ABSTRACT

Title of Dissertation:	TOWARD A PSYCHOLINGUISTIC MODEL OF IRONY COMPREHENSION
	Rachel Michelle Adler, Doctor of Philosophy, 2018
Dissertation directed by:	Jared M. Novick, Assistant Professor, Hearing and Speech Sciences Yi Ting Huang, Assistant Professor, Hearing and Speech Sciences

This dissertation examines how listeners reach pragmatic interpretations of irony in real-time. Over four experiments I addressed limitations of prior work by using fine-grained measures of time course, providing strong contexts to support ironic interpretations, and accounting for factors known to be important for other linguistic phenomena (e.g., frequency). Experiment 1 used a visual world eye-tracking paradigm to understand how comprehenders use context and frequency information to interpret irony. While there was an overall delay for ironic utterances compared to literal ones, the speed of interpretation was modulated by frequency. Participants interpreted frequent ironic criticisms (e.g., "fabulous chef" about a bad chef) more quickly than infrequent ironic compliments (e.g., "terrible chef" about a good chef). In Experiment 2A, I tested whether comprehending irony (i.e., drawing a pragmatic inference) differs from merely computing the opposite of an utterance.

The results showed that frequency of interpretation (criticisms vs. compliments) did not influence processing speed or overall interpretations for opposites. Thus, processing irony involves more than simply evaluating the truth-value condition of an utterance (e.g., pragmatic inferences about the speaker's intentions). This was corroborated by Experiment 2B, which showed that understanding irony involves drawing conclusions about speakers in a way that understanding opposites does not. Opposite speakers were considered weirder and more confusing than ironic speakers. Given the delay in reaching ironic interpretations (Exp. 1), Experiments 3 and 4 examined the cognitive mechanics that contribute to inhibiting a literal interpretation of an utterance and/or promoting an ironic one. Experiment 3 tested whether comprehending irony engages cognitive control to resolve among competing representations (literal vs. ironic). Results showed that hearing an ironic utterance engaged cognitive control, which then facilitated performance on a subsequent high-conflict Stroop trial. Thus, comprehenders experience conflict between the literal and ironic interpretations. In Experiment 4, however, irony interpretation was not facilitated by prior cognitive control engagement. This may reflect experimental limitations or late-arriving conflict. I end by presenting a model wherein access to the literal and ironic interpretations generates conflict that is resolved by cognitive control. In addition, frequency modulates cue strength and generates delays for infrequent ironic compliments.

TOWARD A PSYCHOLINGUISTIC MODEL OF IRONY COMPREHENSION AND PRODUCTION

by

Rachel Michelle Adler

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Assistant Professor Jared M. Novick, Chair
Assistant Professor Yi Ting Huang, Chair
Professor Colin Phillips
Professor Rochelle S. Newman
Associate Professor Robert L. Slevc

© Copyright by
Rachel Michelle Adler
2018

Acknowledgements

First, I would like to thank my amazing advisors, Jared Novick and Yi Ting Huang. Jared, you have been the most supportive, kind, and encouraging advisor I could ask for. Thank you for always making time for me and being my cheerleader. You represent all of the things about academia that I know I will miss. I would never have made it without you! Yi Ting, thank you for providing endless feedback and input throughout my time at UMD. You have helped shaped me as a scientist, for which I am very grateful.

I would also like to thank my family for providing infinite support and motivating me to keep going whenever the going got tough. Mom, dad, and Danielle, thank you for always believing in me and making me feel loved. And thank you to all of my wonderful siblings – Greg, Samantha, Sabrina, Sydney, and Skylar. Hanging out with you guys never fails to make me smile, and I will always want to spend more time playing Telestrations (female body builder) or One Night Ultimate Werewolf. I love all of you so much!

Of course, thank you to my friends, Alix Kowalski, Alia Lancaster, Zeke Lancaster, Chris Heffner, Eric Pelzl, and Zoe Schlueter, for the many game nights that kept me sane and happy. No one besides a fellow graduate student quite understands what getting your doctorate is like, and you guys have helped me navigate through the tough times. In addition, thank you to Chelsea Ezzo for all of your support and for always lending an ear when I needed to vent. I am also incredibly grateful to my boyfriend, Niko Anderson, for his unconditional support

and encouragement. No one else would put up with my hangry rants or times of panic, and I am lucky to have had you by my side.

I would also like to thank the many people at UMD who have helped make this dissertation possible. My committee members, Bob Slevc, Rochelle Newman, and Colin Philips, have provided useful feedback and helped make this dissertation the best it could be. The staff at the Language Science Center (Shevaun Lewis, Tess Wood, and Caitlin Eaves) helped me to hone my non-academic skills and enabled me to pursue a career outside of academia. Pam Komarek, thank you for always making sure I was on track and funded! I would also like to thank Karly Schwarz and Hannah Sichel for all of their help in designing and running these experiments. Karly, you are the most incredible undergraduate and masters student I've been lucky enough to work with, and I certainly wouldn't have finished in time without your support. And I would like to thank Chris Heffner and Amritha Mallikarjun for being my Ironic Ike and Literal Lucy – you are the best enthusiastic/sarcastic speakers around!

Finally, I have been supported by a variety of funding sources, including a Dissertation Assistantship from NACS and a Dean's Fellowship from HESP. In addition, this material is based upon work supported by the National Science Foundation under Grant No.1449815.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Speakers use irony to express information contradictory to what they say. For example, if a speaker says, "What a fabulous chef Fred is," we might conclude that Fred cooks well (literal interpretation). However, if we just saw Fred make a mess, we would instead infer that he is a terrible chef (ironic interpretation). Speakers are ironic for a range of social purposes: testing and bolstering common ground (Brown, 1995), saving face when making criticisms (Dews & Winner, 1995; Jorgensen, 1996), increasing politeness (Attardo, 2001), and alienating others (Colston, 1997). For a comprehender, accurately interpreting irony requires the consideration of context and the speaker's goals. For example, in order for the comprehender to understand that "fabulous" is being used ironically, he must consider the mess Fred made (the context). In addition, the comprehender must take into account the identity of the speaker, who may have a tendency to use irony and is unlikely to be complimenting Fred. Thus, listeners have to interpret irony by way of a pragmatic inference. When the speaker makes a positive statement about Fred's cooking, the listener must consult the context to recognize this is a false statement.

According to Grice (1975), the mismatch between context and a speaker's utterance leads the listener to generate an inference that the speaker must have actually meant the opposite of what they said. This is because speakers are cooperative and generally do not utter false statements (the Maxim of Quality). When the listener hears the speaker describe Fred as "fabulous," he consults the context (Fred's mess) and realizes that the speaker's utterance is false. The violation of the Maxim of Quality, combined with the listener's assumption that the speaker is

being cooperative, leads the listener to draw a pragmatic inference that the speaker must have meant something other than what he said: namely, the ironic interpretation. In this way, listeners make use of this assumption of cooperation to correctly interpret ironic utterances.

While it is known that listeners must consult the context and speaker goals to understand irony, *when* and *how* they do so is unknown. The experiments described in this dissertation will help us to better understand how comprehenders process irony as well as the semantics-pragmatics interface more broadly. In particular, these experiments address key limitations of existing work on irony comprehension, such as by using more fine-grained measures of time course, providing strong contexts to support ironic interpretations, and considering factors known to be important for other linguistic phenomena, such as frequency. By overcoming these limitations, the present dissertation helps explain how comprehenders use context to reach ironic interpretations.

In the remainder of this chapter I will first discuss the social functions that irony serves. I will then provide an overview of the existing accounts of irony comprehension, as well as their strengths and weaknesses. To address these weaknesses, I will present two possible accounts of irony processing that will be tested in this dissertation. Finally, I will preview the four experiments of this dissertation.

Social functions of irony

Given that irony may lead to comprehension difficulty, one might wonder why speakers bother to use irony to begin with. Irony carries with it certain

pragmatic functions that literal language may not convey. One pragmatic function that irony bestows is to enhance relationships and group affiliation. Irony builds solidarity by conveying negative judgments about others (Attardo, 2001). This creates an "in-group feeling" (p. 173), wherein the speaker and listener feel more attached and familiar with each other. Irony thus creates the sense between interlocutors that "you and I are the same" (Lakoff, 1990, p. 173). One reason for this may be that irony can serve as a way to test the amount of shared knowledge (Brown, 1995). As a result, irony use can highlight and bolster common ground: the set of beliefs, knowledge, and experience shared by the speaker and listener. For example, a speaker may use irony to comment on a problem relevant to both the speaker and listener, thereby highlighting the shared nature of the experience (and strengthening the interlocutors' bond). The role of irony in strengthening relationships seems to be further supported by the fact that individuals in closer relationships are more likely to use irony (Pexman & Zvaigzne, 2004; Sally, 2003). In one study, participants were asked to read brief stories ending in an ironic utterance and had to rate the appropriateness of the utterance (Kreuz, Kassler, Coppenrath, & Allen, 1999). For example, the speaker might say, "you sure were the hit of the party!" after the listener fell asleep at a party. Participants rated ironic utterances as more appropriate when there was greater common ground between the speaker and listener (e.g., a husband and wife as opposed to two strangers). Thus, having more common ground between interlocutors seems to make ironic utterances more appropriate.

Additionally, irony can be used for criticisms in order to save face for the speaker and make the speaker seem less rude (Dews & Winner, 1995; Jorgensen,

1996; cf. Colston, 1997). Saving face fulfills the speaker's desire to be liked and respected (Brown, 1995). For example, Dews and Winner (1995) had subjects read short stories ending in ironic or literal criticisms (e.g., "you're so considerate" said to someone who just stole something). Participants rated ironic criticisms as less critical than literal ones, as well as less likely to negatively affect the speaker-listener relationship. Furthermore, because using irony to make a criticism is less face damaging than outright aggression, it can be used to make criticisms more polite (Attardo, 2001; Boylan & Katz, 2013). Indeed, comprehenders rate ironic criticisms as more polite and positive than literal ones (Boylan & Katz, 2013). Similarly, irony can be used for humor, and comprehenders perceive ironic speakers as being funnier (Dews, Kaplan, & Winner, 1995; Gibbs, Bryant, & Colston, 2014; Kumon-Nakamura, Glucksberg, & Brown, 1995; Matthews, Hancock, & Dunham, 2006). Irony can therefore be used for a range of social purposes to influence the relationship between interlocutors.

Current accounts of irony comprehension

While it is known that the social functions of irony can increase group affiliation, it is less clear how listeners reach ironic interpretations through the coordination of various sources of evidence (e.g., context and what the utterance literally means). Work on irony comprehension has largely focused on whether the literal interpretation of an ironic utterance is necessarily initiated and completed prior to the ironic one. According to the *standard pragmatic view*, comprehending irony occurs in stages (Cutler, 1976; Dews & Winner, 1999; Giora et al., 2007). The listener first accesses the context-independent literal interpretation of an ironic

utterance. Then, if there is a mismatch with the literal interpretation and the context (e.g., speaker identity, prior events), the listener reaches the ironic interpretation by computing the opposite of the literal meaning of the utterance. On this view, context is not used until later in processing, after the literal interpretation has already been reached. For example, after hearing "What a fabulous chef Fred is," the listener would first interpret the utterance as meaning that Fred cooks well. The listener would then consult the context (e.g., Fred's mess) and revise the initial (incorrect) literal interpretation. Because there are multiple stages involved, the standard pragmatic view predicts that ironic utterances will be understood more slowly than literal ones.

In contrast, the *direct access view* posits that context interacts with lexical processing early on (Gibbs, 1986; Ivanko & Pexman, 2003). That is, if the ironic interpretation of an utterance is supported by the context, the listener can access this interpretation without the need to first access the literal meaning. Thus, the listener would observe Fred's mess (and perhaps the speaker's likelihood to use irony) and would therefore reach the ironic interpretation immediately upon hearing "fabulous." In contrast to the standard pragmatic view, the direct access view claims that ironic utterances should be understood as quickly as (or even faster than) literal ones.

Finally, the *graded salience hypothesis* combines aspects of the standard pragmatic and direct access views. According to this hypothesis, the more salient interpretation of an utterance is always accessed first (Giora, 1997). An utterance is "salient" if it is encoded in the mental lexicon, that is, if it can be interpreted without considering contextual information. The salience of a word or phrase may be

influenced by its frequency, familiarity, or conventionality (Giora, Fein, & Schwartz, 1998). The ironic meaning of familiar ironies is lexicalized and therefore accessed directly, but unfamiliar ironies are processed in stages: salient literal meaning first, then ironic. For example, if "what a fabulous chef" is a familiar irony (e.g., based on frequency of use), then listeners should only access the negative, ironic interpretation of the utterance (Fred is a bad chef). However, if "what a fabulous chef" is unfamiliar, then participants should first access the literal interpretation (Fred is a good chef), and then the ironic one (Fred is a bad chef). Thus, while this account predicts that sometimes irony is accessed in stages (as in the standard pragmatic view), it may also be accessed directly under certain circumstances.

Empirical evidence for comprehension accounts

Evidence for these three accounts of irony processing is mixed. Some work seems to support the standard pragmatic view (Cutler, 1976; Dews & Winner, 1999; Giora et al., 2007). For example, Dews and Winner (1999) had participants read 2-3 sentence stories that ended in one character making either a literal or an ironic comment. The participants' task was to judge whether the speaker meant something positive or something negative. Dews and Winner found that participants took longer to judge ironic utterances compared to literal ones. This delay seems to suggest that comprehending irony requires additional stages (i.e., accessing the literal interpretation). Importantly, because participants' reaction times were only measured for the judgments they made after reading the entire sentence, it is difficult to determine the precise cause of the delay. For example, while it could be the case that the literal interpretation preceded the ironic one, it is also possible that the two

interpretations were accessed simultaneously. This particular paradigm cannot disentangle these two possibilities. In addition, due to the design of the study, the observed delay could either arise from the processing of the ironic utterance itself, or from the task of making a positive/negative judgment about ironic utterances.

Other work has focused on the speaker's linguistic tendencies in order to establish context for irony. For example, Giora et al. (2007) presented subjects with dialogues between two friends. In each dialogue, one speaker produced an ironic utterance and then later, produced another utterance (target sentence) that could be interpreted ironically or literally, based on the dialogue context. If context is used early, then participants should the read ironic utterances as quickly as the literal ones. However, if context is not integrated until later in processing (leading to a multi-step process), reading times for ironic utterances should be slower than for literal ones. The results supported the latter alternative. Specifically, target sentence reading times were slower following ironic-biasing dialogues compared to literal-biasing ones, suggesting that context effects may be delayed during comprehension (leading to delayed irony comprehension). Critically, however, it is possible that the contextual cues provided were not strong enough for the subjects to use. Subjects did not receive any explicit information about the speakers and they encountered new speakers with each dialogue.

Multiple studies have also supported the direct access account of irony comprehension. In one of the earliest psycholinguistic studies of irony comprehension, Gibbs (1986) had participants read stories that ended in literal or ironic utterances. He found that participants were faster to read ironic comments

(e.g., “You are a fine friend”) compared to their literal counterparts (e.g., “You are a bad friend”), which was interpreted to mean that comprehenders do not need to access the literal interpretation prior to the ironic one. Importantly, whole sentence reading times may be too coarse grained to accurately answer questions about real-time processing; a more fine-grained measure of online interpretations may be necessary. Similarly, Ivanko and Pexman (2003) had participants complete a self-paced reading task with stories ending in literal or ironic statements. In line with Gibbs’ (1986) findings, participants read the ironic statements as quickly as the literal ones given a strongly supportive context. However, the conclusions we can draw from these results are somewhat limited for two reasons. First, the authors compared reading times for positive adjectives (in the ironic condition) with negative adjectives (in the literal condition). Given the fact that negative words tend to be read more slowly than positive ones (Ivanko & Pexman, 2003; Kuchinke et al., 2005; Schact & Sommer, 2009), this could lead to artificially faster reading times for the ironic condition. The second reason is that all of the contexts were negative, which meant that all of the ironic utterances were ironic criticisms (the more frequent type of irony). It is possible then that their findings might be different for less frequent ironic compliments, as frequency has well-known effects on the time course of comprehension at multiple levels of representations (Duffy, Morris, & Rayner, 1988; Forster & Chambers, 1973; Rayner & Raney, 1996).

Finally, there is also evidence for the graded salience account. For example, Giora and Fein (2007) had participants complete a lexical decision task after reading familiar or less familiar ironies in irony- or literal-biasing contexts. In irony-biasing

contexts, less familiar ironies were interpreted literally initially (150ms post-offset) and ironically later (1,000ms post-offset). However, familiar ironies were interpreted both literally and ironically initially (150ms post-offset) and later (1,000ms post-offset). Thus, they argue that salient meanings are always processed initially, regardless of contextual information. Importantly, these findings may also be consistent with the standard pragmatic view, where the literal interpretation is reached prior to the ironic one. Specifically, it is possible that for familiar ironies, the literal analysis was reached before 500ms, rather than both the literal and ironic interpretations being reached simultaneously. Distinguishing between these possibilities requires a more temporally fine-grained measure of online irony processing. Filik, Leuthold, Wallington, and Page (2014) conducted a later experiment using eye-tracking while reading. Participants read short stories containing a target utterance that was either ironic or literal (as determined by the prior context), and was disambiguated by a single word. They manipulated the materials such that half of the ironic utterances were familiar and half were unfamiliar. In line with the graded salience account, Filik et al. found that compared to their literal counterparts, unfamiliar ironies were read more slowly than familiar ones. The delay for unfamiliar ironies was reflected in gaze duration on the critical disambiguating word as well as the post-critical region. Filik et al. argue that the effects observed in the post-critical region indicate that participants were reanalyzing the target sentence as being ironic. Critically, it is difficult to determine whether the post-critical region delays were due to a reanalysis of an initial literal interpretation, or the simultaneous activation of both the literal and ironic interpretations. In

addition, Filik et al. did not take into account irony frequency in their experiment. Based on the sample stimuli provided in their paper, all of the contexts were negative, which meant that all of the ironic utterances were the more frequent ironic criticisms. It is possible that the difference in reading times for familiar and unfamiliar ironies could disappear if the ironies used were the less frequent ironic compliments. Finally, it is hard to identify a clear-cut pattern of results that could be used to falsify the graded salience hypothesis. Because "salience" isn't clearly defined or necessarily measurable, it is difficult to determine how the theory could be falsified at all, which therefore diminishes its potential theoretical contribution to the field.

Thus, the extent to which these various findings support the standard, direct, or graded salience accounts of irony comprehension is unclear. In particular, there are open questions about the time-course of irony interpretation (i.e., when listeners use various evidential cues to inform an ironic analysis of an utterance). Furthermore, existing work has not adequately addressed other factors that may influence the speed of irony comprehension, such as the frequency of certain types of irony (i.e., ironic criticisms vs. compliments).

Moving beyond traditional models

Although traditional models of irony processing consider the relationship between the literal and ironic interpretations, they make contradicting predictions regarding the role of different sources of information over time. That is, they do not explain how linguistic and contextual information are accessed and integrated on a fine-grain scale during real-time processing. For example, there may be circumstances in which listeners might need to use the semantic analysis to guide use

of pertinent contextual cues (e.g., speaker identity). In addition, these accounts do not always consider how other properties of irony, such as frequency, are considered during irony comprehension. There are two forms of irony that differ in their frequency. When a positive utterance is uttered to convey a negative sentiment about an individual, this is known as an *ironic criticism* (e.g., saying that Fred is a fabulous chef after he made a mess in the kitchen). A speaker produces an *ironic compliment*, by contrast, when they use a negative statement to ironically describe a successful individual (e.g., if Fred had made a beautiful cake, a speaker could say, "What a terrible chef Fred is"). Ironic criticisms are generally more frequent or conventional than ironic compliments (Gibbs, 2000). It is well known that frequency information is important in literal language comprehension (Duffy et al., 1988; Forster & Chambers, 1973; Rayner & Raney, 1996). However, it is unknown whether listeners track and make use of frequency information for pragmatic phenomena like irony.

