean. This omitted 1.8% of trials. RTs were analyzed through linear mixed-effects models, with subjects as random intercepts. RTs were significantly longer in incongruent trials than congruent trials (t = 4.28, p < .001), suggesting that they engendered conflict, as expected.

INSERT FIGURE 7 AND TABLE 3 ABOUT HERE

Next, we examined effects of cognitive-control engagement on sentence interpretation by comparing Target preference over time regions of interest (trial n) and how this varied with the preceding Stroop task (trial n-1). Eye movements were analyzed in the same manner as in Experiment 1. Missing looks accounted for 7.4% of fixations. Figure 7 and Table 3 illustrate that fixations were near chance during the Pre-adjective region and converged on correct referents in the Pronoun region. There was no effect or interaction with previous Stroop trials in the Pre-adjective region (p's > .15). Importantly, during the Adjective-Noun region, Target preference was significantly greater for literal interpretations compared to ironic ones, leading to a main effect of current sentence type. There was no effect or interaction with previous Stroop trials (p's > .10). During the Pronoun region, Target preference was numerically but not significantly greater when sentence trials were preceded by incongruent compared to congruent Stroop. There was no effect or interaction with current sentence trials (p's > .30).

In Experiment 3a, we replicated the patterns observed in Experiment 1, where participants were delayed in interpreting ironic compared to literal sentences. To determine whether increased cognitive control would attenuate the delay, we manipulated engagement status through preceding Stroop trials. However, we did not observe such an interaction, which contrasts with studies demonstrating the impact of cognitive control on syntactic ambiguity resolution when multiple meanings are available early in interpretation (Ness et al., 2025; Novick et al., 2005; Novick et al., 2010; Hsu et al., 2021; Ovans et al., 2022; Thothathiri et al., 2018). In Experiment 3b, we explored whether accessing the ironic interpretation might generate *late* conflict with the literal meaning, after listeners first pursue initial semantic analysis. If this is the case, cognitive-control engagement should not impact the time-course of comprehending irony, as only the literal interpretation is available early in comprehension. Instead, the conflict between literal and ironic interpretations may arise downstream, after semantic analysis serves as a basis for deducing relevant cues for generating irony.

4.2 Experiment 3b

4.2.1 Participants

Forty-two UMD undergraduates participated for $5 pay or course credit. One participant was excluded for having participated in a similar experiment, and nine more due to experimenter error. We excluded five participants with low task performance on sentence trials and four more with low task performance on Stroop trials. Data analysis was conducted over the remaining 23 participants.

4.2.2 Procedures and Materials

The procedures and materials were identical to Experiment 3a, except that task order for critical sequences was reversed. On half the trials, interpreting simple literal sentences (trial n-1) should not engage cognitive control, and this establishes a baseline for performance on congruent or incongruent Stroop (trial n). On the other half, interpreting ironic sentence trials (trial n-1) is hypothesized to increase cognitive-control engagement, albeit late (after initial literal semantic analysis), and we tested whether there would be effects on the Stroop task (trial n). Similar to Experiment 3a, participants performed 192 total trials: 96 Stroop trials and 96 sentence-interpretation trials. Half of the trials corresponded to 12 tokens of the four critical-sequence conditions, and the other half corresponded to filler trials that decreased the predictability of sequences.

4.2.3 Results

The data were analyzed in a manner similar to Experiment 3a. We first examined Target preference by sentence interpretation for each time region, to establish whether listeners experienced the same delay in interpreting irony compared to literal sentences. Missing looks accounted for 8.2% of fixations. Figure 8 illustrates that Target preference was around chance in the Pre-adjective region and converged on referents in the Pronoun region. During the Adjective-Noun region, Target preference was greater for literal interpretations compared to ironic ones, leading to a main effect of sentence type as in Experiments 1 and 3a ($t = 2.11$, $p < .05$). In contrast, effects of sentence type were absent in the Pre-adjective and Pronoun regions, as expected (p's > .15).

Next, we examined the effects of interpreting ironic sentences (trial n-1) on the Stroop task (trial n) by comparing RTs on congruent and incongruent trials following sentence types on the previous trial (trial n-1). Approximately 3.4% of trials were excluded from analyses due to no response. Another 2.8% of trials were excluded as outliers. Figure 9 and Table 4 illustrate that RTs were faster on congruent compared to incongruent trials, leading to a main effect of current Stroop type ($t = 3.60$, $p < .01$). However, there was no additional effect of or interaction with previous sentence type (p's > .50).

INSERT FIGURE 8 AND FIGURE 9 AND TABLE 4 ABOUT HERE

4.2.4 Combined analyses of Experiments 3a and 3b

To address concerns about the insufficient power of Experiments 3a and 3b to detect interactions between previous (n-1) and current trials (n), we conducted a post-hoc analysis that combined data from both experiments and included cross-task sequences from filler trials. This was important due to data loss exacerbated by the transition from pre- to post-COVID analyses. Experiment 3a had four fewer subjects to analyze than Experiment 1; Experiment 3b had nine fewer subjects to analyze than Experiment 1. By pooling the data, we aimed to perform a post-hoc analysis of trial (n) performance based on trial (n-1), increasing the likelihood of detecting such effects, if such effects are present.[2]

[2] We note two caveats with this approach. First, while critical trials for the sentence task were all positive adjectives, filler trials were all negative adjectives. Since irony is delayed for negative adjectives as well in Experiment 1, including these trials in our analysis provides another avenue to assess conflict processes in interpretation. Second, pooling data across experiments yields unbalanced sequences across conditions. For Stroop-to-Sentence sequences, there were 1449 literal sentences preceded by congruent Stroop, 1647 literal sentences preceded by incongruent Stroop, 1452 ironic sentences preceded by congruent Stroop,

We repeated our analyses to determine the extent to which a larger dataset would be sensitive to interpretive conflicts associated with processing irony. First, we analyzed sentence trials (n) in Stroop-to-Sentence sequences to evaluate the presence of early-arriving conflict. During the Pre-adjective region, Figure 10 and Table 5 illustrate that Target preference was unexpectedly greater for literal interpretations compared to ironic ones, and this was numerically (but not significantly) related to previous Stroop type (t = 1.90, p > .05). Target preference for literal sentences was numerically greater than ironic ones when preceded by congruent Stroop (t = 1.88, p > .05) but not incongruent Stroop (t = 1.03, p > .30). During the Adjective-Noun region, Target preference was significantly greater for literal interpretations compared to ironic ones, leading to a main effect of current sentence trial (t = 3.70, p < .001). There was no additional effect or interaction with previous Stroop trials (p’s > .20). During the Pronoun region, Target preference was greater for literal interpretations compared to ironic ones, and this was numerically related to previous Stroop type (t = 1.88, p > .05). Planned comparisons revealed that this interaction was opposite from what would be predicted by cognitive-control engagement. Target preference for literal sentences was numerically greater than ironic ones when preceded by *congruent* Stroop (t = 1.83, p > .05) but unaffected by incongruent Stroop (t = 0.29, p > .70). Since the pooled analysis replicates delays in interpreting ironic compared to literal sentences with no interaction with prior Stroop, this suggests that literal and ironic interpretations are not yet in conflict in early processing.

INSERT FIGURE 10 AND TABLE 5 ABOUT HERE

Next, we analyzed Stroop trials (n) in Sentence-to-Stroop sequences to evaluate the presence of *late*-arriving conflict. Figure 11 and Table 6 illustrate that RTs were faster on congruent compared to incongruent trials, leading to a main effect of Stroop type (t = 5.38, p < .001). Importantly, this difference

and 1050 ironic sentences preceded by incongruent Stroop. For Sentence-to-Stroop sequences, there were 565 congruent Stroop trials preceded by literal sentences, 657 congruent Stroop trials preceded by ironic sentences, 655 incongruent Stroop trials preceded by literal sentences, 444 incongruent Stroop trials preceded by ironic sentences across participants.

was greater when Stroop trials were preceded by literal compared to ironic sentences, leading to an interaction between prior sentence and current Stroop type (t = 3.50, p < .001). Planned comparisons revealed typical Stroop effects when prior sentences were literal (t = 6.54, p < .001), but no Stroop effects when the prior sentences were ironic (t = 1.24, p > .20). This provides evidence of cross-task adaptation of cognitive control from ironic sentences to incongruent Stroop trials, and suggests that interpreting irony involves resolving conflict, which engages cognitive-control operations that facilitates performance on Stroop incongruent trials—a canonical measure of control.

INSERT FIGURE 11 AND TABLE 6 ABOUT HERE

4.3 Discussion

Experiment 3 tested whether observed delays in interpreting irony arise due to conflict between the literal and ironic meanings. We found that a listener's cognitive-control state (manipulated via prior Stroop) did not modulate how quickly they interpreted ironic utterances, suggesting that the literal and ironic meanings of a sentence may not be concurrently active from the earliest moments of interpretation. However, as participants switched from processing literal to ironic meaning, we found evidence that cognitive-control engaged, and impacted subsequent Stroop task performance. Although no effects were found in Experiment 3b alone, the combined Experiment 3ab analysis increased power, revealing effects consistent with late conflicts between literal and ironic interpretations. These patterns are inconsistent with prominent characterizations of irony as conventionalized in the lexicon and directly retrievable during comprehension (Gibbs, 1986; Giora & Fein, 1999). Instead, since conflicts between an utterance's literal meaning and the speaker's intended meaning only emerge after literal interpretations are available, these findings provide converging evidence that ironic interpretations are generated in real time. Doing so allows listeners to appreciate the pragmatic force of irony by relishing mismatches between an utterance's literal meaning and a speaker's intent.

In Experiment 4, we examine the speaker-related cues that support real-time inferences of irony. It is well documented that listeners readily harness their knowledge about speakers to inform sentence interpretation (Arnold et al., 2007; Fairchild & Papafragou, 2018; Fairchild et al., 2020; Grodner &

Sedivy, 2011;Van Berkum et al., 2008). When told facts about a speaker (e.g., s/he has object agnosia, is autistic, is a child, is a L2 speaker, is a nerd), they immediately adjust expectations of likely utterance meanings with no additional training. Listeners also categorize speakers based on language use. On average, ironic speakers are judged to be more aggressive, offensive, and angry compared to those who use literal utterances (Climie & Pexman, 2008; Leggitt & Gibbs, 2000; Toplak & Katz, 2000). These effects are further mediated by knowledge about speakers (Dews & Winner, 1995; Dews et al., 1995; Pexman & Olineck, 2002; Colston & Lee, 2004). However, the research to date has focused on comparing literal vs. ironic utterances, which makes it unclear whether judgements arise from the pragmatic function of irony or more general aversion for negative evaluations or indirect statements. To understand how inferences about speakers' intentions influence utterance interpretation, Experiment 4 will examine how participants rate ironic speakers relative to literal and opposite speakers.

5. Experiment 4: How do listeners evaluate ironic speakers?

5.1 Method

5.1.1 Participants

Two hundred workers on Amazon Mechanical Turk participated for pay. Data from eight participants were excluded for submitting multiple entries. Another five participants were excluded for listing a non-English as their primary language mode of communication. Data analysis was conducted over the remaining 187 participants.

5.1.2 Procedures and Materials

The experiment unfolded over two parts. During the familiarization phase, participants watched eight videos, each showing events involving two different-gender characters (Fred and Sally). After each video, participants heard one of two speakers describe a character. Half the participants (n = 93) were told that one speaker would always be ironic (Ironic Ike or Iris) and the other would always be literal (Literal Lucy or Luke). The other half (n = 94) were told that one speaker would always say the opposite of what s/he means (Opposite Ollie or Olive) and the other would always say what s/he means.

After viewing all videos and descriptions, participants moved onto the test phase. Here, they used a 7-point Likert scale to rate speakers along six dimensions: matter-of-fact, polite, critical, aggressive, weird, and confusing. For each dimension, ratings for both speakers were yoked in presentation. For example, on the dimension of politeness, participants were first asked: “On a scale of 1 to 7, where 1 is NOT very polite and 7 is very polite, how would you describe the female speaker?” They were then asked for the same judgment on the male speaker. The displays and sentences were based on Experiments 1 and 2. We created 16 presentation lists, which counterbalanced the speaker gender in the familiarization phase and randomized question order within and across the rating dimensions in the test phase. Within lists, familiarization phases balanced the number of literal vs. non-literal statements, positive and negative adjectives, and location of the described character.

5.2 Results and Discussion

Figure 12 illustrates ratings across the six dimensions when the familiarization phase featured literal vs. ironic speakers (left) or literal vs. opposite speakers (right). Speaker ratings were analyzed through linear regressions, with familiarization phase (literal/ironic vs. literal/opposite) and speaker type (literal vs. ironic/opposite) as fixed effects. First, we analyzed each dimension separately and assessed which properties were driving global differences. Table 7 illustrates that literal speakers were considered more matter-of-fact and polite compared to non-literal counterparts, leading to main effects of speaker type but no additional interaction with familiarization phase (p’s > .30). In contrast, non-literal speakers were considered more critical compared to literal counterparts, and speakers in the literal/ironic context were considered more critical compared to those in the literal/opposite context. This led to main effects of speaker type and familiarization phase, but no interaction between the two (p’s > .90). Non-literal speakers were rated as more aggressive compared to literal counterparts, leading to main effects of speaker type but no additional interaction with familiarization phase (p’s > .70).

INSERT FIGURE 12 AND TABLE 7 ABOUT HERE

The last dimensions pinpointed how ironic and opposite speakers differ from literal speakers and from each other. Non-literal speakers were weirder compared to literal counterparts, but this difference was greater for opposite (t = 11.78, p < .001) compared to ironic speakers (t = 5.51, p < .001). This led to a main effect of speaker type and an interaction with the familiarization phase. Non-literal speakers were also considered to be more confusing compared to literal counterparts, but this difference was greater for opposite (t = 13.14, p < .001) compared to ironic speakers (t = 8.01, p < .001). This led to a main effect of speaker type and an interaction with the familiarization phase. Together, this demonstrates that when indirectness has a pragmatic function, listeners perceive this to reflect the communicative context —in particular, the speaker's desire to critique or express aggression. When indirectness does *not* convey a pragmatic function, listeners attribute deviations from literal utterances to speaker traits, judging those who use opposite utterances to be weird and confusing. This suggests that listeners perceive irony to be a socially conventionalized way to criticize others, and that listeners' real-time inferences of irony are based on their knowledge of the pragmatic function of these utterances.

6. General Discussion

In this study, we asked how listeners access irony during real-time comprehension. By applying psycholinguistic theories and methods, we aimed to identify the general-purpose mechanisms that allow people to construct interpretations that are both stable enough to be shared and flexible enough to fit the communicative context. Unlike the robust frequency effects found with literal sentences, we found that the real-world frequency of ironic critiques vs. compliments did not influence the timing of reference restriction. Unlike the systematicity of interpreting irony, opposites caused widespread delays across all sentences, despite having identical truth conditions as irony. Conflicts between literal and ironic meanings were late arriving during comprehension, suggesting that initial semantic analysis may provide the basis for recruiting context to generate irony. Finally, listeners hold stable beliefs about speakers based on their irony usage, and their understanding of irony's pragmatic functions may support predictions of future language use. Together, these findings suggest that irony involves real-time reasoning over utterance meanings and speaker intent, and signatures of these computations are evident during processing.

Our findings relate to other work examining how irony interacts with information channels (e.g., prosody, facial cues) and interpretive systems (e.g., language, emotion processing) across time scales and individuals. For example, early ERP components reveal that listeners are sensitive to speaker accent and prosody, and these perceptual effects are more prominent when interpreting the less conventional, ironic compliments compared to ironic critiques (Caffarra et al., 2019; Mauchand et al., 2020; Mauchand et al., 2021). Similarly, reading times for irony are affected by subtle changes in the presence of negation and the familiarity of lexical expressions (Filik et al., 2014; Giora et al., 2015). Other studies highlight the probabilistic nature of accessing irony, which update within experimental sessions (Olkoniemi et al., 2016; Olkoniemi et al., 2019; Spotorno & Noveck, 2014), and vary based on individual personality traits (Bruntsch & Ruch, 2017; Filik et al., 2018), social and emotional skills (Filik et al., 2017; Olkoniemi et al., 2019; Spotorno & Noveck, 2014), and cognitive flexibility (Kaakinen et al., 2014; Zajączkowska & Abbot-Smith, 2020). Ongoing research continues to produce evidence for delayed interpretation of irony (Deliens et al., 2018; Filik et al., 2014; Turcan & Filik, 2016) as well as for rapid sensitivity (Caffarra et al., 2019; Kaakinen et al., 2014; Mauchand et al., 2020; Mauchand et al., 2021), and sometimes both in the same study (Filik et al., 2018; Giora et al., 2015; Spotorno et al., 2013). These findings highlight the challenges of interpreting empirical differences in timing, and call for more specific characterizations of the algorithms that underlie successful comprehension of irony.

By situating irony within a generalized architecture for sentence comprehension, our study draws connections to other phenomena within the semantics-pragmatics interface. On the one hand, listeners rapidly adapt to *speaker* characteristics, and adjust their utterance interpretations based on this top-down knowledge (Arnold et al., 2007; Fairchild & Papafragou, 2018; Fairchild et al., 2020; Grodner & Sedivy, 2011; Van Berkum et al., 2008). Our findings suggest that such generalizations can operate over speakers' language use, but require familiarity with relevant categories. Hence, despite being told how to compute truth conditions, processing profiles for interpreting Opposite Ollie were highly unsystematic. In contrast, listener expectations about *utterance* meanings are strikingly slow to update. Even after hearing irony up to 20 times within a restricted context, listeners' reliance on initial semantic analysis remained robust throughout the experiment. This pattern is consistent with converging evidence that listeners implement inferences about speakers' intention in real time (Huang & Snedeker, 2009, 2011, 2018; Gardner et al., 2021; Pogue et al., 2016; Ryskin et al., 2019), even when pragmatic meanings are *more* frequent than semantic counterparts, e.g., scalar implicatures. While listeners can eventually converge on algorithms to directly predict pragmatic meanings, this updating is strikingly slow and only emerges after 20-100 trials of concentrated exposure (Breheny et al., 2013; Degen & Tannenhaus, 2015, 2016; Grodner et al., 2010; Huang & Snedeker, 2018; Gardner et al., 2021; Pogue et al., 2016; Ryskin et al., 2019). Together, this offers additional evidence that the lexicon does not store frequency statistics about pragmatics.

These dual algorithms – fast reasoning about speakers, slow predictions about utterances – are remarkable since they suggest that listeners can attribute usage statistics to distinct generative processes that contribute meanings during comprehension (Huang & Snedeker, 2018; Huang & Ovans, 2021). Notably, these algorithms trace the division of labor proposed by contemporary theories of language, between linguistically encoded meaning of utterances (semantics) and how this meaning is enriched by current context, world knowledge, and speaker goals (pragmatics). This dual architecture contrasts with usage-only accounts that eschew multiple meaning systems in favor of the parsimony of a single level of analysis. This includes the Direct-access accounts of irony, whereby all meanings and their frequencies are encoded in the lexicon (Gibbs, 1986; Giora & Fein, 1999). More recently, uniform architectures have received a great deal of attention as they are the basis for modern large language models (LLMs). These models achieve impressive feats of utterance interpretation by way of aggregating large-scale regularities across wide-ranging written corpora and implementing a simple algorithm for relating current sentence features to next-word predictions. Yet, despite their mastery of formal aspects of linguistic competence (e.g., subject-verb agreement) and conventional usages (e.g., metaphors, indirect speech), current LLMs perform poorly when reasoning about utterances that require pragmatic inferences, including those that involve humor and irony, or require adapting to speaker characteristics, world knowledge, and social cognition (Hu et al., 2022; Mahowald et al., 2024; Domanski et al., 2024; Denning et al., 2025).