Thus, there are a number of limitations to the existing accounts of irony comprehension that need to be addressed. As described above, there are conflicting findings about when and how the ironic interpretation is accessed. While some work indicates that the ironic interpretation may be accessed immediately and directly (e.g., Gibbs, 1986; Ivanko & Pexman, 2003), other work suggests the literal interpretation is sometimes accessed prior to the ironic one (Dews & Winner, 1999; Giora et al., 2007). It is possible that other factors, such as irony frequency, may help account for these discrepancies, though this has yet to be tested. An additional reason for these conflicting findings may be the fact that many studies use only coarse-grained measures of processing (e.g., Gibbs, 1986; Giora & Fein, 2007). As a result, they

may not accurately capture how the literal and ironic interpretations are accessed in real-time. Furthermore, it is important to consider the type of context that listeners are provided in these experiments. The majority of the studies reviewed here present a brief, written story involving two or more characters, followed by an ironic utterance. Strengthening these contexts will reveal whether findings that favor the standard pragmatic view actually characterize irony processing in general, or whether they are driven by contexts that are insufficiently irony-biasing. Finally, to assess whether irony comprehension is slower than literal utterance comprehension, it is necessary to compare ironic utterances to appropriate literal baselines. Since negative adjectives are processed more slowly than positive ones (Kuchinke et al., 2005; Schact & Sommer, 2009), ironic utterances with negative adjectives should be compared to literal utterances with negative adjectives (and positive ironic with positive literal).

Two models of irony comprehension

Two possible processing accounts of irony comprehension are shown in *Figure 1*, which seek to explain how speaker goal information (context) and frequency are used to direct comprehenders' processing commitments for irony. As the figure depicts, interpreting a literal utterance involves connecting the utterance (e.g., "What a fabulous chef Fred is"; white arrows) to its corresponding lexical entries. These entries encode some frequency information (e.g., "fabulous" is frequent, while "terrible" is less frequent). The lexical entries then connect to higher order conceptual representations that represent meaning. However, how frequency and contextual information are used in irony comprehension is up for debate.

One possibility is that both the literal and ironic interpretations of an ironic utterance are available initially, and that contextual information (e.g., speaker goal and identity information) is recruited immediately (Early Access account, see *Figure 1*A). According to this account, irony is stored in the lexicon in a similar way as the two meanings of a homophone. As it does for homophones, frequency would modulate the speed of interpretation of irony. This is because there are extra connections from the lexical entries to the concepts that carry the opposite valence (e.g., "fabulous" to bad). Moreover, some uses of irony are simply more frequent than other uses: speakers are more inclined to use a positive adjective to criticize a negative event (Gibbs, 2000). The thickness of the gray arrows in *Figure 1* represents the frequency of the interpretation: ironic criticisms are more frequent than ironic compliments. On this account, more frequent ironic criticisms would be accessed simultaneously with the literal interpretation; providing context to support the ironic interpretation would wipe out any potential delays. However, there would be a slowdown for less frequent ironic compliments. That is, the comprehender should be slower to reach the ironic interpretation of "terrible" used to compliment Fred, compared to "fabulous" used to criticize Fred. These findings should hold even with a strong, irony-biasing context (e.g., strong knowledge of speaker identity and tendencies to use irony). This would be similar to the findings for word recognition, where both context and frequency influence processing (Duffy et al., 1988; Swinney, 1979).

Alternatively, it is possible that the ironic interpretation is only accessed after the literal one is reached, and contextual information is not integrated until later in

processing (Late Access account, see *Figure 1*B). On this account, ironic interpretations are not stored in the lexicon. Even if contextual cues are present before the speaker says anything (e.g., via speaker identity), comprehenders may have difficulty tracking and coordinating these cues before engaging in semantic analysis. This would be similar to some findings for scalar implicatures (Bott & Noveck, 2004; Huang & Snedeker, 2009; Huang & Snedeker, 2011; Tomlinson, Bailey, & Bott, 2013), where semantic analysis mediates pragmatic inferencing. In this case, the processor would not immediately consult these contextual cues, even with strongly irony-biasing contexts. Because context is only taken into account late in processing, frequent ironic criticisms and less frequent ironic compliments would both be accessed equally slowly. Thus, evidence of an overall delay for irony would suggest that, unlike homophones, ironic interpretations are not stored in the lexicon.

The experiments described in this dissertation test these models. That is, they seek to determine how context (e.g., speaker goals) and frequency influence the speed with which a comprehender reaches an ironic interpretation, and how the semantic and pragmatic interpretations of irony are accessed over time.

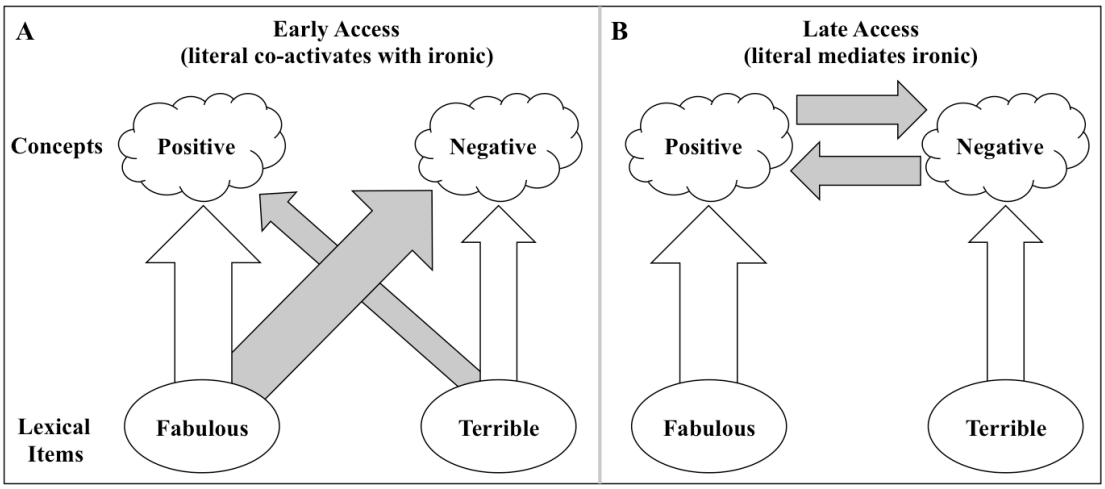

Figure 1. Two possible models of irony comprehension: (A) Early Access and (B) Late Access. White arrows represent literal interpretations; gray arrows represent ironic interpretations. The thickness of the gray arrows represents the frequency of the interpretation. On the Early Access account, the literal and ironic interpretations are accessed simultaneously, and irony frequency mediates processing time. On the Late Access account, the literal interpretation is accessed prior to the ironic one, and irony frequency does not mediate processing time.

Remaining questions and specific aims

As described above, prior research leaves open critical questions about how comprehenders use context and frequency to reach ironic interpretations in real-time. To address these questions it is necessary to use a fine-grained measure of online processing (e.g., eye-tracking during listening) where the speed of irony processing is compared to an appropriate baseline. For example, negative adjectives used ironically should be compared to negative adjectives used literally. In addition, these experiments must manipulate irony frequency by comparing ironic criticisms with irony compliments. Finally, ironic utterances should be presented with strong contextual support to ensure that any observed delays are not simply due to insufficient context.

In this dissertation, Experiment 1 makes use of the visual world eye-tracking paradigm. Participants see events featuring two characters (e.g., Sally bakes a beautiful cake while Fred makes a mess). They then hear either a literal or ironic speaker describe Fred or Sally using a positive or negative adjective ("What a *fabulous*/*terrible* chef s/he is"). Participants are told that one speaker is always literal and one speaker is always ironic (these are distinguished by speaker gender). The participants' task is to select the character that the speaker describes (the Target). While Targets are unambiguously identified by pronoun gender, adjective valence (combined with context) can provide an earlier cue. When the ironic speaker speaks, fixations to the target referent reveal the extent to which frequency guides early interpretation. To briefly foreshadow the results, the overall magnitude of Target looks is greater for literal utterances compared to ironic ones, but there is no effect of frequency on magnitude of looks. In contrast, the rate of interpretation is higher for ironic criticisms than ironic compliments.

It is important to note here that prosody was not manipulated in this experiment (or in any subsequent experiments). The literal recordings for one list became the ironic recordings for another, and so participants heard the same audio files regardless of whether the speaker was literal or ironic. The speakers were told to use an enthusiastic tone, which was potentially consistent with irony but not inconsistent with literal utterances. It was not expected that using the same prosody for ironic and literal utterances would be problematic, given evidence that there may not be an "ironic tone of voice" (Attardo, Eisterhold, Hay, & Poggi, 2003; Bryant & Fox Tree, 2005; Kreuz & Roberts, 1995; Rockwell, 2000) and that listeners do not

rely on particular vocal cues to identify verbal irony (Bryant & Fox Tree, 2005). Indeed, Cutler (1974) wrote that, "if cues from the context are strong enough, no intonational cues are necessary" (p. 117). Furthermore, Kreuz and Roberts (1995) argue that an ironic tone of voice might even be detrimental to comprehension when there is high common ground between the speaker and listener. However, it is possible that if speakers were specifically asked to sound ironic, it would alter the results presented here. If it is the case that there *are* particular acoustic cues to irony that are not captured here, it could mean that the present results might not generalize to those circumstances.

Experiments 2A and 2B follow up on Experiment 1 to test whether participants are actually generating the pragmatic inferences necessary for irony comprehension. It is possible for participants to complete the task while only computing truth conditions (e.g., always do the opposite of what the speaker said), rather than considering speaker goals. In Experiment 2A, the ironic speaker is replaced with a speaker who always "says the opposite" of what he/she means. Thus, while the truth conditions of the utterances are the same, the pragmatic inferences required for interpretation are not. If delays from Experiment 1 are generated by all non-literal interpretations, then the same effects found in Experiment 1 should be found in Experiment 2A. However, if delays found in Experiment 1 are specific to irony (where pragmatic inferences are necessary), the patterns should not hold in Experiment 2A. These findings suggest that the observed delays are in fact specific to irony. In Experiment 2B, I evaluate the social-pragmatic functions of irony by examining the inferences that comprehenders make about ironic speakers and their

goals. This experiment differentiates inferences made about opposite speakers and ironic speakers (who also say the opposite of what they mean, but do so for a social pragmatic purpose). The results indicate that comprehenders draw different conclusions about ironic speakers than opposite speakers, thereby corroborating the different patterns of findings of Experiments 1 and 2A.

Experiments 3 and 4 test the source of delays for irony comprehension found in Experiment 1. In syntactic processing, the simultaneous activation of two interpretations leads to conflict that consequently engages cognitive control (Hsu & Novick, 2016; Kan et al., 2013). Irony may behave like syntactic ambiguity, in which case comprehending irony would generate conflict and lead to the engagement of cognitive control. The engagement of cognitive control would then facilitate performance on a subsequent cognitive control task. This is tested in Experiment 3. In Experiment 4 I examine whether the delay for irony comprehension can be mitigated by the prior engagement of cognitive control. This would again be similar to findings for syntactic ambiguity (Hsu & Novick, 2016).

Taken together, these four experiments will help us to better understand how comprehenders process irony as well as the semantics-pragmatics interface more broadly. In particular, these experiments address some of the prior limitations to previous work by (a) using a fine-grained measure of time course, (b) examining frequent criticisms as well as infrequent compliments, (c) comparing ironic utterances to appropriate literal baselines, and (d) providing participants strong context to better isolate the source of delays for irony.

Chapter 2: Experiment 1

Overview

In order to compare the Early Access and Late Access accounts, the present chapter examines how frequency and context interact over time in verbal irony comprehension. This issue may be informed by work on word recognition, where context and frequency inform comprehenders' real-time processing decisions (Duffy et al., 1988; Swinney, 1979). In a seminal study on homophones, Swinney (1979) found that even when context biased toward one meaning of a homophone, both meanings were activated early in processing. However, only the relevant meaning remained activated given a delay. These findings suggest that context does not have an immediate effect on lexical access, but instead only plays a role in a post-access decision process, where one of the two activated meanings is selected. This could similarly be the case for irony: both the literal and ironic meanings could be accessed initially, and context (e.g., about speaker goals or identity) might only play a role later to select the relevant ironic interpretation. Duffy et al. (1988) compared equibiased homophones, where both meanings are equally frequent (e.g., "pitcher"), to non-equibiased homophones, where one meaning is more frequent than the other (e.g., "port"). They found that, given a strong preceding context, only the relevant meaning of an equibiased homophone was activated. However, there was a slowdown when comprehenders were given context supporting the subordinate meaning of a non-equibiased homophone, indicating that context led to competition between the two activated meanings. Thus, both context and frequency play a role in processing lexical ambiguity (see *Figure 2*).

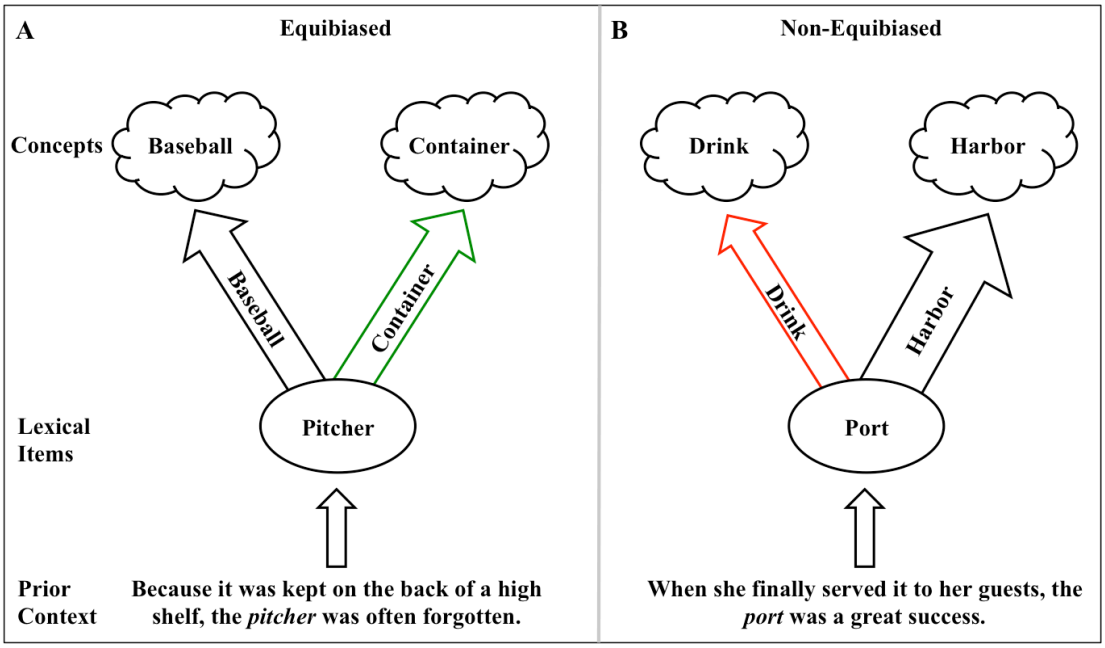

Figure 2. Models of homophone processing for (A) equibiased and (B) non-equibiased nouns given disambiguating context (bottom of diagram). The thickness of the arrows represents the frequency of the interpretation. For equibiased words, context selects the appropriate meaning immediately and there is no increase in processing time (green arrow). For non-equibiased words, context supporting the subordinate meaning leads to the simultaneous activation of the dominant (relevant) meaning and the subordinate (irrelevant) meaning, thereby increasing processing time (red arrow). Note: disambiguating context taken from Duffy et al. (1988).

Findings from the word recognition literature may inform our understanding of how context and frequency interact during irony comprehension. If processing irony is like processing homophones, and ironic representations are stored in the lexicon, then both context (e.g., speaker identity) and frequency (criticisms vs. compliments) should interact in processing, as in the Early Access account. Because criticisms are more frequent, providing strong contextual support should speed up processing time. However, because ironic compliments are less frequent, they should be slower than criticisms even with context. On the other hand, it is possible that

irony will not be processed in the same way as homophones, and that context will only be taken into account late in processing (Late Access account). This would indicate that irony is *not* stored in the lexicon. In this case, the frequency of irony will not affect comprehension speed: frequent ironic criticisms and less frequent ironic criticisms will both be accessed equally slowly. Because both ironic criticisms and compliments must go through the literal interpretation, even adding strong contextual support should not facilitate processing.

There are several requirements for an experiment that could distinguish between the Early and Late Access accounts. First, it is necessary to use a fine-grained measure of time course, such as a visual world eye-tracking paradigm. The visual world eye-tracking paradigm is highly sensitive to probabilistic information about language use and context (e.g., Dahan, Magnuson, & Tanenhaus, 2001; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002). For example, Tanenhaus and colleagues (1995) showed that relevant non-linguistic information (e.g., a listener's visuo-contextual environment) immediately informs real-time reference resolution. A listener's eye movements are closely time-locked to incoming speech, providing important temporal information about listeners' ongoing interpretation commitments that can shed light on the underlying mental architecture that supports language comprehension. Thus, this method is capable of assessing how listeners consult context and frequency information during the real-time interpretations of irony.

In addition, the experiment would need to provide context to sufficiently distinguish between ironic and literal utterances. One way to do this would be to

introduce a literal speaker and an ironic speaker. Prior work indicates that listeners can use speaker identity and linguistic tendencies to aid in reference resolution (Arnold, Hudson Kam, & Tanenhaus, 2007; Grodner & Sedivy, in press; Nappa & Arnold, 2014) and pragmatic inference interpretation (Bergen & Grodner, 2012). Furthermore, explicit information about the speaker has been shown to influence real-time interpretation commitments. For example, Arnold, Pancani, and Rosa (in press) showed that when a speaker is described as being distracted, listeners rely less heavily on acoustic prominence to determine whether a referent has previously been mentioned or is new to the conversation. Similarly, Arnold and colleagues (2007) had participants complete a reference resolution task while listening to instructions from a disfluent speaker. Listeners were biased to interpret the referred to object as unfamiliar when the speaker was disfluent. However, when the speaker was described as having object agnosia that caused difficulty with naming objects, the unfamiliarity bias was reduced. There is also evidence that comprehenders make use of explicit information about a speaker's occupation in deciding whether an utterance is meant ironically as well as in recalling the utterance later (Katz & Pexman, 1997; Pexman & Olineck, 2002).

Finally, this experiment needs to manipulate the frequency of irony. Whether irony is stored in the lexicon (like homophones) can be assessed by testing how frequency modulates interpretation speed. Thus, the experiment must include both ironic criticisms as well as ironic compliments. This can be achieved by manipulating the adjective valence. The use of positive adjectives would indicate an ironic criticism (e.g., "what a fabulous chef he is"), while negative adjectives would

indicate ironic compliments (e.g., "what a terrible chef she is"). In addition, the time course of interpretation for these two types of utterances would need to be compared to appropriate baselines. This is because negative words (e.g., "terrible") tend to be interpreted more slowly than positive ones (e.g., "fabulous"; Kuchinke et al., 2005; Schact & Sommer, 2009). Thus, the experiment must also include literal uses of positive and negative adjectives.

The goal of Experiment 1 was to distinguish between the Early Access and Late Access accounts. To do so, participants were presented with vignettes describing visually depicted events featuring two different-gender characters (*Figure 3*) while their eye movements were tracked. For example, Sally baked a beautiful cake while Fred made a mess. Next, subjects heard an utterance that described Fred or Sally using a positive or negative adjective. The participants' task was to select the character that the speaker described (Target character). Participants were told that these target utterances could be produced by one of two speakers: an ironic speaker and a literal speaker. The ironic and literal speakers were disambiguated by gender so that participants could determine whether the utterance was ironic or literal as soon as the speaker began to speak.