In this light, the ability for human listeners to reverse engineer sources of meanings is striking since a given utterance is a confluence of many factors that can contribute meaning (e.g., speaker, context, words). One way listeners may do so is by pursuing initial semantic analysis as a strong prior for likely meaning, and over time create ad-hoc algorithms for computing interpretations based on accumulating evidence from speaker-specific usage (Huang & Snedeker, 2018; Pogue et al., 2016; Ryskin et al., 2019). On the face of things, this iterative process may appear inefficient, but its chief benefit is its ability to support communication when common ground is low or unknown (e.g., talking to new people, about new topics). Moreover, measurable gaps between the meanings generated by speaker-general vs. speaker- specific algorithms may provide an ongoing probabilistic representation of alignment between listeners and speakers. This notion is consistent with computational and experimental work demonstrating that listeners generate top-down, context-specific predictions *alongside* bottom-up analysis of the signal, and incrementally update weightings of internal representations based on feedback from prediction errors (Chang et al., 2006; Dell & Chang, 2014; Huang & Snedeker, 2018; Eddine et al., 2024). More broadly, algorithms for probabilistic common ground may be useful in interactive communication, where listeners are also speakers in subsequent turns and must plan utterances based on dynamic inferences about shared knowledge (Ferrerira, 2019; Pickering & Garrod, 2013; Yoon & Brown-Schmidt, 2023).

In conclusion, the current study demonstrates that irony involves real-time reasoning over utterance meanings and speaker intentions, and that signatures of both computations are evident during comprehension. Listeners exhibit frequency effects for literal but not ironic meanings, are systematic when interpreting irony but not opposites, experience late-arriving conflicts between literal and ironic meanings because semantic analysis is pursued before the inference is drawn, and reason about irony to generate specific inferences about speaker properties. More broadly, our findings are consistent with an architecture whereby semantics encode past regularities, support stable meanings across contexts, and provides a general-purpose lexicon that is useful when common ground is unknown. Pragmatics, in contrast, encode the idiosyncrasies of talking to particular speakers and support flexible communication within a specific context. Irony offers an informative test case for understanding how this division of labor gives rise to stable and flexible language use, and paves the way for future research that describes the complex algorithms supporting meanings during communication.

Acknowledgments

We are grateful to Karly Schwarz, Hannah Sichel, and research assistants in the Language and Cognition Lab and the Cognition and Language Processing Lab for supporting data collection. We thank Chris Heffner, Amritha Mallikarjun, and Alix Kowalski for their help with stimuli creation. Materials, data, and code are available at https://osf.io/4298j/. This work has benefited from conversations with Bob Slevc, Colin Phillips, Ellen Lau, and audiences at the Trends in Experimental Pragmatics Workshop and the 31st annual CUNY conference on Human Sentence Processing. This work is supported by an MCM award to RA, a William Orr Dingwall Dissertation Fellowship to LKS, an NSF NRT Grant #1449815 to RA and LKS, and NSF-BCS grant to YH (BCS-1844194). Author address: Department of Hearing and Speech Sciences, University of Maryland, College Park, MD 20742. Email address: ythuang1@umd.edu.

References

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition, 76*(1), B13-B26.

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say *thee uh* you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(5), 914-930.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor, 16*, 243-260.

Averbeck, J. M., & Hample, D. (2008). Ironic message production: How and why we produce ironic messages. *Communication Monographs, 75*(4), 396-410.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version, 1.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Investigating the time-course of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes, 28(4),* 443-467.

Bruntsch, R., & Ruch, W. (2017). Studying irony detection beyond ironic criticism: Let's include ironic praise. *Frontiers in Psychology, 8*.

Bryant, G. A., & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor and Symbol, 17*(2), 99-119.

Caffarra, S., Motamed Haeri, A., Michell, E., & Martin, C. D. (2019). When is irony influenced by communicative constraints? ERP evidence supporting interactive models. *The European Journal of Neuroscience, 50*(10), 3566-3577.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113,* 234–272.

Clark, H. H., & Gerrig, R. J. (1984). On the pretense theory of irony. *Journal of Experimental Psychology: General, 113*(1), 121-126.

Climie, E. A., & Pexman, P. M. (2008). Eye gaze provides a window on children's understanding of verbal irony. *Journal of Cognition and Development, 9*(3), 257-285.

Colston, H. L. (1997). Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes, 23*(1), 25-45.

Colston, H. L., & Lee, S. Y. (2004). Gender differences in verbal irony use. *Metaphor and Symbol, 19(4)*, 289-306.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science, 39,* 667–710.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science, 40,* 172–201.

Deliens, G., Antoniou, K., Clin, E., Ostashchenko, E., & Kissine, M. (2018). Context, facial expression and prosody in irony processing. *Journal of Memory and Language, 99*, 35-48.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological*

Sciences, 369, 20120394.

Denning, J. M., Guo, X. H., Snefjella, B., & Blank, I. A. (2025). Do Large Language Models know who did what to whom?. arXiv preprint arXiv:2504.16884.

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? The social functions of irony. *Discourse Processes, 19*(3), 347-367.

Dews, S., & Winner, E. (1999). Obligatory processing of literal and nonliteral meanings in verbal irony. *Journal of Pragmatics, 31*(12), 1579-1599.

Domanski, S., Rudinger, R., Carpuat, M., Shafto, P., & Huang, Y. (2024). Assessing common ground via language-based cultural consensus in humans and large language models. *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* Rotterdam, Netherlands.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of memory and language, 27*(4), 429-446.

Eddine, S. N., Brothers, T., Wang, L., Spratling, M., & Kuperberg, G. R. (2024). A predictive coding model of the N400. *Cognition, 246,* 105755.

Fairchild, S., & Papafragou, A. (2018). Sins of omission are more likely to be forgiven in non-native speakers. *Cognition, 181*, 80-92.

Fairchild, S., Mathis, A., & Papafragou, A. (2020). Pragmatics and social meaning: Understanding under-informativeness in native and non-native speakers. *Cognition, 200*, Article 104171.

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology, 70(1),* 29-51.

Filik, R., Leuthold, H. Wallington, K., & Page, J. (2014). Testing theories of irony processing using eye-tracking and ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 811-828.

Filik, R., Howman, H., Ralph-Nearman, C., & Giora, R. (2018). The role of defaultness and personality factors in sarcasm interpretation: Evidence from eye-tracking during reading. *Metaphor and Symbol, 33*(3), 148-162.

Gardner, B., Dix, S., Lawrence, R., Morgan, C., Sullivan, A., & Kurumada, C. (2021). Online pragmatic interpretations of scalar adjectives are affected by perceived speaker reliability. *PLoS One, 16*(2), Article e0245130.

Gibbs, R. W., Jr. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General, 115*(1), 3-15.

Gibbs, R. W., Jr. (2000). Irony in talk among friends. *Metaphor and Symbol, 15*(1&2), 5-27.

Giora, R., Drucker, A., Fein, O., & Mendelson, I. (2015). Default sarcastic interpretations: On the priority of nonsalient interpretations. *Discourse Processes, 52*(3), 173-200.

Giora, R., & Fein, O. (1999). Irony: Context and salience. *Metaphor and Symbol, 14*(4), 241-257.

Giora, R., Fein, O., & Schwartz, T. (1998). Irony: Grade salience and indirect negation. *Metaphor and Symbol, 13*(2), 83-101.

Giora, R., Drucker, A., Fein, O., & Mendelson, I. (2015). Default sarcastic interpretations: On the priority of non-salient interpretations. *Discourse Processes, 52(3),* 173–200.

Giora, R., Jaffe, I., Becker, I., & Fein, O. (2018). Strongly attenuating highly positive concepts: The case of default sarcastic interpretations. *Review of Cognitive Linguistics, 16(1),* 19-47.

Grodner, D., Klein, N., Carbary, K., & Tanenhaus, M. (2010). Some and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116,* 42–55.

Grodner, D., & Sedivy, J. C. (2011). 10 The Effect of Speaker-Specific Information on Pragmatic Inferences. In *The processing and acquisition of reference* (Vol. 2327, pp. 239-272). MIT Press.

Hsu, N. S., & Novick, J. M. (2016). Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. *Psychological Science*, 27(4), 572–582.

Hsu, N. S., Kuchinsky, S. E., & Novick, J. M. (2021). Direct impact of cognitive control on sentence processing and comprehension. *Language, Cognition and Neuroscience*, 36(2), 211–239.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801.*

Huang, Y. & Ovans, Z. (2021). Who "it" is influences what "it" does: Discourse effects on children's syntactic parsing. *Cognitive Science, 46*, e13076.

Huang, Y. & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology, 58*, 376-415.

Huang, Y. & Snedeker, J. (2011). 'Logic & Conversation' revisited: Evidence for a division between semantic and pragmatic content in real time language comprehension. *Language and Cognitive Processes, 26*, 1161-1172.

Huang, Y., & Snedeker, J. (2013). The use of lexical and referential cues in children's online interpretation of adjectives. *Developmental Psychology, 49*(6), 1090–1102.

Huang, Y. & Snedeker, J. (2018). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology, 102*, 105-126.

Ivanko, S. L., & Pexman, P. M. (2003). Context incongruity and irony processing. *Discourse Processes, 35*(3), 241-279.

Jorgensen, J. (1996). The functions of sarcastic irony in speech. *Journal of Pragmatics, 26*(5), 613-634.

Kaakinen, J. K., Olkoniemi, H., Kinnari, T., & Hyönä, J. (2014). Processing of written irony: An eye movement study. *Discourse Processes, 51*(4), 287-311.

Kan, I. P., Teubner-Rhodes, S., Drummey, A. B., Nutile, L., Krupa, L., & Novick, J. M. (2013). To adapt or not to adapt: The question of domain-general cognitive control. *Cognition*, 129(3), 637-651.

Katz, A. N., Blasko, D. G., & Kazmerski, V. A. (2004). Saying what you don't mean: Social influences on sarcastic language processing. *Current Directions in Psychological Science, 13*(5), 186-189.

Kim, A. E., Langlois, V. J., Ness, T., Wade, M., & Novick, J. M. (2025). Resolving conflicting interpretations: Theta band oscillations and the role of cognitive control. Neuropsychologia, 109214.