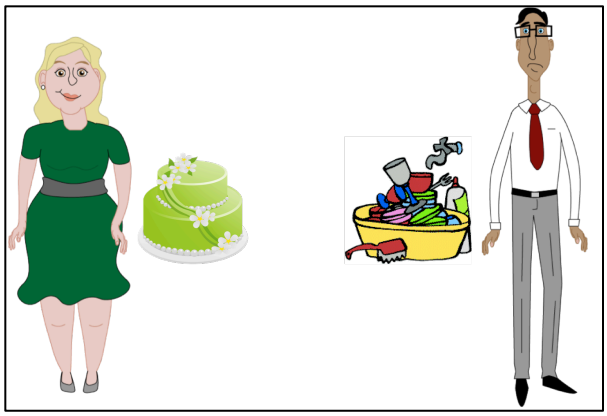

Figure 3. Example display for Experiments 1 and 2A.

Continuing with the example used earlier, the ironic speaker would say either (a) "What a fabulous chef he is" or (b) "What a terrible chef she is." Utterance (a) would constitute an ironic criticism, while (b) would constitute an ironic compliment. The literal speaker would say either (c) "What a fabulous chef she is" or (d) "What a terrible chef he is." In addition to generating the irony frequency manipulation (criticisms vs. compliments), both positive and negative adjectives were used because, as described above, negative words (e.g., "terrible") tend to be interpreted more slowly than positive ones (e.g., "fabulous"; Kuchinke et al., 2005; Schact & Sommer, 2009). Thus, any delays observed for ironic compliments could not simply be attributed to the use of a negative adjective.

While Targets were unambiguously identified by pronoun gender, adjective valence could provide an earlier cue. Critically, when the literal speaker speaks, participants should look to Sally shortly after "*fabulous*" and Fred after "*terrible*." However, looks to Fred should be delayed compared to looks to Sally, given evidence that negative words are interpreted more slowly than positive words. Importantly,

when the ironic speaker speaks, fixations to the target referent will reveal the extent to which frequency guides early interpretation.

Method

Participants

Thirty-five undergraduates from the University of Maryland participated in this experiment for either pay ($5) or course credit. One participant's data were not analyzed due to equipment malfunction, and two more participants were excluded due to low task accuracy (under 75% for one or more conditions). Thus, there were a total of 32 participants included in the analyses (27 female, mean age = 19.3, range = 18-22).

Materials and procedure

Participants were seated in front of a computer and their eye movements were recorded. At the beginning of the experiment, participants were told that on each trial, they would hear brief stories describing two characters, Fred and Sally. For example, Sally baked a beautiful cake while Fred made a mess. These descriptions were accompanied by images on the screen depicting the events (*Figure 3*). Then, participants would hear a new speaker describe of the two characters using a positive or negative adjective ("What a *fabulous*/*terrible* chef s/he is"). One speaker would always be ironic and one speaker would always be literal. Their task was to select the character that the speaker described (Target character). See Appendix A for the complete task instructions. Prior to beginning the experiment, participants completed three practice trials (two literal, one ironic).

The experiment employed a 2 x 2 x 2 design, with adjective valence (positive, negative) and interpretation type (literal, ironic) as within-subjects factors, and speaker gender (ironic male and literal female; ironic female and literal male) as a between-subjects factor (Table 1). Participants saw a total of twenty critical items. Four versions of each critical item were generated by manipulating the interpretation type (ironic or literal) and the adjective valence (positive or negative). This generated four presentation lists, such that each list contained five items in each condition (positive literal, positive ironic, negative literal, negative ironic), and each item appeared once in each list. For example, one participant would hear "What a fabulous chef he is," another would hear "What a fabulous chef she is," a third would hear "What a terrible chef he is," and a fourth would hear "What a terrible chef she is." See Appendix B for a list of all critical items. An additional 24 filler trials were constructed in which both characters were either successful or unsuccessful at a given action. Therefore, for filler items, participants could not identify the Target until the pronoun. The order of presentation of critical and filler trials was randomized across critical trials.

Table 1

Display Crosses Adjective Valence, Interpretation, and Speaker Gender

	Positive adjective ("*fabulous chef*")	Negative adjective ("*terrible chef*")
Literal interpretation	(a) Positive Target	(b) Negative Target
Ironic interpretation	(c) Negative Target	(d) Positive Target

Note. (c) is an ironic criticism and (d) is an ironic compliment. For half the subjects (n = 16), the male speaker was ironic and the female speaker was literal; for the other half, the genders were reversed.

Four additional lists were generated by manipulating the speaker gender. For half of the participants, the male speaker was ironic and the female speaker was literal. For the other half, the female speaker was ironic and the male speaker was literal. The same recordings were used, regardless of speaker identity. That is, the recordings for the four lists where the male speaker was ironic were also used for the four lists where the male speaker was literal. The speakers who pre-recorded the literal and ironic statements were instructed to use an enthusiastic tone of voice, which was felicitous with an ironic interpretation, but did not preclude a literal one. Therefore, all participants heard the same recordings. All of the recordings were produced using a Shure SM-51 microphone in a sound-attenuated room. It was not expected that using the same prosody for ironic and literal utterances would be problematic, given evidence that there is no ironic tone of voice (Attardo et al., 2003; Bryant & Fox Tree, 2005; Kreuz & Roberts, 1995; Rockwell, 2000) and that listeners do not rely on particular vocal cues to identify verbal irony (Bryant & Fox Tree, 2005).

Analysis

Eye movements were divided into three time regions of interest:

(1) *Pre-adjective*: The pre-adjective region began at the onset of the critical utterance and ended just before the onset of the adjective (e.g., "What a"). This region served as a baseline measure of looks to the display before any adjective or pronoun information. In this region there should be approximately equal fixations to the Target and Distractor characters.

(2) *Adjective-noun*: The adjective-noun region began at the onset of the adjective and ended just before the onset of the pronoun (e.g., "fabulous chef"). In this region, the comparison between the four conditions would reveal the reveal the extent to which frequency guides early interpretation. It was expected that in the literal condition, fixations to the Target would be faster for positive adjectives than negative ones. For the ironic condition, there were two alternatives. If context is used early and frequency mediates irony comprehension speed (Early Access account), then there should be an interaction between adjective, interpretation, and time, such that there would be a greater delay for ironic criticisms (compared to literal compliments) than there would be for ironic compliments (compared to literal criticisms). However, if context is used late in processed and frequency does not mediate comprehension speed (Late Access account), then there should be no adjective by interpretation by time interaction. That is, the delay for ironic criticisms and ironic compliments should be approximately equal as compared to the literal uses of those adjectives.

(3) *Pronoun*: The pronoun region began at the onset of the critical utterance and ended at the offset of the statement (e.g., "he is"). In this region, the Target character was unambiguously resolved by the gender of the pronoun. Therefore, it was expected that looks to the Target would be greater than looks to the Distractor, regardless of condition.

For each region, looks prior to 200ms were removed to account for the time it takes to launch a saccade (Allopenna, Magnuson, & Tanenhaus, 1998; Matin, Shao, & Boff, 1993). In addition, all incorrect trials (i.e., where the subject did not click on the Target) were removed from analysis; this corresponded to 1.93% of all trials. I coded the two characters as Target (who the speaker described) and Distractor (the other character on the screen). The primary dependent measure examined the proportion of looks to the Target, which was calculated as Target looks divided by Target plus Distractor looks. The proportion of looks to the Target was averaged across 50ms windows.

Target looks were analyzed in R (version 3.3.2; R Core Team, 2016) with linear mixed effects models using the *lme4* package (version 1.1-12; Bates, Maechler, Bolker, & Walker, 2015). The *lmerTest* package was used to compute p-values using Satterthwaite's approximation for denominator degrees of freedom (version 2.0-32; Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2015). For each region, a linear effects model was constructed containing adjective valence (positive, negative), interpretation type (literal, ironic), and time from adjective onset (in 50ms bins) as fixed effects. (I additionally constructed a model including time, adjective, and speaker gender for ironic trials only. The goal of this analysis was to ensure that listeners did not interpret irony differently when it was produced by a male speaker versus a female one. However, since there was no significant three-way interaction [$p = .36$], speaker gender was not included in any further analyses.) Adjective valence and interpretation type were both deviation coded, while time was included as a continuous factor. For the adjective-noun region, an additional model was

constructed that only included adjective valence (positive, negative) and interpretation type (literal, ironic) to better understand the overall, time-independent magnitude differences.

For all analyses, I first fit maximal models that included both random slopes and intercepts for subjects and items (Barr, Levy, Scheepers, & Tily, 2013). However, when the maximal models did not converge, I constructed all lower-level models that did converge and compared these models using a chi-square likelihood ratio test. The model with the lowest Akaike's information criterion (AIC) value was deemed the best-fitting model and was then used for analysis.

Results

Figure 4 plots the average looks to the Target over time during each region of interest (pre-adjective, adjective-noun, and pronoun) by condition (positive literal, positive ironic, negative literal, negative ironic). *Figure 5* plots the average looks to the Target, collapsing across time, during each region of interest by condition. As the two figures show, the mean proportion of looks to the Target during the pre-adjective region was 0.49 (SE = 0.08). The linear model parameters for the pre-adjective region are shown in Table 2. As expected, there were no significant main effects or interactions during this region (ps > .10).

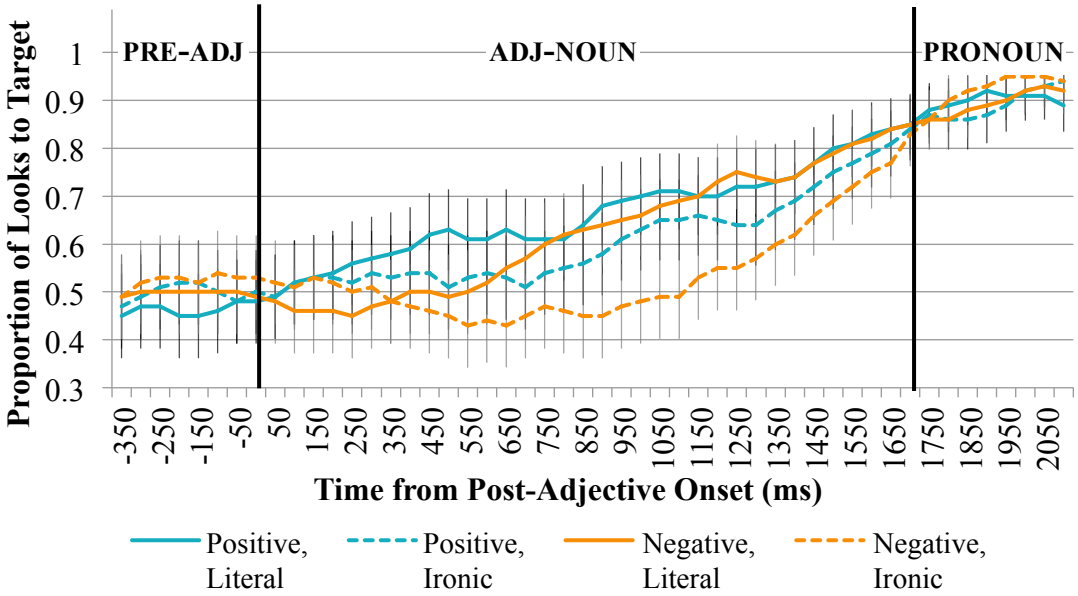

Figure 4. Average proportion of looks to the Target character in 50-ms intervals post-adjective onset by region (pre-adjective, adjective-noun, pronoun), adjective valence (positive, negative), and interpretation type (literal, ironic). The first vertical line represents the adjective onset and the second vertical line represents the average onset time of the pronoun. Bars represent standard errors.

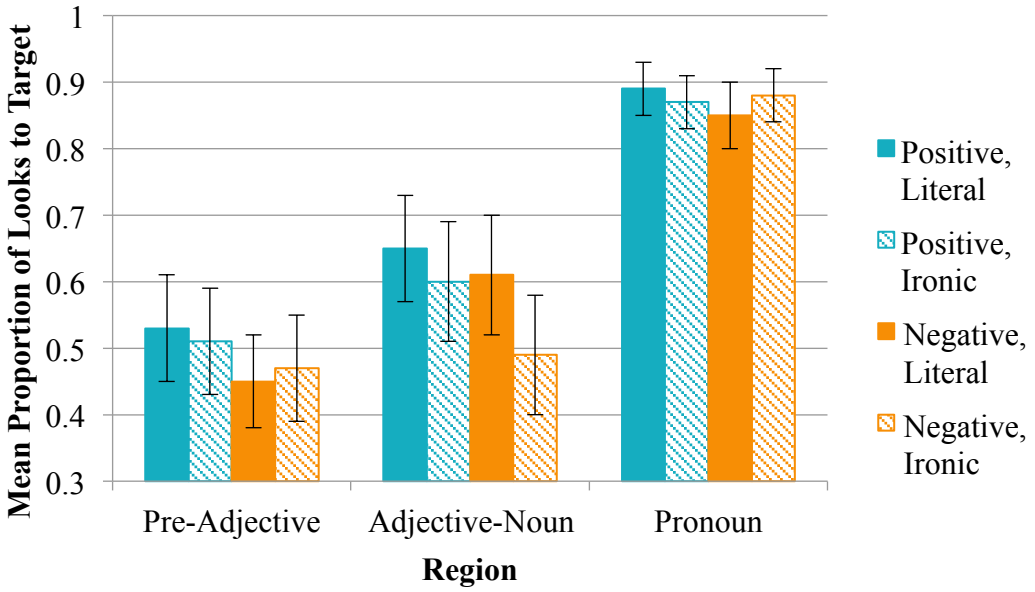

Figure 5. Mean proportion of looks to Target character in by region (pre-adjective, adjective-noun, pronoun) and condition (positive literal, positive ironic, negative literal, negative ironic). Bars represent standard errors.

Table 2

Model Parameters for Pre-Adjective Region: Exp. 1

	Parameters			
Factor	β	SE	t	p
Intercept	0.50	0.03	17.41	< .0001
Adjective	-0.02	0.03	0.87	.39
Interpretation	-0.02	0.03	0.92	.36
Time	0.00	0.00	-0.13	.90
Adjective x Interpretation	0.03	0.05	-0.53	.60
Adjective x Time	-0.01	0.01	-0.85	.40
Interpretation x Time	-0.00	0.01	-0.03	.97
Adjective x Interpretation x Time	0.01	0.01	1.44	.15

Note. Model specification: TargetLooks ~ Adjective * Interpretation * Time + (1|Subject) + (1|Item)

As described above, two linear mixed effects models were constructed for the adjective-noun region. The first model examined the overall magnitude of Target

looks by adjective and interpretation (Table 3, *Figure 5*). As the table and figure show, the proportion of looks to the Target during the critical adjective-noun region was significantly higher when the adjective was positive versus negative (0.62 vs. 0.55). This was confirmed by a significant main effect of adjective, $F(1, 580.51) = 8.42$, $p < .01$. Thus, consistent with prior work on semantic analysis (Kuchinke et al., 2005; Schact & Sommer, 2009), access to positive adjectives was faster than negative adjectives. In addition, the proportion of Target looks was significantly higher when the interpretation was literal versus ironic (0.63 vs. 0.55). This was also confirmed by a significant main effect of interpretation, $F(1, 580.56) = 10.57$, $p < .01$. There was no interaction between adjective and interpretation ($p = .24$). Thus, looks to the Target were decreased for ironic utterances compared to literal ones, regardless of irony frequency.

Table 3

Model Parameters for Adjective-Noun Region without Time: Exp. 1

	Parameters			
Factor	β	SE	t	p
Intercept	0.59	0.02	32.99	< .0001
Adjective	-0.07	0.03	-2.90	< .01
Interpretation	-0.08	0.03	-3.25	< .01
Adjective x Interpretation	-0.06	0.05	-1.19	.24

Note. Model specification: TargetLooks ~ Adjective * Interpretation + (1|Subject) + (1|Item)

The parameters for the model including adjective, interpretation, and time are shown in Table 4. As can be seen in Table 4 and *Figure 4*, the proportion of looks to the Target increased over time across all conditions during the adjective-noun region. This was confirmed by a significant main effect of time, $F(1, 16{,}684) = 358.78$, $p <$

.0001. In addition, there was a significant interaction between interpretation and time: looks to the Target increased more quickly for literal utterances compared to ironic ones, $F(1, 16{,}668) = 13.44$, $p < .001$. Importantly, there was a significant interaction between adjective, interpretation, and time, $F(1, 16{,}671) = 14.00$, $p < .001$. To better explore this three-way interaction, I generated two additional linear mixed effects models: one with positive adjectives only and one with negative adjectives only. For the positive adjective model, there was no interaction between interpretation and time ($p = .83$). Thus, looks to the Target increased at approximately the same rate for positive adjectives used literally and positive adjectives used ironically. In contrast, the negative adjective model revealed a significant time by interpretation interaction, $F(1, 8{,}040.6) = 26.19$, $p < .0001$. In other words, negative ironic utterances (i.e., less frequent compliments) were delayed compared to negative literal utterances. Taken together, these results show no effect of irony frequency in the magnitude of Target looks, but a significant effect of frequency on the rate of increase of Target looks.

Table 4

Model Parameters for Adjective-Noun Region with Time: Exp. 1

	Parameters			
Factor	β	SE	t	p
Intercept	0.48	0.02	28.39	< .0001
Adjective	-0.09	0.03	-3.15	< .01
Interpretation	-0.04	0.01	-3.17	< .01
Time	0.01	0.00	18.13	< .0001
Adjective x Interpretation	0.04	0.03	1.69	.09
Adjective x Time	0.00	0.00	1.77	.08
Interpretation x Time	-0.00	0.00	-3.61	< .001
Adjective x Interpretation x Time	-0.01	0.00	-3.96	< .0001

Note. Model specification: TargetLooks ~ Adjective * Interpretation * Time + (1|Subject) + (1|Item) + (1+Adjective|Subject)

By the pronoun region, participants had largely converged on the Target across conditions: the proportion of looks to the Target was greater than 85% in all conditions. The complete modeling results are shown in Table 5. As in the adjective-noun region, there was a significant interaction between adjective, interpretation, and time, $F(1, 10{,}593) = 9.13$, $p < .01$. This appears to be driven by the fact that Target looks for the negative ironic condition were greater than looks for the negative literal condition, while Target looks in the positive ironic condition were not consistently different from looks in the positive literal condition.

Table 5

Model Parameters for Pronoun Region: Exp. 1

	Parameters			
Factor	β	SE	t	p
Intercept	0.50	0.02	20.83	< .0001
Adjective	-0.28	0.04	-7.00	< .0001
Interpretation	-0.26	0.03	-7.73	< .0001
Time	0.01	0.00	21.34	< .0001
Adjective x Interpretation	-0.20	0.07	-3.02	< .01
Adjective x Time	0.01	0.00	7.58	< .0001
Interpretation x Time	0.01	0.00	7.70	< .0001
Adjective x Interpretation x Time	0.01	0.00	3.02	< .01

Note. Model specification: TargetLooks ~ Adjective * Interpretation * Time + (1|Subject) + (1|Item) + (1+Adj|Item)

Discussion

Prior work on irony leaves open critical questions regarding how frequency and context are used over time during verbal comprehension. According to the Early Access account, context and frequency interact early during processing. Thus, more

frequent ironic criticisms should be processed more quickly (relative to their literal baselines) than less frequent ironic compliments. However, according to the Late Access account, context is only taken into account later in processing. Therefore, irony frequency should not modulate comprehension speed: frequent ironic criticisms and less frequent ironic criticisms should be accessed equally slowly.

The data from this experiment show that in terms of the overall magnitude of Target looks, irony frequency does not mediate interpretation. Target looks were lower for ironic utterances compared to literal ones, regardless of irony frequency. However, the rate analyses indicate that the speed with which ironic utterances are processed is mediated by their frequency. Comprehenders were slower to reach the less conventional use of irony (compliments) than the more conventional use (criticisms), compared to their literal baselines. When irony was used in its more conventional form, there was no difference in the rate of Target looks compared to the literal condition.

These findings do not seem to definitively support the Early *or* Late Access accounts. While irony frequency had an early effect on the rate of Target looks, it did not affect the overall magnitude. This seems to suggest that overall, ironic utterances are interpreted more slowly than literal utterances. This is in line with prior work on irony processing indicating that ironic utterances take longer to comprehend (Filik & Moxey, 2010; Schwoebel, Dews, Winner, & Srinivas, 2000). However, using the more frequent form of irony (criticisms) benefits interpretation speed to a greater degree than less frequent forms of irony (compliments).

The present experiment builds on prior work in two key ways. First, the context provided to participants was very strong. Participants were told in advance that one speaker would always be ironic, thereby generating a perfectly reliable cue. Thus, the delay observed for irony could not be attributed to insufficient contextual support. (And indeed, participant performance did not change over the course of the experiment: in an analysis including adjective, interpretation, and half, there was no three-way interaction [$p = .41$]. Participants did not make better use of the speaker identity cue in the second half compared to the first.) Second, the present experiment manipulated frequency. The rate of interpretation was modulated by irony frequency, suggesting that prior work that ignored irony frequency may not have accurately captured the time course of irony comprehension.

The patterns found here differ from those for other instances where multiple meanings are linked to a single phonological form, such as homophones. For homophones, frequency and context interact early during processing (Duffy et al., 1988). While the present experiment showed early frequency effects in the rate of irony interpretation, comprehenders' use of contextual information (i.e., speaker identity) did not boost the magnitude of ironic interpretations to the same level as literal interpretations. One possible reason for this discrepancy is the level of representations involved. For homophones, both potential meanings exist at the lexical level (e.g., "bank" for the financial institution and "bank" for the side of the river). In contrast, the two possible meanings present for irony occur across different levels of representation: the semantic analysis (literal interpretation) and the pragmatic analysis (ironic interpretation). It is possible that context can only be used

early in processing when the multiple meanings to be disambiguated operate at the same level of linguistic representation, as for homophones.

Importantly, it is possible that participants in this experiment were not necessarily computing the ironic interpretations. That is, rather than treating the ironic speaker as ironic, participants could treat the speaker as always saying the opposite of what they mean. Thus, instead of generating a pragmatic inference about ironic speaker and their reason for being ironic, the participant could simply choose the character that is consistent with the opposite of "fabulous chef." Indeed, irony is often defined as simply saying the opposite of what you mean. However, irony carries certain pragmatic functions that saying the opposite does not, such as being polite (Attardo, 2001; Dews & Winner, 1995; Jorgensen, 1996) or humorous (Dews et al.,1995; Gibbs et al., 2014; Kumon-Nakamura et al., 1995; Matthews et al., 2006). If listeners in the present experiment were simply accessing the opposite meaning, but not the pragmatic interpretation, then the present experiment would not tell us anything about irony comprehension. It is therefore important to determine whether listeners *did* actually draw the pragmatic inference for ironic utterances in this experiment.

Experiment 2A seeks to address this potential limitation by comparing the present results to those obtained for an "opposite" speaker. The exact same paradigm, materials, and procedure from Experiment 1 were used in Experiment 2A, except that instead of one speaker always being ironic, that speaker was described as always saying the opposite of what he/she meant. If the patterns in Experiment 2A are the same as those observed in Experiment 1, then we can conclude that the

Experiment 1 participants were not computing the pragmatic inferences typically involved in irony comprehension. However, if the eye movement patterns are different, it means that the participants in Experiment 1 were not just doing the opposite of what the ironic speaker said. Experiment 2B further addressed this issue by examining the pragmatic inferences comprehenders make about ironic speakers vs. opposite speakers. The goal of this experiment was to determine whether listeners draw pragmatic inferences about ironic speakers that differ from those for opposite speakers. If this difference exists, it would provide additional evidence that irony is more than just computing opposites, and that the listeners in Experiment 1 were making pragmatic inferences about ironic speakers that they would not make for opposite speakers.

Chapter 3: Experiments 2A and 2B

Overview

As described above, the participants in Experiment 1 were slower to comprehend ironic criticisms compared to ironic compliments. Thus, frequency modulated the speed of ironic interpretations. However, the findings are also potentially consistent with the idea that the participants were simply "doing the opposite" of what the ironic speaker said, rather than computing the pragmatic inferences involved in irony comprehension. As a result, it may be the case that the delay for irony and the existence of a frequency effect are simply due to comprehenders computing opposites, not irony. Experiments 2A and 2B were therefore conducted to test whether listeners in the current paradigm interpret the ironic utterances as ironic, not just as opposites.

Experiment 2A sought to determine whether the categorical manipulation of speakers (speakers are either always literal or always ironic) led participants to use strategies, such as always doing the opposite of what the ironic speaker says, rather than actually computing the pragmatic inference generated by irony. Experiment 2A was identical to Experiment 1, except that the ironic speaker was replaced with a speaker who always said "the opposite" of what he/she actually meant. If the participants in Experiment 1 were just doing the opposite of what the ironic speaker said, then it was expected that Experiment 2A would replicate the findings from Experiment 1. That is, there would be a delay for opposites that would be modulated by frequency. However, if comprehending ironic utterances requires pragmatic processes beyond simply computing the truth value (e.g., considering speaker goals),

then there would be a different time course for the opposite conditions in Experiment 2A compared to the ironic conditions in Experiment 1. In particular, while there might still be a delay for opposite utterances compared to literal ones, there would not be an effect of frequency.

The goal of Experiment 2B was to determine the inferences that comprehenders make about speakers who say the opposite of what they mean (opposite speakers), compared to speakers who also say the opposite of what they mean, but do so for a social pragmatic purpose (ironic speakers). Thus, this experiment sought to further test whether comprehenders view ironic speakers in the same way as opposite speakers. To do so, participants were shown brief videos constructed from Experiments 1 and 2A. That is, Fred and Sally were shown completing various actions and then a speaker described one of the two characters. Half of the participants were told that one speaker would always be literal and the other would always be ironic, and the other half were told that one speaker would always be literal and the other would always say the opposite of what he/she meant. As in Experiments 1 and 2A, the speakers were disambiguated by gender.

After viewing four videos with each speaker (four literal plus four ironic, or four literal plus four opposite), the participants were asked to rate each speaker on six adjectives: *critical*, *polite*, *matter-of-fact*, *aggressive*, *confusing*, and *weird*. These adjectives were selected to represent potential goals or communicative functions for each linguistic form. Thus, ironic speakers were expected to be rated highly on *critical* and *polite*, literal speakers were expected to be rated highly on *matter-of-fact* and *aggressive*, and opposite speakers were expected to be rated highly on *confusing*

and *weird*. The degree to which the participants' ratings differed for ironic and opposite speakers would reveal the different inferences listeners make about ironic and opposite speakers.

Experiment 2A Method

Participants

Thirty-nine undergraduates from the University of Maryland participated in this experiment for either pay ($5) or course credit. Seven participants were excluded due to low task accuracy (under 75% for one or more conditions). Thus, there were a total of 32 participants included in the analyses (22 female, mean age = 19.8, range = 18-22).

Materials and procedure

Participants completed the same task in Experiment 2A as they did in Experiment 1, except that the ironic speaker was replaced with an "opposite" speaker who always said the opposite of what they meant. The same recordings were used as in Experiment 1. As in Experiment 1, participants completed three practice trials (two literal, one opposite) before beginning the experiment.

Analysis

Analysis was performed in the same way as in Experiment 1. As in Experiment 1, all incorrect trials (i.e., where the subject did not click on the Target) were removed from analysis; this corresponded to 1.87% of all trials.

Experiment 2A Results

Figure 6 plots the average looks to the Target over time during each region of interest (pre-adjective, adjective-noun, and pronoun) by condition (positive literal, positive opposite, negative literal, and negative opposite). *Figure 7* plots the average looks to the Target during each region of interest (pre-adjective, adjective-noun, and pronoun) by condition (positive literal, positive opposite, negative literal, and negative opposite). As the two figures show, the mean proportion of looks to the Target during the pre-adjective region was 0.49 (*SE* = 0.09). The linear model parameters for the pre-adjective region are shown in Table 6. All main effects and interactions were non-significant except for a time by adjective interaction, $F(1, 4{,}402.2) = 5.35$, $p < .05$. This was unexpected, as there was no adjective information available in this region. However, as shown in *Figure 6*, any differences between conditions were resolved prior to the adjective onset.

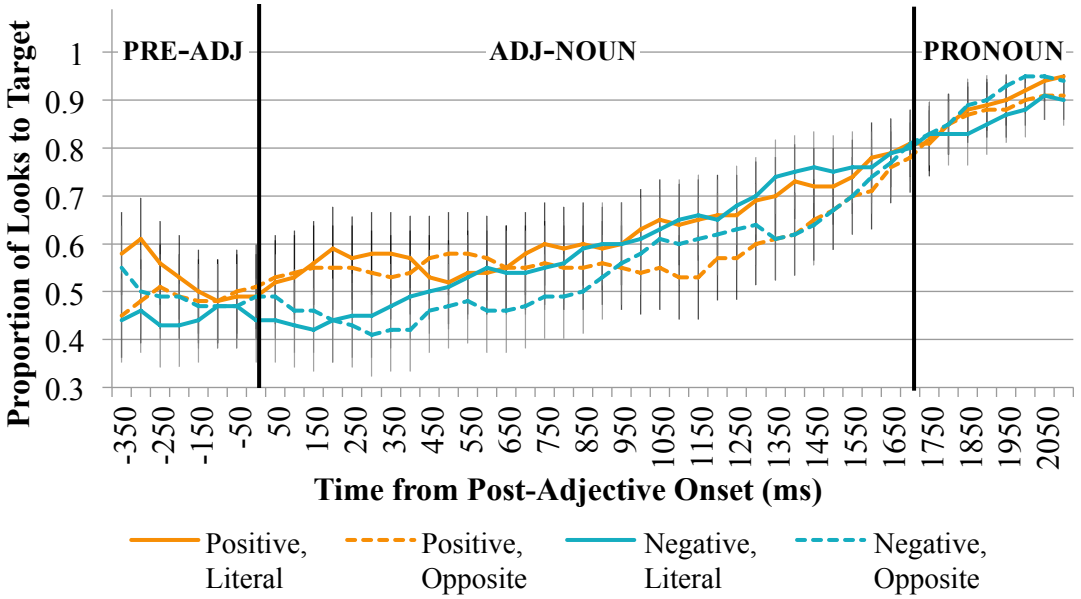

Figure 6. Average proportion of looks to the Target character in 50-ms intervals post-adjective onset by region (pre-adjective, adjective-noun, pronoun), adjective valence

(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.

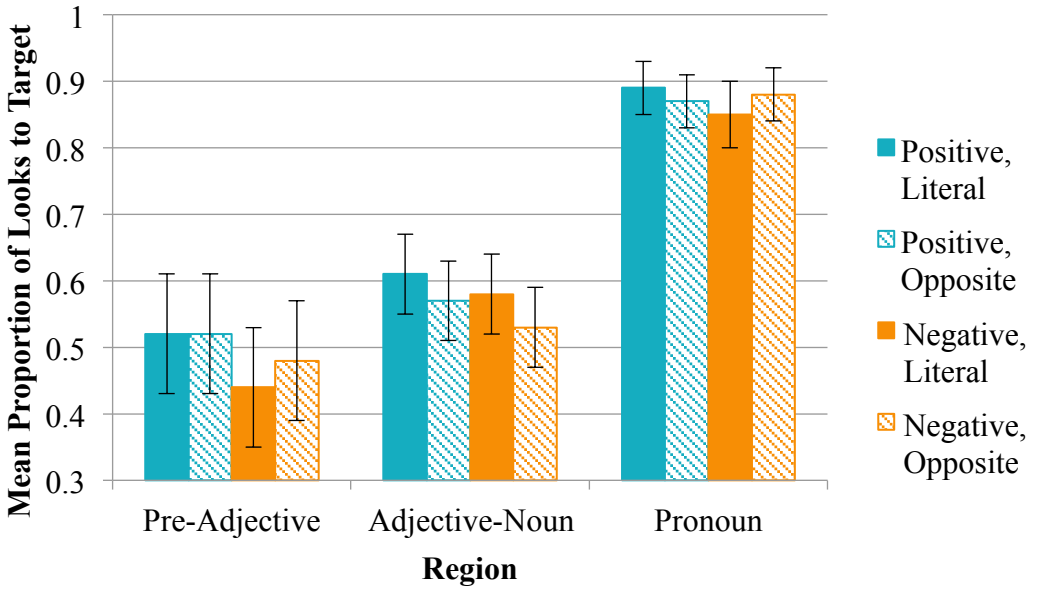

Figure 7. Mean proportion of looks to Target character in by region (pre-adjective, adjective-noun, pronoun) and condition (positive literal, positive opposite, negative literal, negative opposite). Bars represent standard errors.

Table 6

Model Parameters for Pre-Adjective Region: Exp. 2A

	Parameters			
Factor	β	SE	t	p
Intercept	0.48	0.03	16.86	< .0001
Adjective	-0.02	0.04	-0.53	.60
Interpretation	-0.01	0.02	-0.52	.60
Time	0.00	0.00	1.19	.23
Adjective x Interpretation	0.07	0.05	1.37	.17
Adjective x Time	-0.01	0.01	-2.31	.02
Interpretation x Time	0.01	0.01	1.63	.10
Adjective x Interpretation x Time	-0.01	0.01	-0.74	.46

Note. Model specification: TargetLooks ~ Adjective * Interpretation + (1|Subject) + (1|Item) + (1+Adj|Item)

As in Experiment 1, two linear mixed effects models were constructed for the adjective-noun region. The parameters for the model including only adjective and interpretation for the adjective-noun region are shown in Table 7. As the table shows, there were no significant main effects or interactions during this region (*p*s > .10).

Table 7

Model Parameters for Adjective-Noun Region without Time: Exp. 2A

	Parameters			
Factor	β	SE	t	p
Intercept	0.57	0.01	42.63	< .0001
Adjective	-0.04	0.03	-1.45	.15
Interpretation	-0.04	0.03	-1.59	.11
Adjective x Interpretation	-0.02	0.05	-0.29	.78

Note. Model specification: TargetLooks ~ Adjective * Interpretation + (1|Subject) + (1|Item)

The parameters for the model including adjective, interpretation, and time are shown in Table 8. As can be seen in Table 8 and *Figure 6*, the proportion of looks to the Target increased over time across all conditions during the adjective-noun region. This was confirmed by a significant main effect of time, $F(1, 16{,}418) = 273.77$, $p < .0001$. Importantly, there was no significant interaction between adjective, interpretation, and time ($p = 0.96$). That is, the difference between positive literal utterances and positive opposite utterances was equivalent to the difference between negative literal utterances and negative opposite utterances. Thus, in contrast to the results from Experiment 1, frequency does not influence the speed of interpretation for opposites.

Table 8

Model Parameters for Adjective-Noun Region with Time: Exp. 2A

	Parameters			
Factor	β	SE	t	p
Intercept	0.47	0.02	24.84	< .0001
Adjective	-0.10	0.01	-6.88	< .0001
Interpretation	-0.02	0.01	-1.34	.17
Time	0.01	0.00	16.55	< .0001
Adjective x Interpretation	-0.03	0.03	-1.03	.31
Adjective x Time	0.01	0.00	5.06	< .0001
Interpretation x Time	-0.02	0.00	-2.69	.01
Adjective x Interpretation x Time	0.00	0.00	0.05	.96

Note. Model specification: TargetLooks ~ Adjective * Interpretation * Time + (1|Subject) + (1|Item)

By the pronoun region, participants had largely converged on the Target across conditions: the proportion of looks to the Target was greater than 85% in all conditions. The complete modeling results are shown in Table 9. As in the adjective-noun region, there was no interaction between adjective, interpretation, and time (p = 0.41).

Table 9

Model Parameters for Pronoun Region: Exp. 2A

	Parameters			
Factor	β	SE	t	p
Intercept	0.37	0.03	14.48	< .0001
Adjective	-0.05	0.05	-0.99	.32
Interpretation	-0.14	0.04	-4.07	< .0001
Time	0.01	0.00	26.59	< .0001
Adjective x Interpretation	0.10	0.07	1.44	.15
Adjective x Time	0.00	0.00	1.35	.18
Interpretation x Time	0.00	0.00	3.86	< .001
Adjective x Interpretation x Time	-0.00	0.00	-0.83	.41

Note. Model specification: TargetLooks ~ Adjective * Interpretation * Time + (1|Subject) + (1|Item) + (1+Adj|Item)

An additional analysis was performed to compare the results from Experiment 2A to the results from Experiment 1. The goal of this analysis was to statistically test whether the difference between these two experiments was driven by the frequency effect observed in Experiment 1 that was not observed in Experiment 2A. Accordingly, a linear mixed effects model was constructed to compare Target looks on ironic trials (Exp. 1) to Target looks on opposite trials (Exp. 2A). Experiment, time, and adjective were included as fixed effects and subject and item were included as random intercepts. The results of this analysis revealed a significant three-way interaction between adjective, time, and experiment, $F(1, 16{,}388) = 13.46$, $p < .001$. This analysis lends further support to the claim that the frequency effect observed in Experiment 1 was not present in Experiment 2A, and thus that ironic interpretations do differ from opposite ones.

Experiment 2B Method

Participants

A total of 192 native English speakers participated in the study through Amazon Mechanical Turk (68 female, 2 no response; mean age = 32.5, range = 19-65). Participants received $2 as compensation for their participation.

Materials and procedure

Participants were told that they would watch eight videos involving different characters, each of which showing events featuring two different-gender characters

(Fred and Sally). Then participants would hear one of two speakers describe Fred or Sally using a positive or negative adjective. One speaker would always be ironic (or opposite) and one speaker would always be literal (see Appendix C for full instructions). The eight videos represented a subset of the critical items from Experiments 1 and 2A. After viewing all eight videos, subjects were asked to rate each speaker on six adjectives. For example, "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?" The six adjectives were: *critical*, *polite*, *matter-of-fact*, *aggressive*, *confusing*, and *weird*. The adjectives were selected to represent potential qualities associated with being ironic (*polite* or *critical*) and/or associated with saying the opposite of what one means (*weird* or *confusing*).

The experiment employed a 2 x 2 x 2 design, with speaker group (A: literal and ironic speakers, B: literal and opposite speakers), interpretation (literal, non-literal), and speaker gender (ironic male and literal female; literal male and ironic female) as between-subject factors. The eight videos were rotated through four conditions: positive literal, positive ironic/opposite, negative literal, and negative ironic/opposite. Crossing speaker group, interpretation, and speaker gender produced a total of sixteen lists. Subjects were randomly assigned to lists, with an equal number per list.

It was expected that participants' ratings of the speakers would differ by speaker group. This would indicate that the inferences subjects are making about speakers are modulated by the speaker's linguistic tendencies.

Analysis

Participants' ratings were analyzed in R (version 3.3.2; R Core Team, 2016) with linear mixed effects models using the *lme4* package (version 1.1-12; Bates et al., 2015). A separate model was created for each of the six adjectives, with subject as random intercept and speaker group (A: literal and ironic speakers, B: literal and opposite speakers), speaker gender (male, female), and interpretation (literal, non-literal) as fixed effects. The *lmerTest* package was used to compute p-values using Satterthwaite's approximation for denominator degrees of freedom (version 2.0-32; Kuznetsova et al., 2015).