Kowatch, K., Whalen, J. M., & Pexman, P. M. (2013). Irony comprehension in action: A new test of processing for verbal irony. *Discourse Processes, 50*(5), 301-315.

Kreuz, R., & Roberts, R. M. (1995). Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbolic Activity, 10*(1), 21-31.

Leggitt, J. S., & Gibbs, R. (2000). Emotional reactions to verbal irony. *Discourse Processes*, *29*(1), 1-24.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49(4)*, 1494–1502.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perceptions & Psychophysics, 53*(4), 372-380.

Mauchand, M., Caballero, J. A., Jiang, X., & Pell, M. D. (2021). Immediate online use of prosody reveals the ironic intentions of a speaker: Neurophysiological evidence. *Cognitive, Affective, & Behavioral Neuroscience, 21*, 74-92.

Mauchand, M., Vergis, N., & Pell, M. D. (2020). Irony, prosody, and social impressions of affective stance. *Discourse Processes, 57*(2), 141-157.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176-182.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in cognitive sciences, 28(6),* 517-540.

Ness, T., Langlois, V. J., Kim, A. E., & Novick, J. M. (2025). The state of cognitive control in language processing. *Perspectives on Psychological Science*, 20(2), 219–240.

Ness, T., Langlois, V. J., Novick, J. M., & Kim, A. E. (2024). Theta-band neural oscillations reflect cognitive control during language processing. Journal of Experimental Psychology: General, 153(9), 2279-2298.

Novick, J. M., Thompson-Schill, S. L., & Trueswell, J. C. (2008). Putting lexical constraints in context into the visual-world paradigm. *Cognition*, 107(3), 850–903.

Olkoniemi, H., Ranta, H., & Kaakinen, J. K. (2016). Individual differences in the processing of written sarcasm and metaphor: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(3), 433.

Olkoniemi, H., Strömberg, V., & Kaakinen, J. K. (2019). The ability to recognize emotions predicts the time-course of sarcasm processing: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology, 72*(5), 1212-1223.

Ovans, Z., Hsu, N. S., Bell-Souder, D., Gilley, P., Novick, J. M., & Kim, A. E. (2022). Cognitive control states influence real-time sentence processing as reflected in the P600 ERP. *Language, Cognition and Neuroscience*, 37(8), 939–947.

Pexman, P. M., Ferretti, T. R., & Katz, A. N. (2000). Discourse factors that influence online reading of metaphor and irony. *Discourse Processes, 29*(3), 201-222.

Pexman, P. M., & Olineck, K. M. (2002). Understanding irony: How do stereotypes cue speaker intent? *Journal of Language and Social Psychology, 21*(3), 245-274.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36(4),* 329-347.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology, 6*, 2035.

R core team (2020). A language and environment for statistical computing [Internet]. *R Foundation for Statistical Computing.*

Rayner, K., Pacht, J. M., & Duffy, S. A. (1994). Effects of prior encounter and global discourse bias on the processing of lexically ambiguous words: Evidence from eye fixations. *Journal of Memory and Language*, *33*(4), 527-544.

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research*, 29, 483-495.

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information integration in modulation of pragmatic inferences during online language comprehension. *Cognitive Science, 43*(8), e12769.

Schwoebel, J., Dews, S., Winner, E., & Srinivas, K. (2000). Obligatory processing of the literal meaning of ironic utterances: Further evidence. *Metaphor and Symbol, 15*(1&2), 47-61.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4), 447–481.

Spotorno, N., Cheylus, A., Van Der Henst, J. B., & Noveck, I. A. (2013). What's behind a P600? Integration operations during irony processing. *PLoS One, 8*(6), Article e66839.

Spotorno, N., & Noveck, I. A. (2014). When is irony effortful?. *Journal of Experimental Psychology: General, 143*(4), 1649-1665.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.

Thothathiri, M., Asaro, C. T., Hsu, N. S., & Novick, J. M. (2018). Who did what? A causal role for cognitive control in thematic role assignment during sentence comprehension. *Cognition*, 178, 162–177.

Toplak, M., & Katz, A. (2000). On the uses of sarcastic irony. *Journal of Pragmatics*, *32*(10), 1467-1488.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73(2)*, 89–134.

Țurcan, A., & Filik, R. (2016). An eye-tracking investigation of written sarcasm comprehension: The roles of familiarity and context. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(12), 1867-1893.

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580-591.

Van Tiel, B., & Pankratz, E. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa: a Journal of General Linguistics*, *6*(1): 32.

Yoon, S. O., & Brown-Schmidt, S. (2023). Understanding Language use in social contexts: The role of past and present discourse contexts. In R. J. Hartsuiker (Ed.), *Language Production* (pp. 284–303). Routledge.

Zajączkowska, M., & Abbot-Smith, K. (2020). "Sure I'll help—I've just been sitting around doing nothing at school all day": Cognitive flexibility and child irony interpretation. *Journal of Experimental Child Psychology, 199*, Article 104942.

Figure 1. Two hypotheses about how irony is accessed during comprehension. Direct access: Red arrows indicate that ironic meanings are stored in the lexicon along with usage statistics. Arrow size indicates the frequency of meanings and their strength of activation. Literal first: Blue and green arrows indicate that irony is computed in real time and depends on the statistics of literal usage.

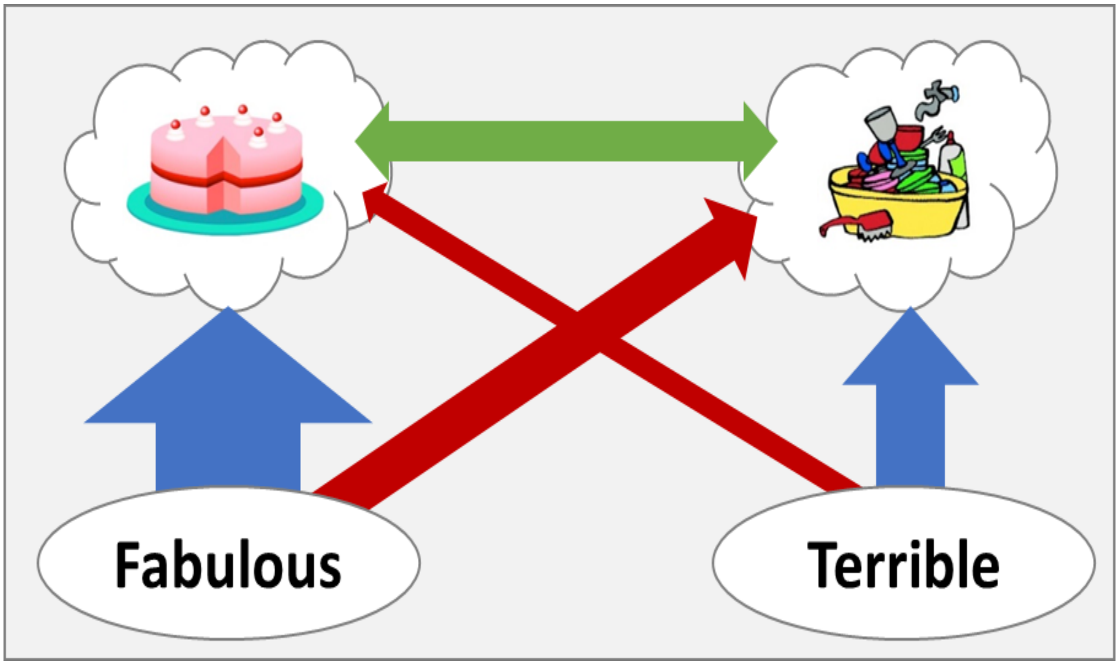


Figure 2. Across experiments, a sample display that was paired with a vignette: "Fred and Sally decided to do some baking. Sally made a beautiful cake. Fred tried his best and made a mess in the kitchen."

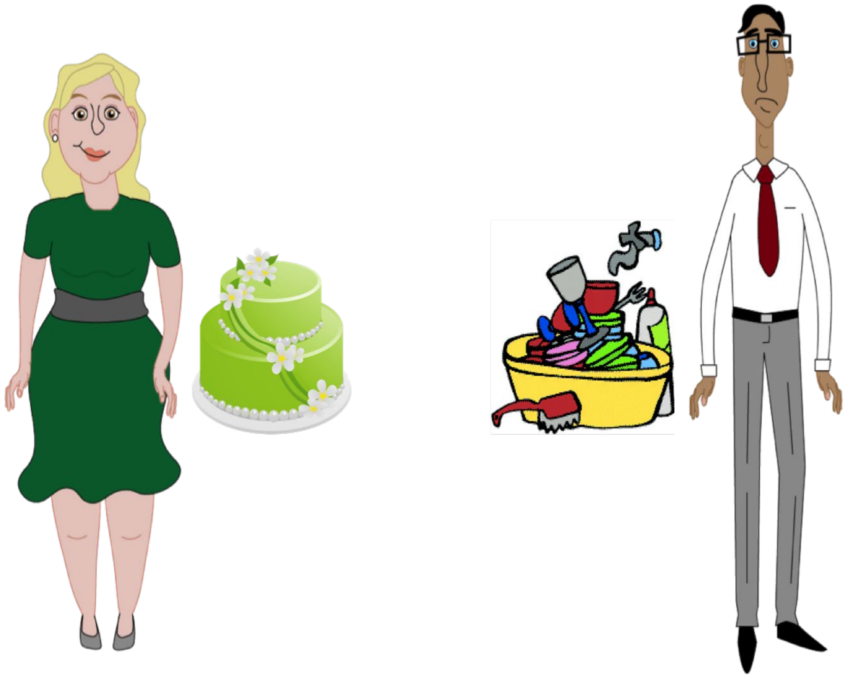


Figure 3. In Experiment 1, Target preference by region (Pre-adjective, Adjective-Noun, Pronoun), adjective valence (positive, negative), and interpretation type (literal, ironic). The first vertical line represents the adjective onset and the second vertical line represents the average onset time of the pronoun. Bars represent standard errors.

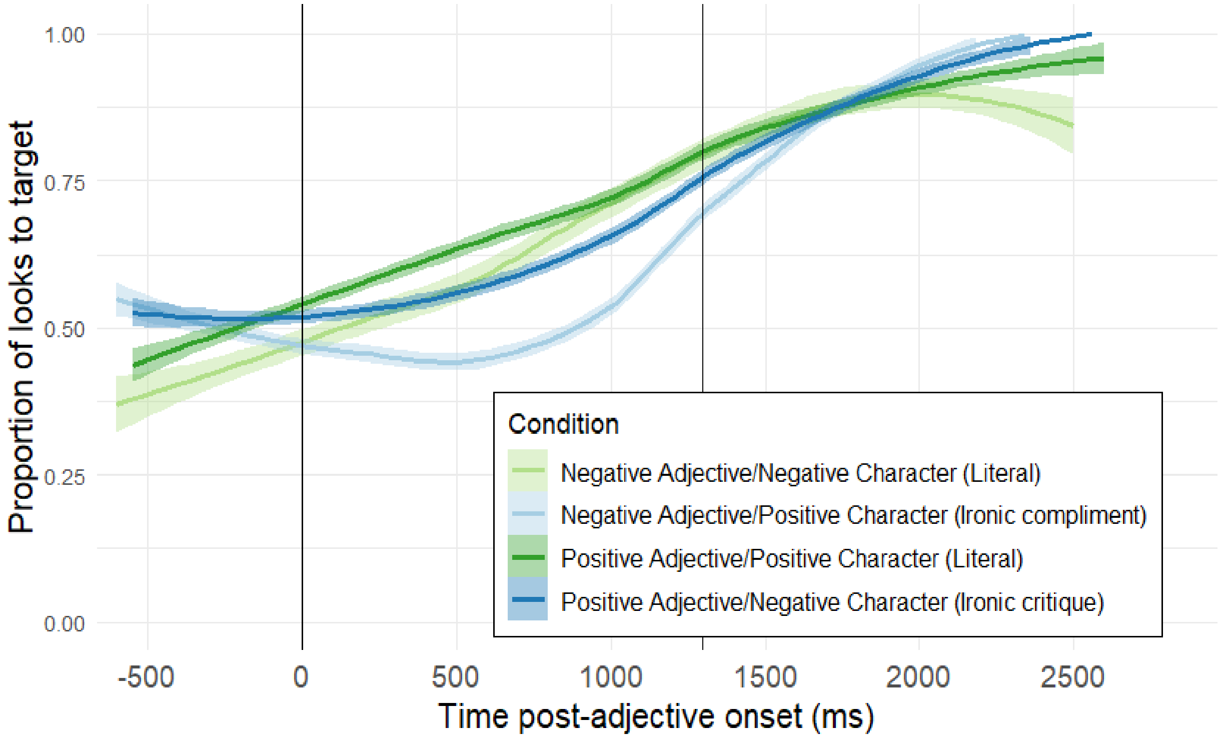


Figure 4. In Experiment 1, Target preference by region (Pre-adjective, Adjective-Noun, Pronoun) and condition (positive literal, positive ironic, negative literal, negative ironic). Bars represent standard errors.

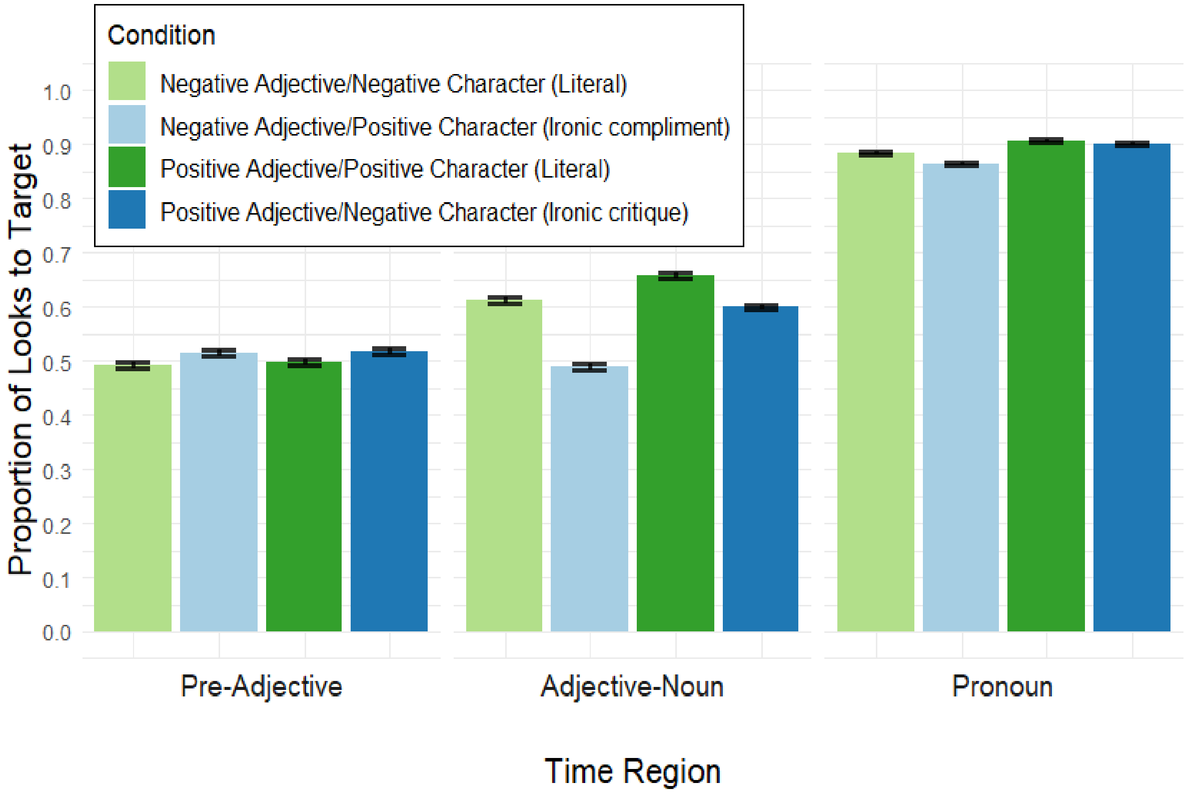


Figure 5. In Experiment 2, Target preference by region (Pre-adjective, Adjective-Noun, Pronoun), adjective valence (positive, negative), and interpretation type (literal, opposite). The first vertical line represents the adjective onset and the second vertical line represents the average onset time of the pronoun. Bars represent standard errors.

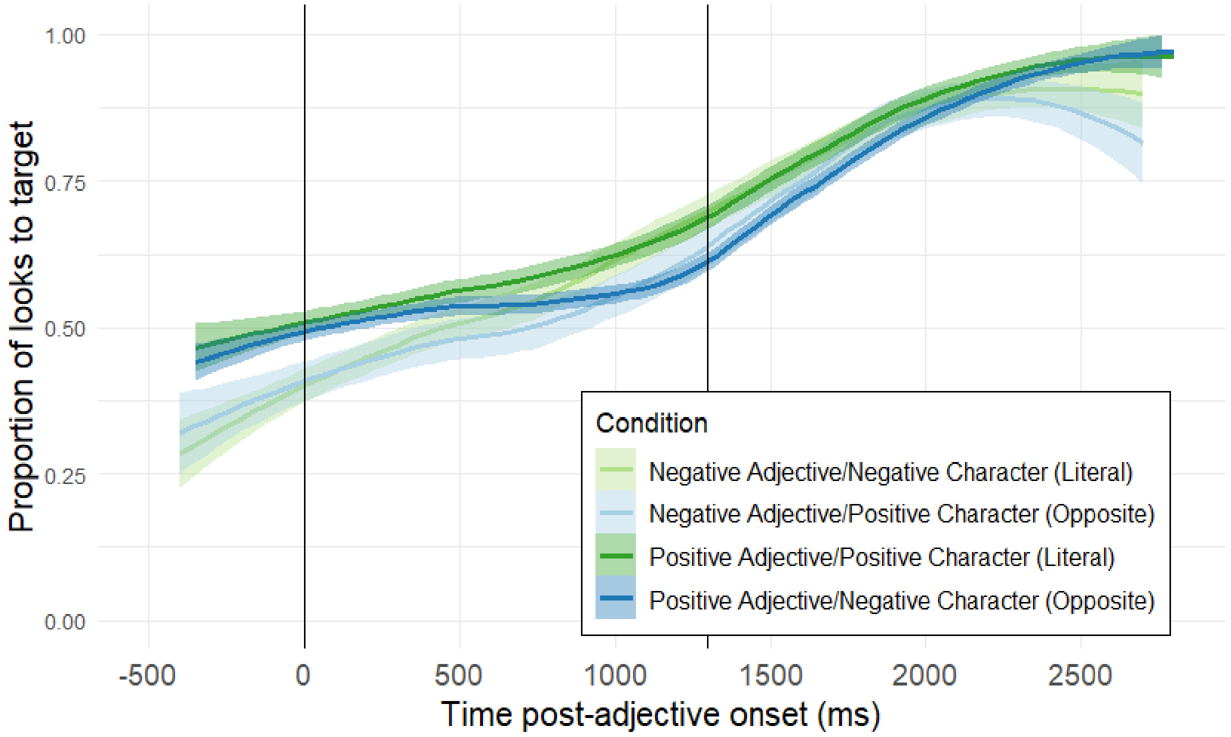


Figure 6. In Experiment 2, Target preference by region (pre-adjective, adjective-noun, pronoun) and condition (positive literal, positive opposite, negative literal, negative opposite). Bars represent standard errors.