Experiment 2B Results

Critical

The mean ratings by adjective are shown in *Figure 8* for Group A and in *Figure 9* for Group B. The model parameters for the adjective *critical* are shown in Table 10. As the table shows, there was a significant main effect of group, such that Group A was rated as more critical overall ($M = 4.70$, $SE = 0.12$) compared to Group B ($M = 4.32$, $SE = 0.13$), $F(1, 374) = 4.72$, $p < .05$. There was additionally a main effect of interpretation: non-literal speakers were rated as more critical ($M = 4.72$, $SE = 0.12$) than literal speakers ($M = 4.30$, $SE = 0.13$), $F(1, 374) = 5.99$, $p < .05$. Finally, there was a significant interaction between interpretation and speaker gender, $F(1, 374) = 6.35$, $p < .05$. This appears to be driven by the fact that for literal speakers, males were judged as more critical than females ($M_{diff} = 0.59$), whereas for non-literal speakers, females were judged as more critical than males ($M_{diff} = 0.29$). Importantly, there was no interaction between interpretation and speaker group. Thus, the

difference between ironic and literal speakers (M_{diff} = 0.45) was equivalent to the difference between opposite and literal speakers (M_{diff} = 0.32).

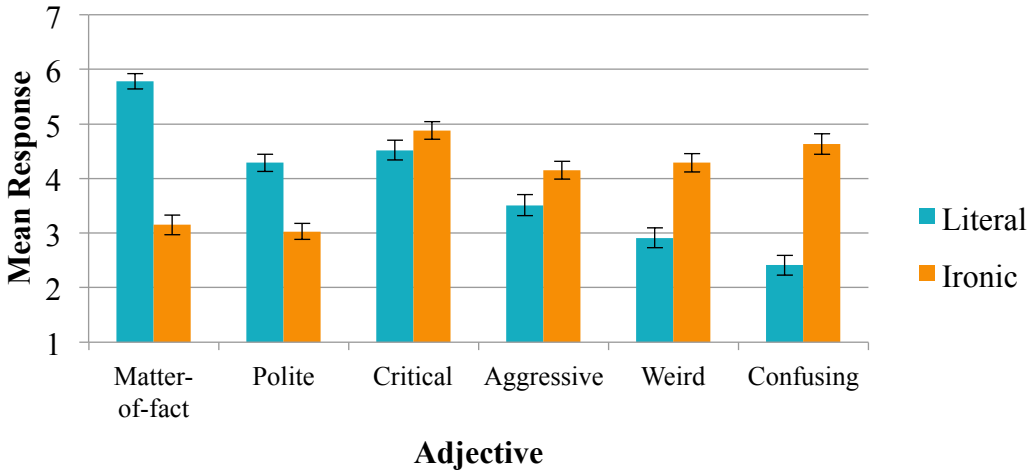

Figure 8. Mean ratings by adjective and speaker (literal and ironic) for Group A. Bars represent standard errors.

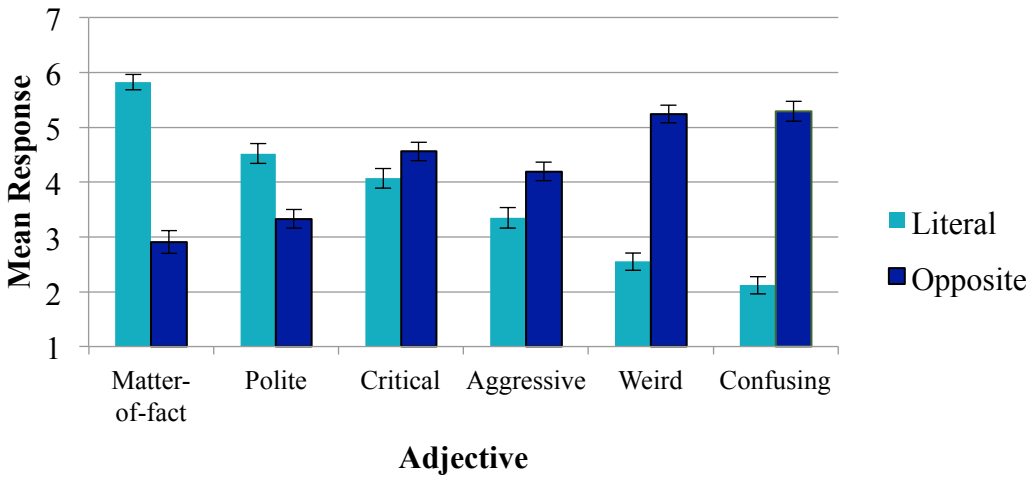

Figure 9. Mean ratings by adjective and speaker (literal and opposite) for Group B. Bars represent standard errors.

Table 10

Model Parameters for Adjective "Critical": Exp. 2B

	Parameters			
Factor	β	SE	t	p

Intercept	4.51	0.09	51.90	< .0001
Group	-0.38	0.17	-2.17	.03
Interpretation	0.43	0.17	2.45	.01
Speaker Gender	0.15	0.17	0.84	.40
Group x Interpretation	0.13	0.35	0.37	.71
Group x Speaker Gender	0.06	0.35	0.18	.86
Interpretation x Speaker Gender	-0.88	0.35	-2.52	.01
Group x Interpretation x Speaker Gender	0.37	0.69	0.54	.59

Polite

The model parameters for the adjective *polite* are shown in Table 11. As the table shows, there was a significant main effect of interpretation: literal speakers were rated as more matter-of-fact (M = 5.80, SE = 0.10) than non-literal speakers (M = 3.03, SE = 0.14), $F(1, 373) = 55.63$, $p < .0001$. There were no other significant main effects or interactions. Thus, the difference in *polite* ratings between ironic and literal speakers (M_{diff} = 1.26) was equivalent to the difference between opposite and literal speakers (M_{diff} = 1.19).

Table 11

Model Parameters for Adjective "Polite": Exp. 2B

	Parameters			
Factor	β	SE	t	p
Intercept	3.79	0.08	46.14	< .0001
Group	0.26	0.16	1.59	.11
Interpretation	-1.23	0.16	-7.56	< .0001
Speaker Gender	-0.28	0.16	-1.69	.09
Group x Interpretation	0.07	0.33	0.21	.83
Group x Speaker Gender	0.34	0.33	1.04	.30
Interpretation x Speaker Gender	0.56	0.33	1.71	.09
Group x Interpretation x Speaker Gender	0.09	0.66	0.13	.89

Aggressive

The model parameters for the adjective *aggressive* are shown in Table 12. As the table shows, there was a significant main effect of interpretation: non-literal speakers were rated as more aggressive ($M = 4.17$, $SE = 0.12$) than literal speakers ($M = 3.43$, $SE = 0.13$), $F(1, 374) = 18.01$, $p < .0001$. Additionally, there was a significant interaction between interpretation and speaker gender, $F(1, 374) = 13.80$, $p < .001$. This appears to be driven by the fact that for literal speakers, males were judged as more aggressive than females ($M_{diff} = 0.84$), whereas for non-literal speakers, females were judged as more critical than males ($M_{diff} = 0.44$). Importantly, there was no interaction between interpretation and speaker group. Thus, the difference in *aggressive* ratings between ironic and literal speakers ($M_{diff} = 0.64$) was equivalent to the difference between opposite and literal speakers ($M_{diff} = 0.84$).

Table 12

Model Parameters for Adjective "Aggressive": Exp. 2B

	Parameters			
Factor	β	SE	t	p
Intercept	3.80	0.09	43.80	< .0001
Group	-0.06	0.17	-0.32	.75
Interpretation	0.74	0.17	4.24	< .0001
Speaker Gender	0.20	0.17	1.15	.25
Group x Interpretation	0.20	0.35	0.58	.56
Group x Speaker Gender	-0.21	0.35	-0.59	.56
Interpretation x Speaker Gender	-1.29	0.35	-3.72	< .001
Group x Interpretation x Speaker Gender	-0.87	0.69	-1.25	.21

Confusing

The model parameters for the adjective *confusing* are shown in Table 13. As the table shows, there was a significant main effect of interpretation, such that non-literal speakers were rated as more confusing ($M = 4.96$, $SE = 0.13$) compared to literal speakers ($M = 2.27$, $SE = 0.12$). There was additionally a main effect of speaker gender: female speakers were rated as more confusing ($M = 3.85$, $SE = 0.16$) than male speakers ($M = 3.37$, $SE = 0.16$). Finally, there was a significant interaction between group and speaker gender, $F(1, 375) = 6.92$, $p < .01$. This appears to be driven by the fact that the difference between literal and ironic speakers (Group A; $M_{diff} = 2.22$) was greater than the difference between literal and opposite speakers (Group B; $M_{diff} = 3.17$). Thus, compared to their literal baselines, opposite speakers were judged as being more confusing than ironic speakers.

Table 13

Model Parameters for Adjective "Confusing": Exp. 2B

	Parameters			
Factor	β	SE	t	p
Intercept	3.61	0.09	40.42	< .0001
Group	0.19	0.18	1.06	.29
Interpretation	2.70	0.18	15.08	< .0001
Speaker Gender	-0.49	0.18	-2.73	.01
Group x Interpretation	0.94	0.36	2.63	.01
Group x Speaker Gender	-0.65	0.36	-1.82	.07
Interpretation x Speaker Gender	-0.08	0.36	-0.22	.82
Group x Interpretation x Speaker Gender	-0.26	0.72	-0.36	.72

Weird

The model parameters for the adjective *weird* are shown in Table 14. As the table shows, there was a significant main effect of interpretation, such that non-literal speakers were rated as weirder ($M = 4.77$, $SE = 0.12$) compared to literal speakers ($M = 2.73$, $SE = 0.12$), $F(1, 375) = 144.60$, $p < .0001$. There was additionally an interaction between interpretation and group, $F(1, 375) = 14.85$, $p < .001$. This is driven by the fact that the difference between literal and ironic speakers (Group A; $M_{diff} = 1.38$) was greater than the difference between literal and opposite speakers (Group B; $M_{diff} = 2.69$). Thus, compared to their literal baselines, opposite speakers were judged as being weirder than ironic speakers.

Table 14

Model Parameters for Adjective "Weird": Exp. 2B

	Parameters			
Factor	β	SE	t	p
Intercept	3.75	0.08	44.19	< .0001
Group	0.29	0.17	1.74	.08
Interpretation	2.04	0.17	12.03	< .0001
Speaker Gender	-0.03	0.17	-0.20	.84
Group x Interpretation	1.31	0.34	3.85	< .001
Group x Speaker Gender	-0.59	0.34	-1.74	.08
Interpretation x Speaker Gender	-0.66	0.34	-1.95	.05
Group x Interpretation x Speaker Gender	0.05	0.68	0.08	.94

Matter-of-fact

The model parameters for the adjective *matter-of-fact* are shown in Table 15. As the table shows, there was a significant main effect of interpretation: literal speakers were rated as more matter-of-fact ($M = 5.80$, $SE = 0.10$) than non-literal

speakers ($M = 3.03$, $SE = 0.14$), $F(1, 374) = 261.23$, $p < .0001$. There were no other significant main effects or interactions. Thus, the difference in *matter-of-fact* ratings between ironic and literal speakers ($M_{diff} = 2.63$) was equivalent to the difference between opposite and literal speakers ($M_{diff} = 2.91$).

Table 15

Model Parameters for Adjective "Matter-of-fact": Exp. 2B

	Parameters			
Factor	β	SE	t	p
Intercept	4.42	0.09	51.51	< .0001
Group	-0.09	0.17	-0.55	.58
Interpretation	-2.77	0.17	-16.16	< .0001
Speaker Gender	-0.08	0.17	-0.48	.63
Group x Interpretation	-0.27	0.34	-0.80	.43
Group x Speaker Gender	0.44	0.34	1.28	.20
Interpretation x Speaker Gender	-0.04	0.34	-0.12	.91
Group x Interpretation x Speaker Gender	-0.12	0.69	-0.18	.86

Summary

It is particularly important to note that there was a significant interaction between speaker group and interpretation only for the adjectives *weird* and *confusing*. Compared to literal speakers, subjects judged opposite speakers as being weirder than ironic speakers, $F(1, 379) = 14.741$, $p < .001$. In addition, subjects rated opposite speakers as more confusing than ironic speakers, compared to literal speakers, $F(1, 379) = 6.803$, $p < .01$. These findings indicate that speakers make different inferences about ironic and opposite speakers. When a speaker does not have a social pragmatic purpose for saying the opposite of what they mean, listeners find the speaker

confusing and weird (compared to ironic speakers, where there *is* a social pragmatic purpose).

Of course, it is also possible that the explicit nature of the task elicited responses that comprehenders do not necessarily generate during real-time comprehension. That is, it is possible that comprehenders need more time to draw conclusions about a speaker's social pragmatic goals, or that explicitly judging a speaker is different from drawing implicit conclusions about their linguistic tendencies. However, existing work indicates that comprehenders can make use of speaker identity information in real-time processing (Bergen & Grodner, 2012; Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Regel, Coulson, & Gunter, 2010; Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008; Yildirim, Degen, Tanenhaus, & Jaeger, 2016), including more explicit speaker identity information (Arnold et al., in press; Gibbs et al., 1991; Grodner & Bergen, 2012; Katz & Pexman, 1997; Pexman & Olineck, 2002). Together, these data make this alternative explanation less likely. Of course, it remains an empirical question that should be tested in future work.

Discussion

In Experiment 1, overall looks to the Target were lower for ironic utterances compared to literal ones, though the rate of increasing Target looks was modulated by frequency. Importantly, it is possible that rather than computing the pragmatic inferences involved in irony comprehension, the participants were instead just doing the opposite of what the ironic speaker said. The goal of Experiments 2A and 2B was

to test this by examining the time course of opposites interpretation (Exp. 2A) and comparing the inferences made about ironic vs. opposite speakers (Exp. 2B).

The participants in Experiment 2A were slower to comprehend opposite compared to literal utterances. More importantly, there was no interaction between adjective, interpretation, and time. This finding conflicts with the results from Experiment 1, where there was an adjective by interpretation by time interaction. The difference between Experiments 1 and 2A suggests that the data from Experiment 1 do not simply reflect a strategy like computing the opposite of what the ironic speaker says. Rather, the Experiment 1 eye movement data represent the time course of the pragmatic inference required for the interpretation of ironic utterances. That is, listeners make inferences about ironic speaker goals that go above and beyond truth conditions, and the generation of these inferences is influenced by the frequency of interpretation.

The goal of Experiment 2B was to determine whether listeners make different inferences about ironic and opposite speakers. Observing such a difference would suggest that the participants in Experiment 1 were making inferences about the ironic speakers that differed from those made by participants in Experiment 2A about opposite speakers. Indeed, the results from Experiment 2B indicate that listeners make different inferences about literal, ironic, and opposite speakers. Of particular note is that compared to ironic speakers, participants found opposite speakers to be weirder and more confusing. Knowing that a speaker is saying the opposite of what they mean for a *reason* (i.e., when the speaker is being ironic) shapes how the comprehender views that speaker. Participants used this knowledge as the contextual

basis for the ironic speaker's utterance, but did not do so for the opposite speakers. This indicates that comprehenders *do* consider speaker goals when processing irony.

The results described thus far indicate that comprehenders can use frequency information relatively early to aid in irony processing, but that there is still an overall delay for irony. This suggests that comprehenders might be experiencing some competition between the two possible interpretations: literal and ironic. If this is the case, we might expect that there would be some measurable conflict between these two opposing interpretations. Experiment 3 more directly tests the source of the delays for irony comprehension observed in Experiment 1 by using a paradigm that can detect conflict. Specifically, this experiment tests whether conflict in irony processing arises during the activation of the ironic and literal interpretations.

Chapter 4: Experiment 3

Overview

In Experiment 1, overall looks to the Target were higher for literal utterances compared to ironic ones. Even though irony frequency facilitated ironic criticisms compared to compliments, there was still an overall delay for irony, regardless of frequency. The present experiment more directly tests the source of the delays for irony comprehension observed in Experiment 1. Specifically, this experiment tests whether comprehending ironic utterances leads to conflict between competing representations (literal and ironic).

In syntactic processing, the simultaneous activation of two interpretations leads to conflict that consequently engages cognitive control (Hsu & Novick, 2016; Kan et al., 2013). Cognitive control supports the processing of multiple, competing representations (Botvinick, Braver, Barch, Carter, & Cohen, 2001). Prior studies have shown that cognitive control plays a role in a range of linguistic tasks, including processing syntactic ambiguities (January, Trueswell, & Thompson-Schill, 2009; Kan et al., 2013), picture naming tasks with high competition items (Novick, Kan, Trueswell, & Thompson-Schill, 2009), and verb generation and stem completion tasks (Botvinick et al., 2001).

For example, January et al. (2009) used fMRI to show within-subject overlap in the brain regions associated with both syntactic and non-syntactic conflict. That is, the regions activated when reading ambiguous sentences were also activated when participants performed a Stroop task. Novick et al. (2009) followed up on these

findings by specifically looking at a patient with damage to the left inferior frontal gyrus (LIFG). This region has been shown to be involved in conflict resolution across a variety of tasks (e.g., Milham, Banich, Claus, & Cohen, 2003; Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997; Kan & Thompson-Schill, 2004). Novick et al. (2009) found that this patient demonstrated difficulty in a high-conflict verbal fluency task as well as when comprehending temporarily ambiguous sentences. Together, these and other studies indicate that cognitive control plays a role in a range of linguistic tasks. While cognitive control plays a role in conflict processing at the lexical and syntactic levels (Botvinick et al., 2001; Hsu & Novick, 2016; January et al., 2009; Kan et al., 2013; Novick et al., 2009), there is no existing evidence as to whether it may also play a role at the level of pragmatics or semantics.

In this experiment, participants performed two tasks using the conflict adaptation paradigm. One task was the Stroop task, a task that engages cognitive control. The other task was the visual world eye-tracking task used in Experiment 1. Participants' performance on the Stroop task was analyzed as a function of the preceding visual world sentence trial. If irony processing generates conflict between the literal and ironic interpretations, then interpreting ironic utterances should engage cognitive control and facilitate subsequent performance on a subsequent high-conflict (incongruent) Stroop trial. This would be similar to syntactic ambiguity resolution, where conflict arises due to the simultaneous activation of the possible interpretations (Hsu & Novick, 2016; Kan et al., 2013). In contrast, if processing irony does not induce conflict, then participants should perform equally well on incongruent Stroop trials preceded by ironic and literal utterances.

Method

Participants

Forty-two undergraduates from the University of Maryland participated in this experiment for either pay ($5) or course credit. Of those, ten participants were removed from analysis for getting more than 20% of sentence trials incorrect ($N = 5$), not inputting a response for more than 20% of Stroop trials ($N = 4$), or for having participated in a prior irony comprehension experiment ($N = 1$). Thus, there were a total of 32 participants included in the analyses (18 female, mean age = 19.42, range = 18-24).

Materials and procedure

As in Experiment 1, participants were seated in front of a computer and their eye movements were recorded. At the beginning of the experiment, participants were told that they would complete two tasks: a sentence task and a Stroop task. Participants performed 192 trials of two interleaved tasks: a Stroop task ($n = 96$) and the visual world eye-tracking task with literal and ironic utterances used in Experiment 1 ($n = 96$). Before beginning the experiment, participants completed three practice sentence trials (two literal, one ironic) and 144 practice Stroop trials.