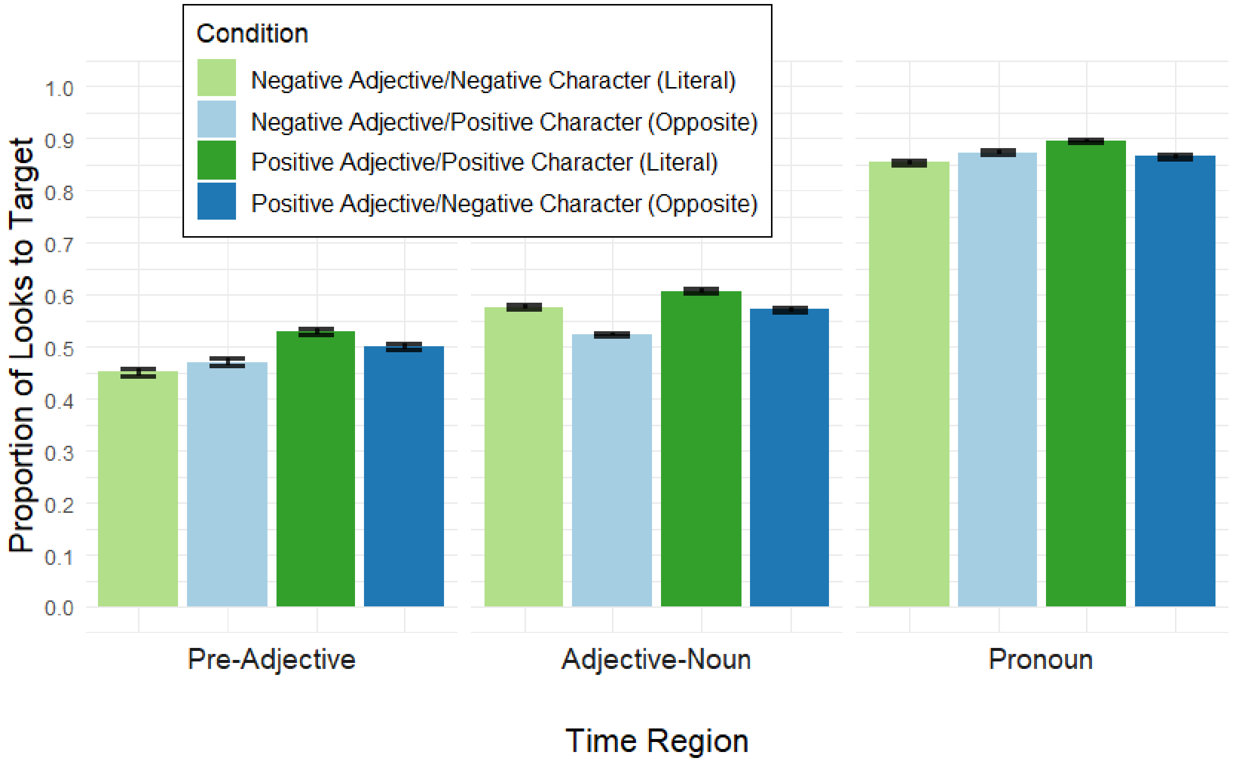


Figure 7. In Experiment 3a, Target preference by region (Pre-adjective, Adjective-Noun, Pronoun) and condition (congruent-literal, congruent-ironic, incongruent-literal, incongruent-ironic). Bars represent standard errors.

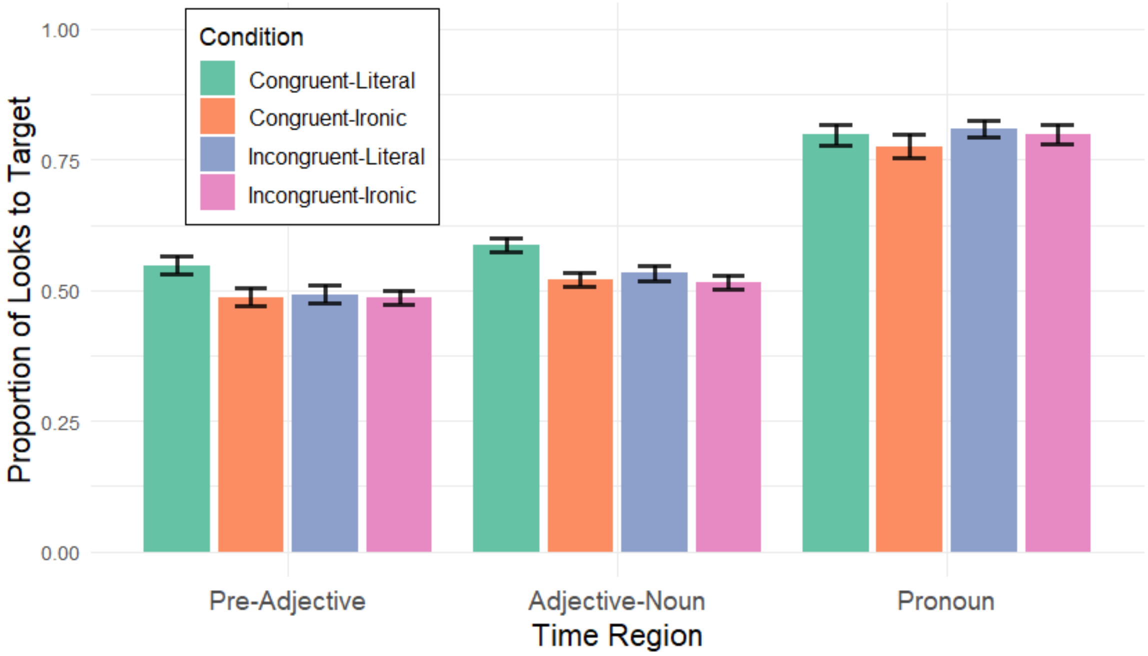


Figure 8. In Experiment 3b, Target preference by region (Pre-adjective, Adjective-Noun, Pronoun) and condition (literal, ironic). Bars represent standard errors.

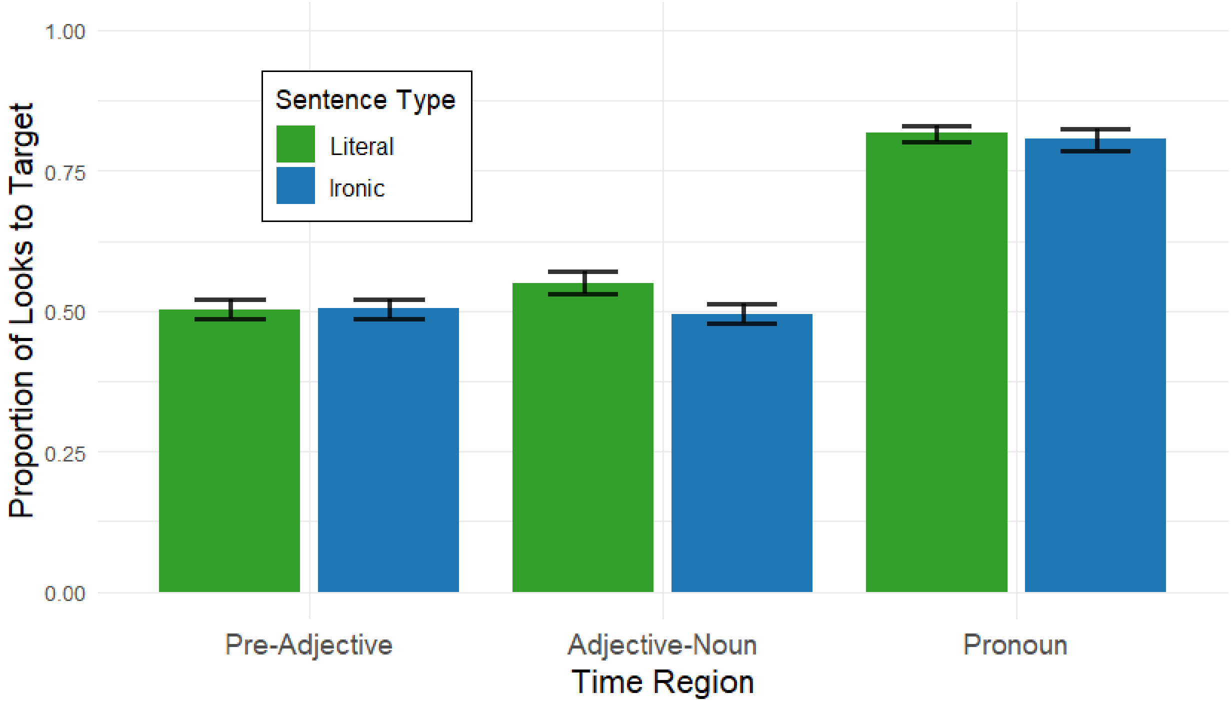


Figure 9. In Experiment 3b, reaction times by previous sentence trial (literal, ironic) and current Stroop trial (congruent, incongruent). Bars represent standard errors.

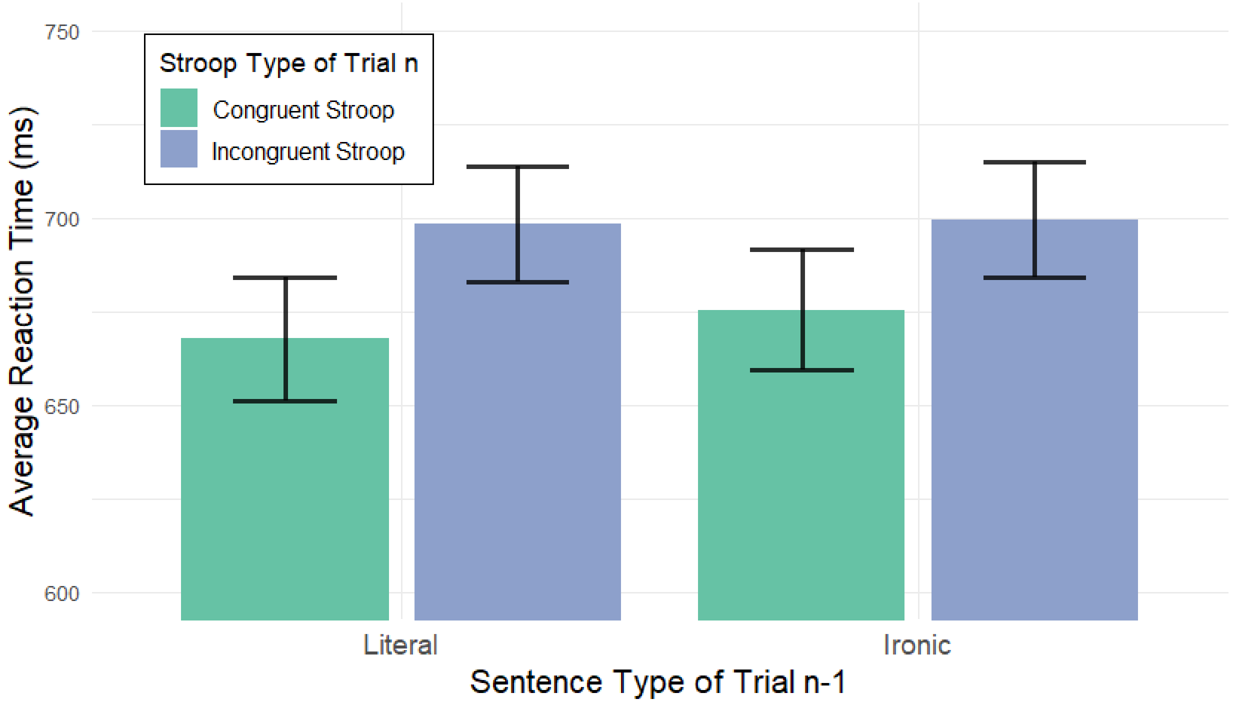


Figure 10. In Experiment 3ab, Target preference by region (Pre-adjective, Adjective-Noun, Pronoun) and condition (congruent-literal, congruent-ironic, incongruent-literal, incongruent-ironic). Bars represent standard errors.

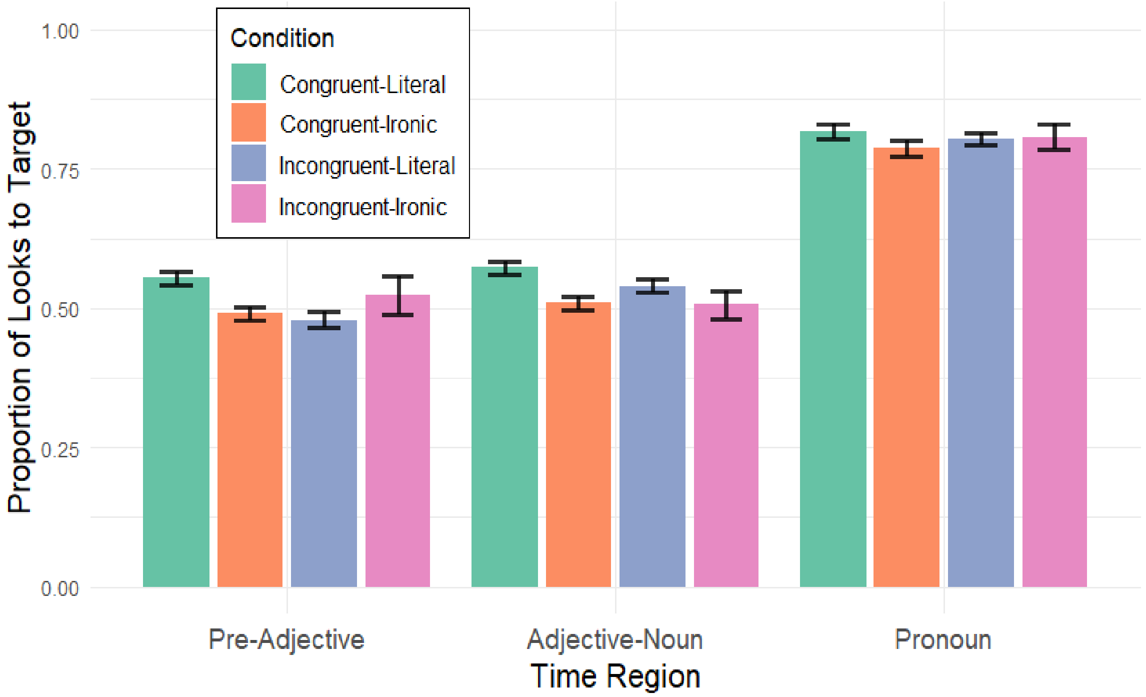


Figure 11. In Experiment 3ab, Stroop reaction times by previous sentence trial (literal, ironic) and current Stroop trial (congruent, incongruent). Bars represent standard errors.

New plot:

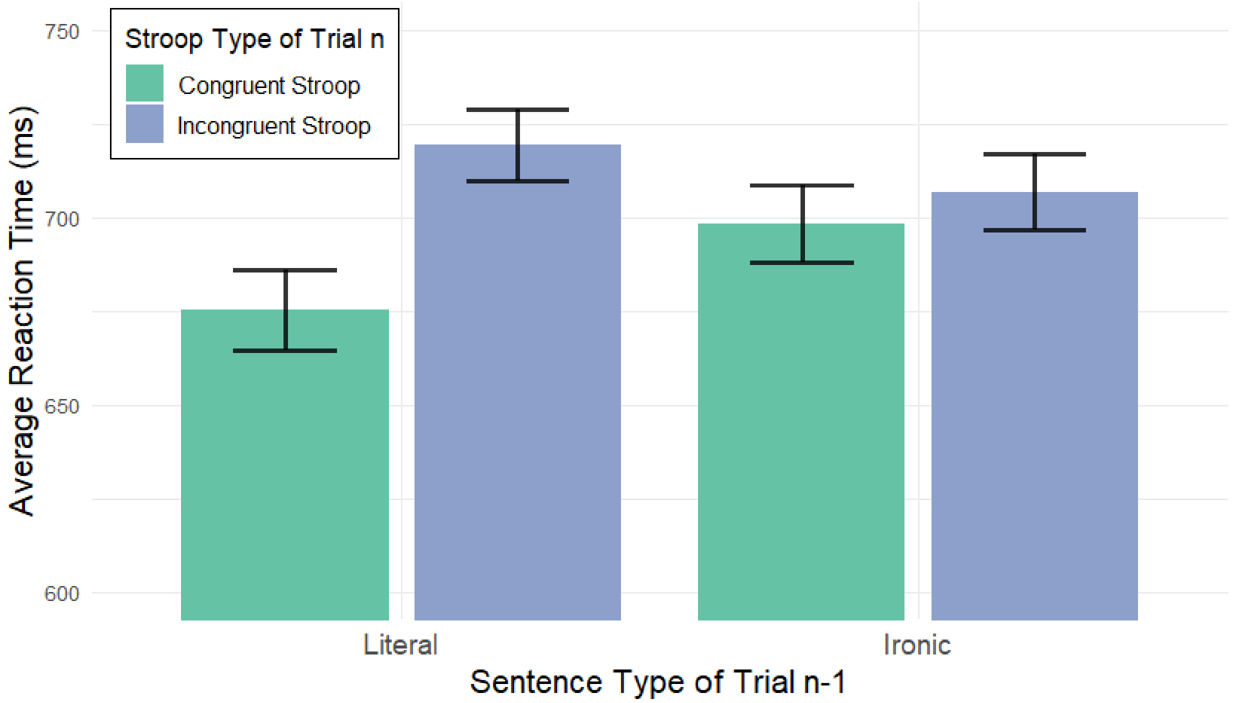


Figure 12. Mean ratings of dimensions by speaker. (A) Literal speakers are in green and ironic speakers are in blue, (B) Literal speakers are in green and opposite speakers are in blue. Bars represent standard errors.

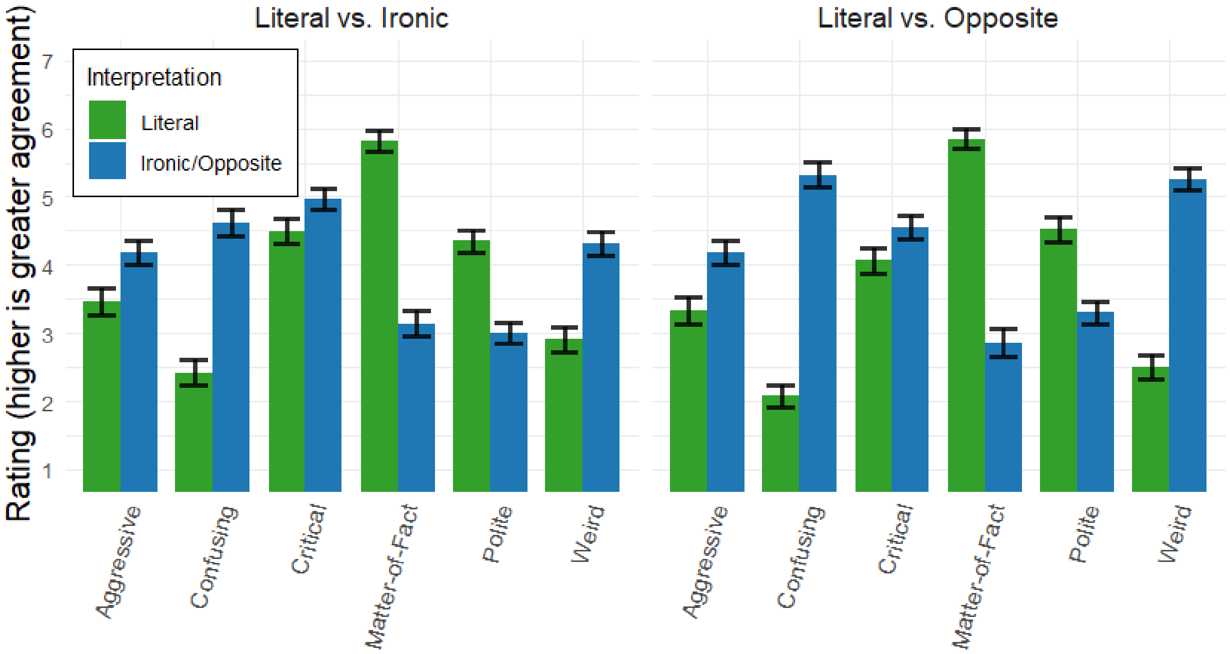
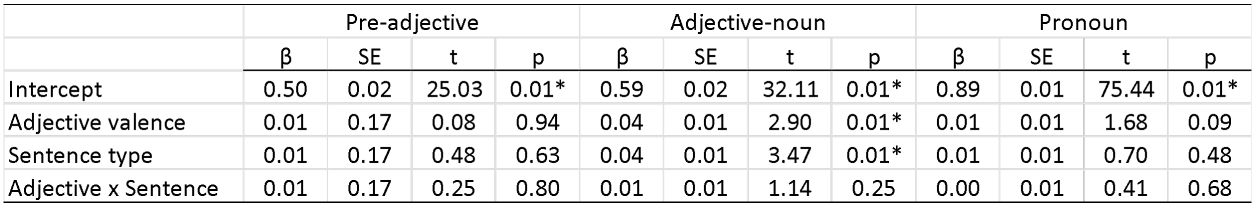
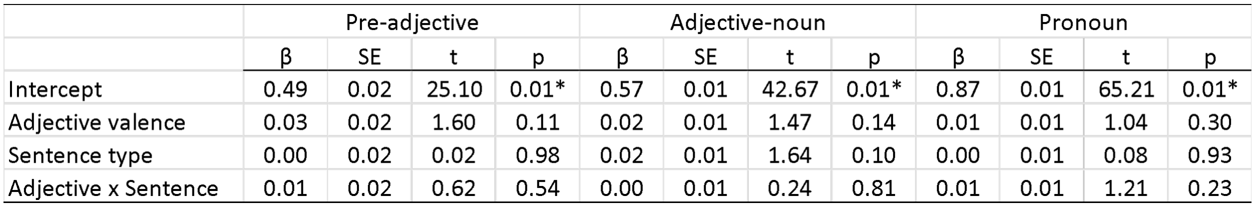