In the Stroop task, participants used a three-button mouse to indicate the ink color in which color names were printed on the computer screen. The response set consisted of three colors: blue, yellow, and green. Half of the trials were congruent ($n = 48$) and half were incongruent ($n = 48$). On congruent trials, the color names matched the ink color ("blue" in blue ink, "yellow" in yellow ink, and "green" in green ink). On incongruent trials, the color names mismatched the ink color. Only

response-ineligible color names were used on incongruent trials, so that the color names did not match any of the colors in the response set ("orange," "brown," and "red" in blue, yellow, or green ink). Using only response-ineligible color names on incongruent trials means that these trials elicit conflict primarily at the representational level rather than the response level (Milham et al., 2001). Thus, the conflict elicited by the Stroop task exists on the same level as the conflict elicited by the sentence processing task, where only representational conflict (i.e., between the ironic and literal interpretations) should be elicited.

On the sentence trials, participants were given the same instructions as those used in Experiment 1. Again, their task was to select the character that the speaker described (Target character). To test for cross-task conflict adaptation—namely, whether irony comprehension engages cognitive control procedures that facilitate listeners' subsequent ability to resolve conflict—the Stroop trials were pseudo-randomly interleaved with the language-comprehension trials. The experiment employed a 2 x 2 x 2 design, with current Stroop congruency (congruent, incongruent) and previous sentence interpretation type (literal, ironic) as within-subjects factors, and speaker gender (ironic male and literal female; ironic female and literal male) as a between-subjects factor. Each participant saw 12 instances each of four critical conditions: literal language comprehension trials with congruent Stroop trials (literal-congruent pairings), literal language comprehension trials with incongruent Stroop trials (literal-incongruent pairings), ironic language comprehension trials with congruent Stroop trials (ironic-congruent pairings), and ironic language comprehension trials with incongruent Stroop trials (ironic-

incongruent pairings). In other words, literal or ironic sentences (trial *n-1*) preceded either congruent or incongruent Stroop items (trial *n*). Thus, sentence trial type (*n*-1) manipulates the engagement status of cognitive control to observe its immediate effects on Stroop trial performance (trial *n*).

Two versions of each critical item were created by manipulating the interpretation type (ironic or literal). This generated two presentation lists, such that each list contained 24 items in each condition (literal, ironic), and each item appeared once in each list. See Appendix D for a list of all critical items. Two additional lists were generated by manipulating the speaker gender. For half of the participants, the male speaker was ironic and the female speaker was literal. For the other half, the female speaker was ironic and the male speaker was literal. As in Experiment 1, the speakers who pre-recorded the literal and ironic statements were instructed to use an enthusiastic tone of voice, which was felicitous with an ironic interpretation, but did not preclude a literal one. Therefore, all participants heard the same recordings.

A total of 48 filler visual world trials and 48 filler Stroop trials were included to ensure that on any given trial, participants could not predict the identity of the upcoming task. That is, Stroop trials were preceded by other Stroop trials (n = 43) and visual world trials (n = 53) an approximately equal number of times. Similarly, visual world trials were preceded by Stroop trials (n = 53) and other visual world trials (n = 42) an approximately equal number of times. The filler visual world trials contained a negative adjective instead of a positive one, so that participants could not predict the adjective valence.

Each trial (Stroop and sentence) began with a 500ms fixation, which was then replaced with either a Stroop or sentence stimulus. Both trial types were followed by a 1,000ms inter-trial interval. On Stroop trials, the word remained on the screen until the participant made a response or 1,000ms had passed, whichever occurred first. On sentence trials, the character images remained on the screen until the participant made a response. Participants were only able to make a response after the critical utterance was completed (the average utterance length was 2,671.9ms). The timing used in this experiment was the same as the timing used by Hsu and Novick (2016) and Kan et al. (2013). We know that cognitive control effects operate across linguistic and non-linguistic tasks within this time frame, so any null findings cannot be explained by timing alone.

Analysis

All data were analyzed in R (version 3.3.2; R Core Team, 2016). For Stroop reaction times, a linear mixed effects model was constructed using the *lme4* package (version 1.1-12; Bates et al., 2015). The *lmerTest* package was used to compute p-values using Satterthwaite's approximation for denominator degrees of freedom (version 2.0-32; Kuznetsova et al., 2015). The model included prior sentence type (ironic, literal), current Stroop type (congruent, incongruent), and their interaction as fixed effects. Current Stroop type and prior sentence type were both deviation coded. For all analyses, I first fit maximal models that included both random slopes and intercepts for subjects (Barr et al., 2013). However, when the maximal models did not converge, I used simpler models that only included subject as random intercept.

Prior to analysis, all trials with reaction times greater than 2 SD from the overall mean were removed; this comprised 1.7% of the Stroop trials. In addition, all incorrect trials were removed from analysis; this corresponded to 6.05% of critical trials. Because the reaction times were not normally distributed, they were log-transformed prior to analysis.

Accuracy on the Stroop task was analyzed with a mixed effects logistic regression model using the *lme4* package (version 1.1-12; Bates et al., 2015) and the *lmerTest* package to compute p-values (version 2.0-32; Kuznetsova et al., 2015). Again, participant was included as a random intercept and prior sentence type and current Stroop type, as well as their interaction, were included as fixed effects. Current Stroop type and prior sentence type were both deviation coded.

The visual-world eye-tracking data were pre-processed in the same way as in Experiment 1. Eye movements were divided into three time regions of interest: pre-adjective ("What a"), adjective-noun ("fabulous chef"), and pronoun ("he is"). For each region, looks prior to 200ms were removed to account for the time it takes to launch a saccade (Allopenna et al., 1998; Matin et al., 1993). In addition, all incorrect trials (i.e., where the participant did not click on the Target) were removed from analysis (4.13% of trials). The two characters were coded as Target (who the speaker described) and Distractor (the other character on the screen). The primary dependent measure consisted of the proportion of looks to the Target, calculated as Target looks divided by Target plus Distractor looks. The proportion of looks to the Target was averaged across 50ms windows.

Target looks were analyzed in R (version 3.3.2; R Core Team, 2016). For each region, a paired t-test was performed comparing Target looks on literal trials to ironic trials.

Results

Stroop task

All model parameters for reaction time are included in Table 16**Error! Reference source not found.**. As shown in *Figure 10*, there was a significant main effect of Stroop trial type: participants were faster on congruent trials (M = 673.3 ms, SE = 81.6) compared to incongruent trials (M = 704.1 ms, SE = 68.9), $F(1, 93) = 26.59$, $p < .0001$. (Note: analyses were performed on log-transformed data, but raw RTs are reported here for clarity of interpretation.) More importantly, there was a significant interaction between current Stroop type and previous sentence type: participants were faster on incongruent Stroop trials preceded by ironic utterances (M = 695.2 ms, SE = 68.7) compared to incongruent Stroop trials preceded by literal utterances (M = 713.0 ms, SE = 69.1), $F(1, 93) = 5.29$, $p < .05$. Thus, the standard Stroop effect is smaller when Stroop trials follow ironic utterances (Figure 9 left bars, $p = .19$) compared to when Stroop trials follow literal ones (right bars, $p < .001$). It is important to note that the paired comparisons did not indicate a significant difference between incongruent trials preceded by a literal utterance (IC) and incongruent trials preceded by an ironic utterance (II; $p = .26$). However, the pattern of results—the reduced conflict effect after an ironic utterance (and faster reaction times on II vs. CI trials)—seems to indicate that comprehending an ironic utterance does facilitate performance on a subsequent incongruent Stroop trial.

Table 16

Model Parameters for Stroop Reaction Time: Exp. 3

	Parameters			
Factor	β	SE	t	p
Intercept	6.51	0.02	349.74	< .0001
Current Trial	-0.05	0.01	-5.16	< .0001
Previous Trial	0.00	0.01	0.32	.75
Current Trial x Previous Trial	-0.04	0.02	-2.30	.02

Note. Model specification: Reaction Time ~ Current Trial * Previous Trial + (1|Subject)

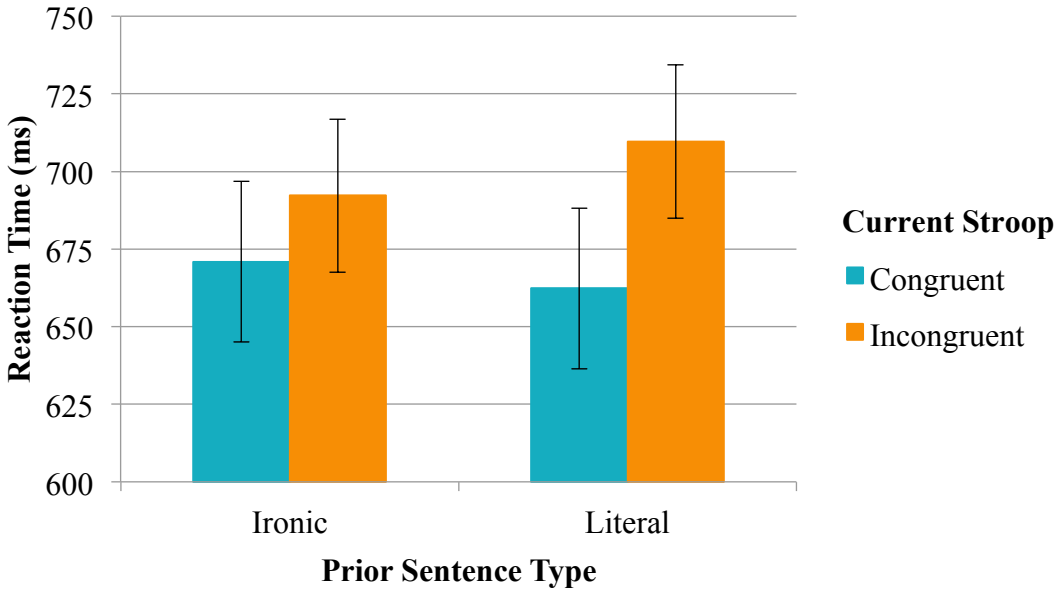

Figure 10. Reaction time by current Stroop trial type (congruent, incongruent) and prior sentence type (literal, ironic). Note: while log-transformed data were used for analysis, raw reaction times are shown here and in text for illustration purposes. Bars represent standard errors.

All model parameters for accuracy are included in Table 17. There was no main effect of current trial type (p = .90), nor was there an interaction between current and previous trial type (p = .53; see *Figure 11*). However, there was a main

effect of prior sentence type, $F(1, 1359) = 2.09$, $p < .05$. On average, participants were more accurate on Stroop trials preceded by a literal sentence trial ($M = 95.3\%$, $SE = 3.7$) compared to an ironic trial ($M = 92.6\%$, $SE = 4.6$).

Table 17

Model Parameters for Stroop Accuracy: Exp. 3

	Parameters			
Factor	B	SE	z	p
Intercept	2.77	0.12	23.80	< .0001
Current Trial	0.03	0.23	0.13	.90
Previous Trial	0.49	0.23	2.09	.04
Current Trial x Previous Trial	0.29	0.47	0.62	.53

Note. Model specification: Accuracy ~ Current Trial * Previous Trial + (1|Subject)

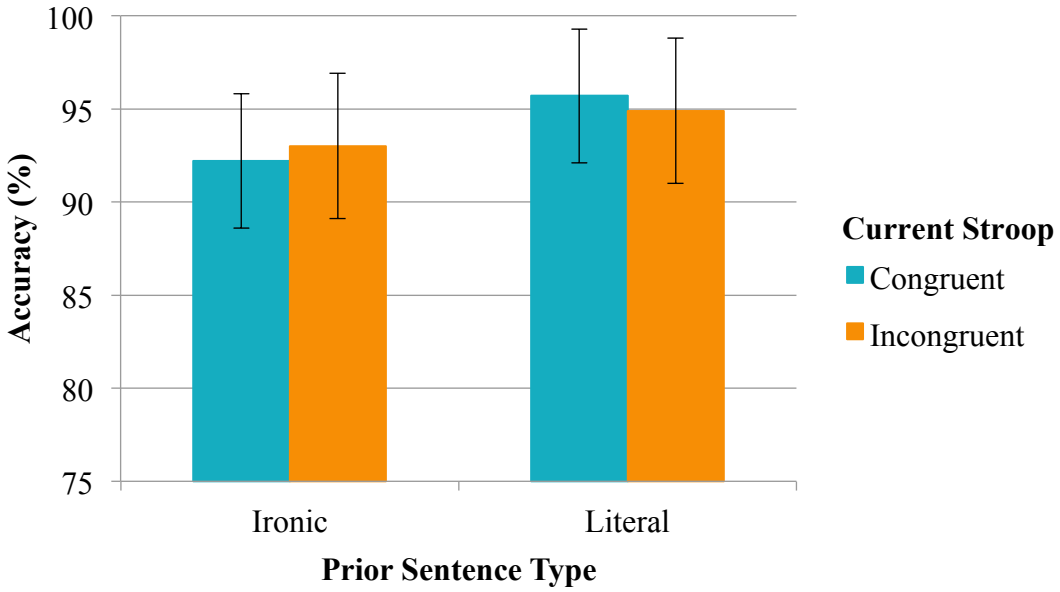

Figure 11. Accuracy by current Stroop trial type (congruent, incongruent) and prior sentence type (ironic, literal). Bars represent standard errors.

Sentence task

There was no significant difference between Target looks for literal and ironic utterances in either the pre-adjective region ($p = .15$) or the pronoun region ($p = .06$).

However, looks to the Target during the adjective-noun region were significantly greater for literal utterances (M = 0.62, SE = 0.09) than for ironic utterances (M = 0.58, SE = 0.09), $t(31) = -3.49$, $p < .01$. This is consistent with the findings from Experiment 1. In addition, participants were significantly more accurate on literal trials (M = 96.9%, SE = 3.0) compared to ironic trials (M = 94.3%, SE = 4.1), $t(31) = -3.10$, $p < .01$.

Discussion

The results from Experiment 1 indicated that looks to the Target were lower for ironic utterances compared to literal ones. The goal of Experiment 3 was to test the source of this difference. In particular, this experiment tested whether irony induces conflict and therefore engages cognitive control. To do so, participant performance on a Stroop task was examined as function of the prior sentence type. If comprehending irony generates competition between two potential interpretations (literal and ironic), then performance on incongruent Stroop trials should be facilitated by preceding ironic utterances (but not literal ones). This would be similar to syntactic ambiguity resolution, where conflict arises due to the simultaneous activation of the possible interpretations (Hsu & Novick, 2016; Kan et al., 2013).

The present findings corroborate those from Experiment 1: during the adjective-noun region, looks to the Target were greater for literal utterances than for ironic ones. More importantly, the results indicate that participants were faster on incongruent Stroop trials that were preceded by ironic utterances compared to literal ones. Thus, ironic utterances generated conflict that was resolved by cognitive control. The engagement of cognitive control then facilitated performance on a

subsequent high-conflict Stroop trial. These results suggest that, like syntactic ambiguity, ironic utterances leads to the activation of competing representations that must be resolved.

These findings also seem to corroborate work on irony processing using event-related brain potentials (ERPs). For example, Filik, Leuthold, Wallington, & Page (2014) had participants read familiar and unfamiliar ironic utterances while measuring their ERPs. Familiarity was determined using a pre-test, where participants rated how familiar they were with the ironic utterances (importantly, they did not distinguish between criticisms and compliments). Filik et al. found that both familiar and unfamiliar ironies elicited a P600-like effect. They argue that this positivity reflects ongoing conflict between the literal and ironic interpretations. This indicates not only that ironic utterances generated conflict, but also that the conflict was sustained long after the disambiguating word. Thus, it is not surprising that in the present experiment, the conflict elicited by interpreting an ironic utterance improved performance on a subsequent cognitive control task.

Work on syntactic ambiguity resolution has demonstrated both adaptation from Stroop trials to sentence trials (Hsu & Novick, 2016) and from sentence trials to Stroop trials (Kan et al., 2013). One might expect that irony might behave similarly. Specifically, while this experiment shows adaptation from ironic utterances to Stroop trials, we might also observe conflict adaptation from Stroop trials to sentence trials. This is tested in Experiment 4.

Chapter 5: Experiment 4

Overview

Experiment 3 provided evidence that comprehending irony facilitates the subsequent resolution of conflict on a Stroop task. That is, the Stroop effect was mitigated when the participant had just heard an ironic utterance compared to a literal one. While Experiment 3 focused on how the interpretation of irony affects cognitive control, Experiment 4 looks at how cognitive control engagement affects the process of interpreting irony. That is, Experiment 4 asks whether the engagement of cognitive control facilitates the subsequent processing of ironic utterances. This would be similar to work by Hsu and Novick (2016) showing that resolving Stroop conflict facilitates the subsequent processing of temporarily ambiguous utterances.

Hsu and Novick (2016) had participants complete interleaved Stroop trials and sentence trials. On sentence trials, participants viewed a scene and carried out verbal instructions while their eye movements were recorded. Participants either heard temporarily ambiguous instructions (as in 1 below) or unambiguous instructions (as in 2 below).

(1) Put the frog on the napkin onto the box.

(2) Put the frog that's on the napkin onto the box.

The visual scenes contained four objects. For example, the scene might contain a frog on a napkin (target), a box (correct goal), an empty napkin (incorrect goal) box, and a horse (competitor). Hsu and Novick found that when an ambiguous sentence trial followed an incongruent Stroop trial, participants made fewer action errors than if the prior trial had been congruent. In addition, the proportion of looks to the

correct goal (the box) were greater on ambiguous trials preceded by an incongruent Stroop trial compared to a congruent one. Thus, comprehenders resolved the conflict in the ambiguous sentence trials more quickly and accurately when cognitive control was already engaged by the preceding high-conflict Stroop trial.

In this experiment, Stroop trials and visual world eye-tracking trials were interleaved as in Experiment 3 (and Hsu & Novick, 2016). However, rather than looking at performance on Stroop trials, the present experiment looked at eye movement patterns as a function of the prior Stroop trial type. If interpreting irony is like resolving temporary syntactic ambiguity, then ironic utterances should be processed more quickly when preceded by incongruent Stroop trials compared to congruent trials.

Method

Participants

Forty-five undergraduates from the University of Maryland participated in this experiment for either pay ($5) or course credit. Of those, thirteen participants were removed from analysis for getting more than 80% of sentence trials incorrect ($N = 7$), getting more than 80% of Stroop trials incorrect ($N = 3$), not inputting responses for more than 20% of Stroop trials ($N = 2$), or for more than 20% eye-tracking track loss ($N = 1$). Thus, a total of 32 participants were included in the analyses (25 female, mean age = 19.87, range = 18-29).

Materials and procedure

Participants performed 192 trials of two interleaved tasks: the Stroop task used in Experiment 3 (n = 96) and the visual world eye-tracking task with literal and ironic utterances used in Experiments 1 and 3 (n = 96). As in Experiment 3, participants completed three practice sentence trials (two literal, one ironic) and 144 practice Stroop trials before beginning the experiment.

On the sentence trials, participants were given the same instructions as those used in Experiment 3. Again, their task was to select the character that the speaker described (Target character). To test for cross-task conflict adaptation—namely, whether the engagement of cognitive control facilitates listeners' irony comprehension—the Stroop trials were pseudo-randomly interleaved with the language-comprehension trials. The experiment employed a 2 x 2 x 2 design, with current sentence interpretation type (literal, ironic) and Stroop congruency (congruent, incongruent) as within-subjects factors, and speaker gender (ironic male and literal female; ironic female and literal male) as a between-subjects factor. Each participant saw 12 instances each of four critical conditions: congruent Stroop trials with literal language comprehension trials (congruent-literal pairings), incongruent Stroop trials with literal language comprehension trials (incongruent-literal pairings), congruent Stroop trials with ironic language comprehension trials (congruent-ironic pairings), and incongruent Stroop trials with ironic language comprehension trials (incongruent-ironic pairings). In other words, congruent or incongruent Stroop items (trial *n-1*) preceded either literal or ironic sentence items (trial *n*). Thus, Stroop trial

type (*n*-1) affects the engagement status of cognitive control to observe its immediate effects on sentence trial performance (trial *n*).