Table 1. In Experiment 1, fixed effects (Adjective valence × Sentence Type) in linear mixed-effects models of Target looks by region (Pre-adjective, Adjective-Noun, Pronoun).



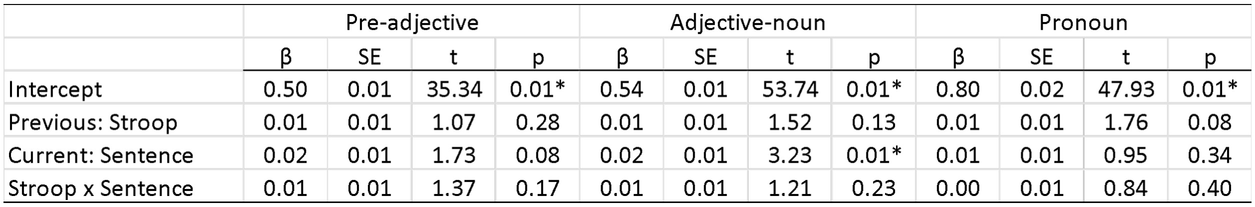
	Pre-adjective				Adjective-noun				Pronoun			
	β	SE	t	p	β	SE	t	p	β	SE	t	p
Intercept	0.50	0.02	25.03	0.01*	0.59	0.02	32.11	0.01*	0.89	0.01	75.44	0.01*
Adjective valence	0.01	0.17	0.08	0.94	0.04	0.01	2.90	0.01*	0.01	0.01	1.68	0.09
Sentence type	0.01	0.17	0.48	0.63	0.04	0.01	3.47	0.01*	0.01	0.01	0.70	0.48
Adjective x Sentence	0.01	0.17	0.25	0.80	0.01	0.01	1.14	0.25	0.00	0.01	0.41	0.68

Table 2. In Experiment 2, fixed effects (Adjective valence × Sentence Type) in linear mixed-effects models of Target looks by region (Pre-adjective, Adjective-Noun, Pronoun).



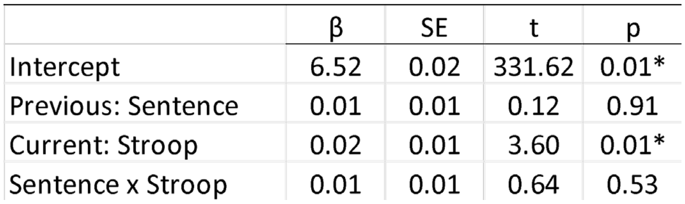
	Pre-adjective				Adjective-noun				Pronoun			
	β	SE	t	p	β	SE	t	p	β	SE	t	p
Intercept	0.49	0.02	25.10	0.01*	0.57	0.01	42.67	0.01*	0.87	0.01	65.21	0.01*
Adjective valence	0.03	0.02	1.60	0.11	0.02	0.01	1.47	0.14	0.01	0.01	1.04	0.30
Sentence type	0.00	0.02	0.02	0.98	0.02	0.01	1.64	0.10	0.00	0.01	0.08	0.93
Adjective x Sentence	0.01	0.02	0.62	0.54	0.00	0.01	0.24	0.81	0.01	0.01	1.21	0.23

Table 3. In Experiment 3a, fixed effects (Adjective valence × Sentence type) in linear mixed-effects models of Target looks by region (Pre-adjective, Adjective-Noun, Pronoun).



	Pre-adjective				Adjective-noun				Pronoun			
	β	SE	t	p	β	SE	t	p	β	SE	t	p
Intercept	0.50	0.01	35.34	0.01*	0.54	0.01	53.74	0.01*	0.80	0.02	47.93	0.01*
Previous: Stroop	0.01	0.01	1.07	0.28	0.01	0.01	1.52	0.13	0.01	0.01	1.76	0.08
Current: Sentence	0.02	0.01	1.73	0.08	0.02	0.01	3.23	0.01*	0.01	0.01	0.95	0.34
Stroop x Sentence	0.01	0.01	1.37	0.17	0.01	0.01	1.21	0.23	0.00	0.01	0.84	0.40

Table 4. In Experiment 3b, fixed effects (Previous: Sentence Type x Current: Stroop type) in regression model of log reaction times.

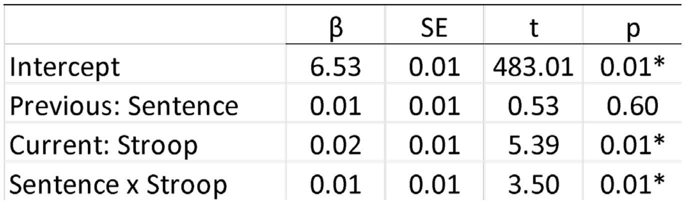


	β	SE	t	p
Intercept	6.52	0.02	331.62	0.01*
Previous: Sentence	0.01	0.01	0.12	0.91
Current: Stroop	0.02	0.01	3.60	0.01*
Sentence x Stroop	0.01	0.01	0.64	0.53

Table 5. In Experiment 3ab Stroop-to–Sentence, fixed effects (Adjective valence × Sentence Type) in linear mixed-effects models of Target looks by region (Pre-adjective, Adjective-Noun, Pronoun).

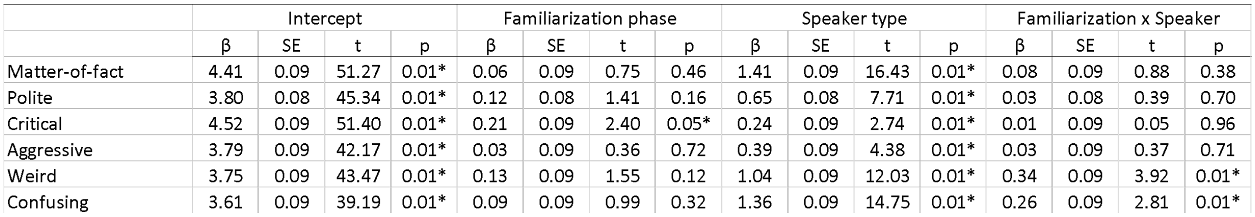
	Pre-adjective				Adjective-noun				Pronoun			
	β	SE	t	p	β	SE	t	p	β	SE	t	p
Intercept	0.51	0.02	30.22	0.01*	0.53	0.01	59.72	0.01*	0.80	0.02	68.72	0.01*
Previous: Stroop	0.01	0.01	0.87	0.38	0.01	0.01	1.18	0.24	0.01	0.01	1.12	0.26
Current: Sentence	0.01	0.01	1.49	0.14	0.02	0.01	3.71	0.01*	0.01	0.01	1.11	0.27
Stroop x Sentence	0.02	0.01	1.90	0.06	0.01	0.01	1.23	0.22	0.01	0.01	1.89	0.06

Table 6. In Experiment 3ab Sentence-to-Stroop, fixed effects (Previous: Sentence Type x Current: Stroop type) in regression model of log reaction times.



	β	SE	t	p
Intercept	6.53	0.01	483.01	0.01*
Previous: Sentence	0.01	0.01	0.53	0.60
Current: Stroop	0.02	0.01	5.39	0.01*
Sentence x Stroop	0.01	0.01	3.50	0.01*

Table 7. In Experiment 4, fixed effects (Familiarization phase × Speaker Type) in linear mixed-effects models of speaker ratings across six dimensions.



	Intercept				Familiarization phase				Speaker type				Familiarization x Speaker			
	β	SE	t	p	β	SE	t	p	β	SE	t	p	β	SE	t	p
Matter-of-fact	4.41	0.09	51.27	0.01*	0.06	0.09	0.75	0.46	1.41	0.09	16.43	0.01*	0.08	0.09	0.88	0.38
Polite	3.80	0.08	45.34	0.01*	0.12	0.08	1.41	0.16	0.65	0.08	7.71	0.01*	0.03	0.08	0.39	0.70
Critical	4.52	0.09	51.40	0.01*	0.21	0.09	2.40	0.05*	0.24	0.09	2.74	0.01*	0.01	0.09	0.05	0.96
Aggressive	3.79	0.09	42.17	0.01*	0.03	0.09	0.36	0.72	0.39	0.09	4.38	0.01*	0.03	0.09	0.37	0.71
Weird	3.75	0.09	43.47	0.01*	0.13	0.09	1.55	0.12	1.04	0.09	12.03	0.01*	0.34	0.09	3.92	0.01*
Confusing	3.61	0.09	39.19	0.01*	0.09	0.09	0.99	0.32	1.36	0.09	14.75	0.01*	0.26	0.09	2.81	0.01*