Again, two versions of each critical item were created by manipulating the interpretation type (ironic or literal). This generated two presentation lists, such that each list contained 24 items in each condition (literal, ironic), and each item appeared once in each list. Two additional lists were generated by manipulating the speaker gender. For half of the participants, the male speaker was ironic and the female speaker was literal. For the other half, the female speaker was ironic and the male speaker was literal. As in Experiments 1 and 3, the speakers who pre-recorded the literal and ironic statements were instructed to use an enthusiastic tone of voice, which was felicitous with an ironic interpretation, but did not preclude a literal one. Therefore, all participants heard the same recordings.

A total of 48 filler visual world trials and 48 filler Stroop trials were included to ensure that on any given trial, participants could not predict the identity of the upcoming task. That is, in total, Stroop trials were preceded by other Stroop trials (n = 37) and visual world trials (n = 58) an approximately equal number of times. Similarly, visual world trials were preceded by Stroop trials (n = 58) and other visual world trials (n = 38) an approximately equal number of times. The filler visual world trials contained a negative adjective instead of a positive one, so that participants could not predict the adjective valence.

The timing was the same as Experiment 3. Each trial began with a 500ms fixation and each trial was followed by a 1,000ms inter-trial-interval. This is the same timing that was used by Hsu and Novick (2016) and Kan et al. (2013). It is

important to note that the time between the end of a Stroop trial and the start of a subsequent sentence trial was shorter than the time between the end of a sentence trial and the start of a subsequent Stroop trial. This is because the average utterance length (2,671.9ms) is longer than the maximum amount of time allowed for a Stroop response (1,000ms). As a result, if adaptation from Stroop to sentence trials is not observed, it cannot be attributable to a longer time from Stroop to sentence trials.

Analysis

All data were analyzed in R (version 3.3.2; R Core Team, 2016). For the visual-world eye-tracking task, analysis proceeded in the same way as in Experiment 1. For each region, looks prior to 200ms were removed to account for the time it takes to launch a saccade (Allopenna et al., 1998; Matin et al., 1993). In addition, all incorrect trials (i.e., where the subject did not click on the Target) were removed from analysis; this corresponded to 3.42% of all trials. I coded the two characters as Target (who the speaker described) and Distractor (the other character on the screen). The primary dependent measure examined the proportion of looks to the Target, which was calculated as Target looks divided by Target plus Distractor looks. The proportion of looks to the Target was averaged across 50ms windows.

For each region, a linear mixed effects model was constructed using the *lme4* package (version 1.1-12; Bates et al., 2015). The *lmerTest* package was used to compute p-values using Satterthwaite's approximation for denominator degrees of freedom (version 2.0-32; Kuznetsova et al., 2015). Each model included current sentence type (literal, ironic), previous Stroop type (congruent, incongruent), and time from adjective onset (in 50ms bins) as fixed effects. Current and previous trial type

were both deviation coded, while time bin was included as a continuous factor. For the adjective-noun region, an additional model was constructed that only included previous Stroop type (congruent, incongruent) and current sentence type (literal, ironic) to better understand the overall, time-independent effects. For all sentence trial analyses, I first fit maximal models that included both random slopes and intercepts for subjects and trials (Barr et al., 2013). However, when the maximal models did not converge, I used simpler models that only included subject and trial as random intercepts.

As will be described below, the present experiment did *not* find conflict adaptation effects from the Stroop task to the sentence task. In order to ensure that this experiment did not differ drastically from Experiment 3, I additionally analyzed the Stroop data in the same way as in Experiment 3. It is important to note that this experiment was not designed to test performance on Stroop trials as a function of the prior sentence type. The number of trials per condition was not equal; there were 21 literal to congruent trials, 27 literal to incongruent trials, 30 ironic to congruent trials, and 18 ironic to incongruent trials. However, these analyses could help shed light on why there was no conflict adaptation effect on sentence trials, as would have been expected.

For the Stroop reaction times, a linear mixed effects model was constructed using the *lme4* package (version 1.1-12; Bates et al., 2015). The *lmerTest* package was used to compute p-values using Satterthwaite's approximation for denominator degrees of freedom (version 2.0-32; Kuznetsova et al., 2015). The model included prior sentence type (ironic, literal), current Stroop type (congruent, incongruent), and

their interaction as fixed effects. Current Stroop type and prior sentence type were both deviation coded. As above, I first fit maximal models that included both random slopes and intercepts for subjects (Barr et al., 2013). However, when the maximal models did not converge, I used simpler models that only included subject as random intercept. Prior to analysis, all trials with reaction times greater than 2 SD from the overall mean were removed; this comprised 4.0% of the Stroop trials. In addition, because the reaction times were not normally distributed, they were log-transformed prior to analysis.

Results

Sentence task

Figure 12 plots the average looks to the Target, collapsing across time, during each region of interest by condition (magnitude analyses). *Figure 13* plots the average looks to the Target during each region of interest by condition and time (rate analyses). As *Figure 12* shows, the mean proportion of looks to the Target during the pre-adjective region was 0.50 (SE = 0.08). The linear model parameters for the pre-adjective region are shown in Table 18. Unexpectedly, there were two significant interactions during this region. First, there was a significant interaction between previous trial and time, $F(1, 9{,}947.5) = 3.99$, $p < .05$. This appears to be driven by the fact that looks to the Target increased more quickly when the previous Stroop trial was congruent versus incongruent. There was additionally a significant three-way interaction between current trial, previous trial, and time, $F(1, 9{,}945.3) = 8.77$, $p < .01$. Importantly, as *Figure 13* shows, these pre-adjective effects are primarily in the CC condition (literal sentence-congruent Stroop), which is not critical to identifying

conflict adaptation. Rather, conflict adaptation effects would be seen in the II (vs. CI) condition, which does not exhibit these early effects.

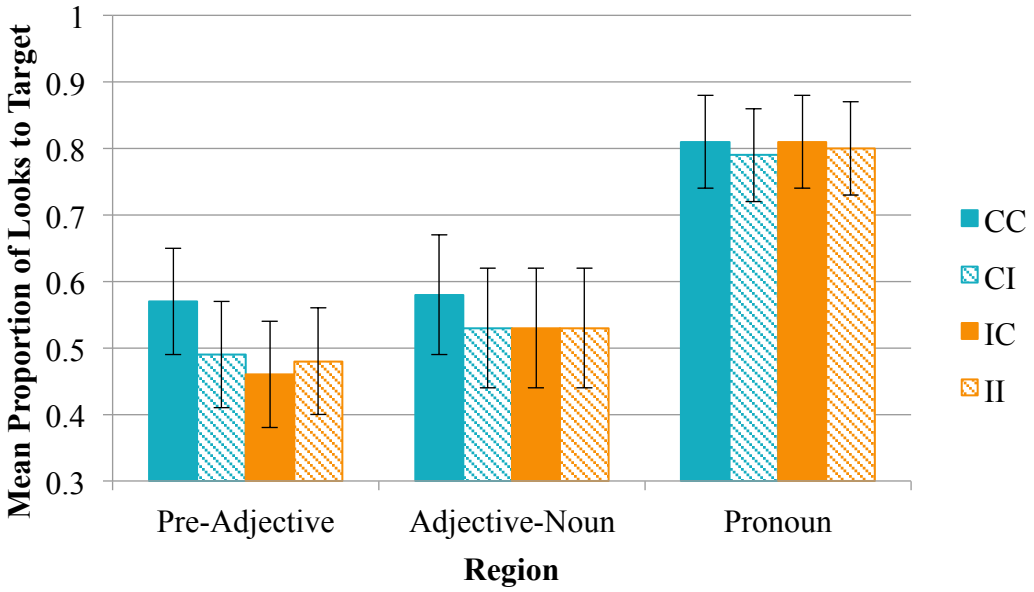

Figure 12. Mean proportion of looks to Target character in by region (pre-adjective, adjective-noun, pronoun) and condition (congruent-congruent [CC], congruent-incongruent [CI], incongruent-congruent [IC], incongruent-incongruent [II]). Bars represent standard errors.

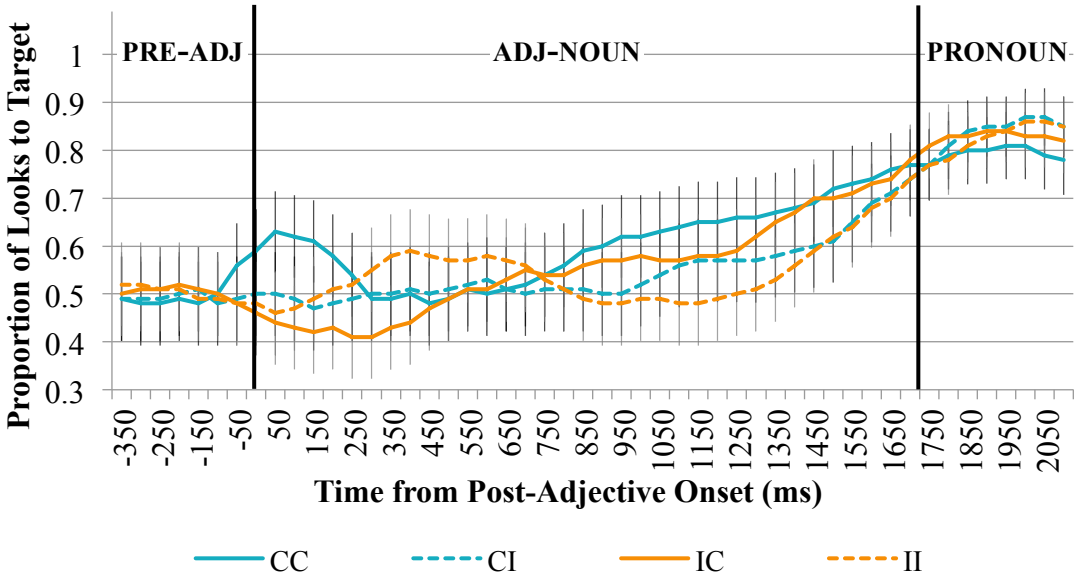

Figure 13. Average proportion of looks to the Target character in 50-ms intervals post-adjective onset by region (pre-adjective, adjective-noun, pronoun) and condition (congruent-congruent [CC], congruent-incongruent [CI], incongruent-congruent [IC], incongruent-incongruent [II]). The first vertical line represents the adjective onset and the second vertical line represents the average onset time of the pronoun. Bars represent standard errors.

Table 18

Model Parameters for Pre-Adjective Region: Exp. 4

	Parameters			
Factor	β	SE	t	p
Intercept	0.51	0.02	20.44	< .0001
Current Trial	0.02	0.05	0.45	0.65
Previous Trial	0.05	0.05	0.92	0.36
Time	-0.00	0.00	-0.86	0.39
Current Trial x Previous Trial	0.03	0.10	0.29	0.77
Current Trial x Time	0.00	0.00	0.59	0.55
Previous Trial x Time	0.01	0.00	2.00	0.05
Current Trial x Previous Trial x Time	0.03	0.01	2.97	< .01

Note. Model specification: TargetLooks ~ Current Trial * Previous Trial + (1|Subject) + (1|Item)

As described above, two linear mixed effects models were constructed for the adjective-noun region. The parameters for the magnitude model including only current and previous trial type are shown in Table 20. As *Figure 12* and Table 19 indicate, the proportion of looks to the Target was significantly higher for literal trials ($M = 0.56$, $SE = 0.09$) than ironic trials ($M = 0.53$, $SE = 0.09$). This was confirmed by a significant main effect of current trial type, $F(1, 48.07) = 4.64$, $p < .05$. This is in line with the results from Experiments 1 and 3, where the proportion of Target looks on literal trials was greater than Target looks on ironic trials. However, there was no main effect of previous Stroop trial type ($p = .16$), nor an interaction between current and previous trial type ($p = .22$).

Table 19

Model Parameters for Adjective-Noun Region without Time: Exp. 4

	Parameters			
Factor	β	SE	t	p
Intercept	0.54	0.01	67.64	< .0001
Current Trial	0.03	0.01	2.15	< .05
Previous Trial	0.02	0.01	1.42	.16
Current Trial x Previous Trial	0.04	0.3	1.25	.22

Note. Model specification: TargetLooks ~ Current Trial * Previous Trial + (1|Subject) + (1|Item)

Figure 13 plots the average looks to the Target over time during each region of interest (pre-adjective, adjective-noun, and pronoun) by current sentence type (literal, ironic) and previous Stroop type (congruent, incongruent; rate analyses). The parameters for the rate model including current trial, previous trial, and time are shown in Table 20. As Table 20 and *Figure 13* show, there was a main effect of time:

the proportion of looks to the Target increased over time across all conditions, $F(1, 36{,}580) = 180.34$, $p < .0001$. In addition, the results indicate that there was a significant interaction between current sentence trial and time, $F(1, 36{,}580) = 124.85$, $p < .0001$. Looks to the Target in the literal condition increased more quickly than in the ironic condition. Again, this is in line with the findings from Experiment 1. Similarly, there was a significant interaction between previous Stroop trial and time, $F(1, 36{,}575) = 9.61$, $p < .01$. Target looks increased more quickly when the previous trial was congruent compared to incongruent. Finally, there was a significant three-way interaction between current trial, previous trial, and time, $F(1, 36{,}569) = 27.36$, $p < .0001$. Importantly, this three-way interaction does not seem to reflect conflict adaptation. If adaptation were present, we would expect looks to the Target in the II condition (incongruent Stroop-ironic sentence) to increase more quickly than looks to the Target in the CI condition (congruent Stroop-incongruent sentence). *Figure 13* suggests that this is not the source of the interaction: the dotted orange line (II) does not increase more quickly than the dotted blue line (CI). Thus, contrary to expectations, comprehending irony is *not* facilitated by the prior engagement of cognitive control.

Table 20

Model Parameters for Adjective-Noun Region with Time: Exp. 4

	Parameters			
Factor	β	SE	t	p
Intercept	0.49	0.01	44.23	< .0001
Current Trial	-0.06	0.02	-3.73	< .001
Previous Trial	0.00	0.02	-0.30	0.77
Time	0.00	0.00	13.43	< .0001
Current Trial x Previous Trial	0.13	0.03	3.86	< .001

Current Trial x Time	0.01	0.00	11.17	< .0001
Previous Trial x Time	0.00	0.00	3.10	< .001
Current Trial x Previous Trial x Time	-0.01	0.00	-5.23	< .0001

Note. Model specification: TargetLooks ~ Current Trial * Previous Trial * Time + (1|Subject) + (1|Item)

By the pronoun region, participants had largely converged on the Target across conditions: the proportion of looks to the Target was equal to or greater than 80% in all conditions. The complete modeling results are shown in Table 21. By this region, there was no longer a three-way interaction between current trial, previous trial, and time (p = .10).

Table 21

Model Parameters for Pronoun Region: Exp. 4

	Parameters			
Factor	β	SE	t	p
Intercept	0.47	0.02	24.16	< .0001
Current Trial	0.21	0.03	7.92	< .0001
Previous Trial	0.02	0.03	0.68	0.50
Time	0.01	0.00	30.10	< .0001
Current Trial x Previous Trial	-0.08	0.05	-1.52	0.13
Current Trial x Time	-0.01	0.00	-8.77	< .0001
Previous Trial x Time	-0.00	0.00	-1.00	0.32
Current Trial x Previous Trial x Time	-0.00	0.00	1.66	0.10

Note. Model specification: TargetLooks ~ Current Trial * Previous Trial * Time + (1|Subject) + (1|Item)

Stroop task

Because adaptation from Stroop to sentence trials was not observed, an additional analysis of the Stroop trials was performed. As shown in *Figure 14*, there was a significant main effect of current Stroop type: participants were faster on

congruent trials (M = 620.2 ms, SE = 24.5) compared to incongruent trials (M = 639.6 ms, SE = 24.0), $F(1, 93) = 24.86$, $p < .0001$. (Note: analyses were performed on log-transformed data, but raw RTs are reported here for clarity of interpretation.) More importantly, there was a significant interaction between current Stroop type and previous sentence type: participants were faster on incongruent Stroop trials preceded by ironic utterances (M = 670.0, SE = 21.9) compared to incongruent Stroop trials preceded by literal utterances (M = 697.0 ms, SE = 23.0), $F(1, 93) = 19.15$, $p < .0001$. Thus, while there was no conflict adaptation from Stroop trials to sentence trials, there *was* adaptation from sentence trials to Stroop trials. This lends further support to the conclusions from Experiment 3, where conflict adaptation from sentence to Stroop trials was observed.

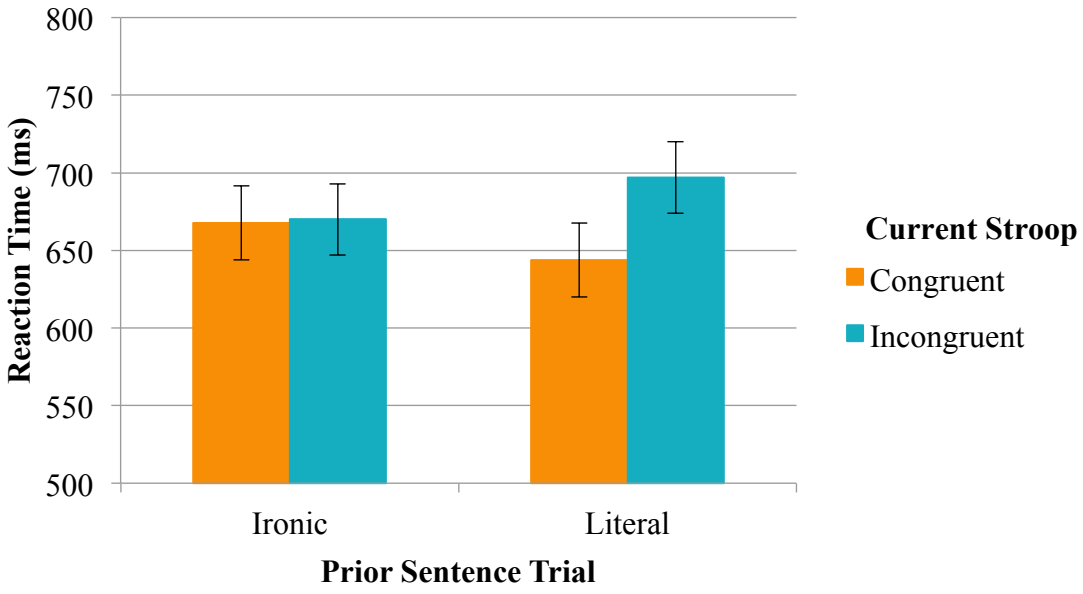

Figure 14. Reaction time by current Stroop trial type (congruent, incongruent) and prior sentence type (literal, ironic). Note: while log-transformed data were used for analysis, raw reaction times are shown here and in text for illustration purposes. Bars represent standard errors.

Discussion

The goal of Experiment 4 was to determine how cognitive control engagement affects the process of interpreting irony. In particular, this experiment addressed whether conflict adaptation occurs from Stroop trials to ironic sentence trials. Indeed, work on syntactic ambiguity has shown adaptation both from temporarily ambiguous sentence to Stroop trials (Kan et al., 2013) and from Stroop to sentence trials (Hsu & Novick, 2016). Experiment 3 demonstrated conflict adaptation from ironic sentence trials to Stroop trials. However, contrary to expectations, adaptation was not observed from Stroop trials to sentence trials in this experiment.

Of course, caution should be used when interpreting a null result. However, there are a number of possible explanations for why adaptation was not observed in this experiment. First, it is possible that comprehending irony does not induce

conflict. This is somewhat unlikely, given that adaptation was observed (in both Experiments 3 and 4) from sentence trials to Stroop trials. In addition, the timing used in this experiment was the same as that used by Hsu and Novick (2016), so it is also unlikely that the lack of adaptation is due to differences in timing. A second possibility has to do with how the setup of Experiment 4 differs from that used by Hsu and Novick (2016). In Hsu and Novick's study, participants used the mouse to act out the movement of objects on the screen. While Hsu and Novick's displays contained four objects (target, correct goal, incorrect goal, distractor), the present experiment only contained two possible referents: Fred and Sally. It is possible that the two-alternative forced choice design used here leads to different looking patterns that are not affected by prior cognitive control engagement in the same way. Another possible explanation is that the increased memory demands in this experiment could have limited or masked the role of cognitive control engagement on the sentence trials. While the participants in this experiment were incorrect 3.42% of the time on sentence trials, the rate was only 1.93% in Experiment 1. Furthermore, many more participants had to be dropped for excessively low accuracy rates in Experiment 4 (n = 7) than Experiment 1 (n = 2). These lower accuracy rates probably arise from the fact that participants had to remember the Stroop task buttons as well as the speaker identities. If participants were not accessing the speaker identities quickly (due to increased memory demands), then the engagement of cognitive control may not have had an effect on their interpretations. *Figure 13* seems to corroborate this possibility, as Target looks early on (prior to 750ms) were particularly noisy.

In addition, a final possibility is that the lack of adaptation observed in this experiment could be due to theoretical, rather than methodological, causes. For example, it is possible that irony is processed sequentially, rather than simultaneously. That is, the listener might first activate the literal interpretation and then the ironic one (indeed, the overall delay for irony in Experiment 1 lends support to this idea). In this case, conflict would likely arise later during the ironic utterance, after the literal interpretation was already accessed. It is therefore possible that the resulting engagement of cognitive control would carry over to the subsequent Stroop trial (as in Experiment 3), but that adaptation would not occur in the opposite direction. That is, the time from the end of the sentence to the next Stroop trial could be shorter (approximately 1,500ms) than the time from the end of the Stroop trial to late in the next sentence (approximately 4,000ms). This would result in adaptation from sentence to Stroop, but not from Stroop to sentence. In addition, the particular task demands could play an important role as well. For the Stroop task, the goal is to inhibit one representation (the word itself) in favor of another (the color of the word ink). In contrast, some researchers have suggested that successfully comprehending irony requires activating and maintaining both interpretations, potentially for comparison, which may lead to late or ongoing conflict (Filik et al., 2014; Giora & Fein, 2007).

There are a number of ways to test the above possibilities. First, the setup could be manipulated such that four characters are presented instead of two. This would align more closely with the setup used by Hsu and Novick (2016). In addition, participants could complete more practices trials to ensure they accurately

remembered the Stroop buttons as well as the speaker identities. In this experiment, participants only completed three practice sentence trials (as they did in Experiment 1) and 18 practice Stroop trials. This is quite different from Hsu and Novick's study, where participants completed 144 practice Stroop trials and four sentence trials (with the option for more, if the participant wanted). Of course, Hsu and Novick did not manipulate speaker type, so there was less of a memory load than in this experiment. Perhaps with practice participant performance would reduce any memory-induced difficulty. Thus, more work needs to be done to better understand the role of conflict and its resolution in irony comprehension.

Chapter 6: General Discussion

Summary of findings

It is well known that speakers use irony for a broad range of pragmatic purposes, such as testing and bolstering common ground (Brown, 1995), saving face when making criticisms (Dews & Winner, 1995; Jorgensen, 1996) and increasing politeness (Attardo, 2001). However, how irony is understood in real-time is a topic of debate. The goal of this dissertation was to investigate how context and frequency influence irony comprehension by systematically testing (a) when and how comprehenders use contextual information during irony processing (Exps. 1 & 2A), (b) the inferences listeners make about ironic speakers (Exp. 2B), and (c) whether irony induces conflict between incompatible representations (i.e., literal and ironic) of utterance meaning (Exps. 3 & 4). In Experiment 1, I showed that there was no effect of frequency on the overall proportion of Target looks during the region of interest. Target looks were greater for literal utterances than ironic ones, regardless of frequency. This suggests that the interpretation of ironic utterances is delayed overall compared to literal ones. However, the speed with which comprehenders reached ironic interpretations was modulated by frequency: participants were faster to interpret more frequent ironic criticisms than less frequent ironic compliments. In Experiment 2A, I showed that the way in which comprehenders understand irony differs from how they understand opposites. In particular, the frequency of interpretation (criticisms vs. compliments) did not influence processing speed or overall interpretations for opposites. Thus, processing irony involves more than simply computing the opposite of the utterance, instead requiring listeners to draw the

pragmatic inferences made about the speaker's intentions. Similarly, the data from Experiment 2B indicate that comprehenders view the social-pragmatic goals of ironic speakers differently from opposite speakers. This is further evidence that understanding irony involves drawing conclusions about speakers and their goals in a way that understanding (or computing) opposites does not. In Experiments 3 and 4 I tested whether and how comprehending irony induces representational conflict. Experiment 3 showed that comprehenders experience conflict between the literal and ironic interpretations when interpreting irony, corroborating the findings from Experiment 1. Hearing an ironic utterance engaged cognitive control, which then facilitated performance on a subsequent high-conflict Stroop trial. This is consistent with work on syntactic ambiguity resolution, where interpreting a temporarily ambiguous sentence facilitates performance on a subsequent high-conflict task (Kan et al., 2013). Thus, cognitive control is engaged to adjudicate between the incompatible interpretations (literal and ironic), providing further evidence that listeners do not immediately access irony without somehow consulting the literal meaning. In Experiment 4, however, conflict adaptation was *not* observed from Stroop trials to sentence trials, in contrast to prior work on syntactic ambiguity resolution (Hsu & Novick, 2016; see Chapter 5 for discussion).

To briefly summarize, the overall findings from this dissertation show that the frequency of irony modulates how quickly the listener reaches the ironic interpretation: frequent ironic criticisms are faster than infrequent ironic compliments. However, even for conventional uses, ironic interpretations appear to compete with the literal meaning of the utterance. Activation, as measured by Target looks, is

lower overall for ironic interpretations compared to literal ones. To resolve this competition, the system draws on cognitive control procedures in a way that resembles how the system avoids misinterpretation at lower levels (Hsu & Novick, 2016; Kan et al., 2013).

In the remainder of this chapter I will examine four topics. First, I discuss the implications this work has for traditional models of irony processing and the kind of model that this dissertation suggests. Next, I talk about the relationship between the present findings and prior work on other phenomena at the semantics-pragmatics interface. In the following section, I touch on how this work speaks to topics related to the lexicon and social cognition. I then talk about the implications that the present findings have for more general theories of sentence processing and language comprehension more broadly. Finally, I discuss limitations to the experiments in this dissertation as well as directions for future work.

Traditional theories of irony processing

The present findings can speak to the traditional accounts of irony described in Chapter 1. To briefly review, there are three standard accounts for irony comprehension. According to the *standard pragmatic view*, comprehending irony occurs in stages (Cutler, 1976; Dews & Winner, 1999; Giora et al., 2007). The listener first accesses the literal interpretation of an ironic utterance. Then, if there is a mismatch between the literal interpretation and the context of the speaker's utterance, the listener computes the ironic interpretation. Thus, context is not used until later in processing. In contrast, the *direct access view* argues that context interacts with lexical processing early on (Gibbs, 1986; Ivanko & Pexman, 2003).

Here, if the context is sufficiently supportive, the listener can access the ironic interpretation of the utterance without needing to access the literal meaning first. Finally, the *graded salience hypothesis* distinguishes between more and less salient ironies, where salience is determined by frequency, familiarity, or conventionality (Giora, 1997; Giora et al., 1998). For salient ironies, the ironic meaning is lexicalized and therefore accessed directly. However, less salient ironies are processed in stages: literal meaning first, then ironic.

The findings in this dissertation do not seem to support any one account completely, but the graded salience hypothesis offers a relatively good fit. While the standard pragmatic view predicts irony will be delayed, it does not take irony frequency into account. Similarly, the direct access account would explain why ironic criticisms are processed at the same rate as literal utterances (the rate effect in Experiment 1). However, neither of these accounts perfectly explains the data – specifically, that irony will be delayed overall, but that the rate of interpretation will be modulated by frequency. In contrast, the graded salience hypothesis considers the role of frequency, something that the other accounts do not. However, there are still some gaps. For example, the overall delay for even the more frequent form of irony observed here seems to suggest that irony is not actually lexicalized, as the graded salience hypothesis would predict. If it were, ironic criticisms should be processed as quickly as literal utterances. It is also unclear how frequency maps onto the account's idea of "salience." For example, a study conducted to test this account used stimuli that were pre-tested for familiarity. The target phrases were presented to participants who rated them on how familiar they were as ironic statements (Filik et al., 2014).

However, frequency as defined by the criticisms/compliment distinction was not manipulated. Thus, some of the "familiar" ironies were negative ("What a shame") while others were positive ("You are so tactful"). Indeed, the lack of a clear definition of salience makes this particular theory difficult to falsify. To better evaluate this model's fit to the data presented here, it would be useful to directly manipulate and test the effects of frequency and familiarity. This would be a productive direction for future work.

A model that accommodates my results might look as follows. When an ironic utterance is produced, the literal interpretation is accessed immediately. The listener then consults the context to determine whether the utterance is appropriate. (It is also possible that these processes happen in tandem, I stake no claim here about sequential vs. simultaneous access. However, see the below section on sentence processing for a discussion of how the present experiments might speak to this issue.) This context may include situational context (the current situation and prior events), within-speaker information (whether a speaker has a tendency to be ironic), or across-speaker information (the greater frequency of ironic criticisms vs. compliments). Given that the literal interpretation of the utterance is *not* appropriate in the context, the listener generates the ironic interpretation. Importantly, as the magnitude effect in Experiment 1 shows, there will be an overall delay for irony. In Experiment 1, providing strong context (in the form of speaker identity information) for irony did not eliminate delays, even for the more frequent criticisms. This indicates that unlike homophones, irony is not stored in the mental lexicon. If it were, providing context for irony would eliminate the delay (particularly for the more frequent criticisms),

similar to how context for equibiased homophones leads to activation for only the relevant meaning (Duffy et al., 1998). Thus, regardless of whether the ironic utterance is a criticism or compliment, there will be an overall delay. However, if the utterance is a criticism, it will be processed relatively quickly given the increased frequency of prior experience (i.e., the across-speaker context). Because the cue to ironic criticisms is stronger (due to greater frequency across speakers), the increase in activation for the ironic interpretation rises at the same rate as the literal interpretation (i.e., the rate effect observed in Experiment 1). If the ironic utterance is a compliment, its relatively infrequent use makes it a weaker cue, therefore generating an additional processing delay. Regardless of frequency, the listener must then either revise or inhibit their initial literal interpretation using cognitive control. Thus, the (sequential or simultaneous) activation of the two interpretations—literal and ironic—generates conflict, thereby leading to an overall delay for irony interpretation.

The semantics-pragmatics interface

This dissertation addresses the semantics-pragmatics interface and how these two levels of representation relate to one another. The findings discussed here suggest that for irony, the activation of the literal and ironic interpretations generates conflict. For ironic compliments, this conflict manifests as delays both in the magnitude of Target looks as well as the growth rate of Target looks over time. For ironic criticisms, the conflict only manifests itself in the rate of Target looks, not magnitude.

Similar results have been obtained for other pragmatic phenomena like scalar implicatures. Scalar terms like *some* have two possible interpretations: a semantic

meaning ("some and possibly all") and a pragmatic meaning ("some but not all"). For example, the semantic analysis of utterance (1) below would be that Fred ate some and possibly all of the cake. Indeed, (2) would be felicitous with this interpretation. However, the pragmatic interpretation would be that Fred ate some, but not all, of the cake. Here, (3) would be a felicitous follow-up.

(1) Fred ate some of the cake.

(2) In fact, Fred ate all of the cake.

(3) But Fred did not eat all of the cake.

Scalar implicatures share some properties with irony: namely, the generation of a semantic analysis and a pragmatic inference. As with irony, a great deal of work on scalar implicatures has been devoted to understanding when and how listeners reach these two interpretations. That is, given sufficient context, can a listener reach the pragmatic interpretation without first going through the semantic analysis? Or must semantic analysis be completed prior to pragmatic inferencing?

Some work supports the latter alternative: the semantic analysis must be processed prior to the pragmatic one (Bott & Noveck, 2004; Huang & Snedeker, 2009; Huang & Snedeker, 2011; Tomlinson et al., 2013). For example, Huang and Snedeker (2009) presented participants with utterances like "Point to the girl that has some of the socks," while tracking their eye movements. The display contained a girl with a subset of an item (e.g., two of four total socks) and another girl that had all of an item (e.g., four of four total soccer balls). Thus, there was a brief period of ambiguity, starting from *some* and ending before the end of the noun (*-ks*). While the semantic meaning of *some* was compatible with either girl, the pragmatic meaning

was only compatible with the girl with the socks. Huang and Snedeker found that participants were slower to look at the girl with the socks compared to trials with unambiguous terms like *all*, *two*, and *three*. The delay for the pragmatic meaning of *some* indicates that listeners perform the semantic analysis prior to generating the pragmatic inference. This is compatible with the current findings from Experiment 1.

However, other work suggests that the pragmatic meaning of *some* can be accessed immediately. For example, Grodner, Klein, Carbary, and Tanenhaus (2010) used a visual world eye-tracking paradigm to determine when listeners reach the pragmatic interpretation of *some*. Participants heard utterances containing *some* (e.g., "Click on the girl who has summa the balloons") and had to select the character (Target) on the screen corresponding to the utterance. (Note that Grodner and colleagues used "summa" instead of "some" to provide an earlier phonetic signal to make the timing more comparable to literal controls.) The display contained one girl with a subset of an item (e.g., two of four balloons) and another girl that had all of an item (e.g., four of four balls). Other trials included non-scalar quantifiers like *alla* and *nunna*. Grodner et al. found that participants looked to the Target as quickly for *summa* as they did for *nunna* and *alla*, indicating that accessing the pragmatic interpretation of *some* does not induce processing delays. This suggests that listeners can reach the pragmatic interpretation of *some* without having to first access the semantic one.

This dissertation adds new evidence to the literature on the semantics-pragmatics interface. Similar to Huang and Snedeker's (2009) findings, looks to the Target were greater overall for literal utterances than ironic ones. Thus, at least in

magnitude, it appears that pragmatic inferencing is delayed compared to semantic analysis. Whether or not this indicates sequential or simultaneous access of the literal and ironic interpretations is a topic for future work (but see the below section on sentence processing for some conjectures). Work by Giora and Fein (2007) can inform this debate to some degree, however. As described in Chapter 1, Giora and Fein had participants complete a lexical decision task after reading ironic utterances in irony- or literal-biasing contexts. They found that less familiar ironies were interpreted literally initially (150ms post-offset) and ironically later (1,000ms post-offset). This is consistent with the present findings, where less frequent ironic compliments are processed more slowly than their literal counterparts. This is also consistent with some of the work on scalar implicatures, where semantic analysis occurs prior to pragmatic inferencing (Bott & Noveck, 2004; Huang & Snedeker, 2009; Huang & Snedeker, 2011; Tomlinson et al., 2013). However, Giora and Fein also showed that more familiar ironies were interpreted both literally and ironically initially. This could explain the difference in Target look magnitude for ironic criticisms and their literal counterparts that was observed here. The simultaneous access of the literal and ironic interpretations would generate conflict, which would then need to be resolved using cognitive control. This, in turn, would aid in resolving the subsequent Stroop trial conflict.

The results described here seem to correspond best to an account where the interpretation of scalar implicatures is delayed compared to literal controls. In this dissertation, the interpretation of irony is similarly delayed. However, because there are different forms of irony that vary in frequency, a more nuanced account must be

offered. Such an account must explain both the overall delay in irony, as well as the difference in interpretation rate attributable to irony frequency (see the above section on traditional theories of irony processing).

Social cognition & the lexicon

The finding that both forms of irony are more difficult to comprehend than their literal counterparts is rather surprising given the abundance of psycholinguistic research showing early effects of context on interpretation (Altmann & Steedman, 1988; Altmann, Garnham, & Dennis, 1992; Tanenhaus et al.,1995; Trueswell & Tanenhaus, 1991). In addition, work using a fast priming paradigm suggests that detailed argument information is accessible very early in comprehension for verbs (Trueswell & Kim, 1998) and nouns (Novick, Kim, & Trueswell, 2003). However, even when comprehenders are presented with a frequent form of irony (represented as across-speaker context), and given a perfectly reliable cue to irony (within-speaker context – i.e., speaker identity), there is an overall delay in processing compared to literal utterances. This is unlikely to be due to the explicit nature of the speaker identity information, as other work has indicated that comprehenders can use explicitly-provided information about speakers to aid in real-time comprehension (Arnold et al., 2007; Arnold et al., in press; Gibbs et al., 1991; Grodner & Bergen, 2012; Katz & Pexman, 1997; Pexman & Olineck, 2002). In addition, telling the listeners different information about the speaker (e.g., ironic in Exp. 1, opposite in Exp. 2A) alters their eye-movement patterns. Furthermore, as evidenced by the findings in Experiment 3, comprehending frequent ironic criticisms still generates conflict. These findings seem to suggest that, unlike homophones, irony may not be

stored pre-compiled in the lexicon. If it were, providing a perfectly reliable cue to interpretation (in the form of speaker identity) should lead to interpretation speeds equivalent to literal meanings. For example, upon hearing "fabulous," the listener immediately activates its phonological, semantic, and (perhaps) argument structure properties. If irony were lexicalized, this rapid recognition process would also include the ironic interpretation of "fabulous" (e.g., "terrible"). However, because this dissertation demonstrates that irony is delayed, it suggests that it is not stored in the lexicon.

In addition, it is unlikely that the delay for irony is due to a general delay in integrating or consulting speaker identity information. In fact, there is a great deal of work indicating that comprehenders track and use information about speaker identity and goals to guide interpretations (Bergen & Grodner, 2012; Brown-Schmidt et al., 2008; Regel et al., 2010; Van Berkum et al., 2008; Yildirim et al., 2016). For example, Van Berkum et al. (2008) had participants listen to utterances whose content did not match inferences about the identity of the speaker (e.g., "If I only looked like Britney Spears" in a male voice). They found that these mismatched utterances generated an ERP as early as 200-300ms after the critical word onset. Thus, listeners rapidly make use of speaker identity information when comprehending speech.

It is also improbable that the delay for irony is due to prosody effects. As described above, prosody was not manipulated in any of the present experiments. However, there is work indicating that there is no particular ironic tone of voice (Attardo et al., 2003; Bryant & Fox Tree, 2005; Kreuz & Roberts, 1995; Rockwell, 2000) and that listeners do not rely on specific vocal cues to identify verbal irony

(Bryant & Fox Tree, 2005). Thus, even though the ironic and literal speakers did not differ in prosody, it should not account for the irony delay.

While the overall delay for irony may not be attributable to the slow access to speaker identity information, it could be due to the necessity of performing semantic analysis in addition to pragmatic inferencing. Whether or not these processes happen sequentially or simultaneously, it seems that both must necessarily take place. This also explains the fact that even the more conventional uses of irony generate conflict that must be resolved (Exp. 3). The literal and ironic interpretations of the utterance compete, leading to delays and the engagement of cognitive control.

This raises the question: if irony is not stored in the lexicon, where is information about irony frequency maintained? One possibility is that this information is simply represented as communicative inferences. Indeed, we know that listeners track and use a range of contextual information to interpret irony. First, they track speaker identity and linguistic tendencies. The fact that listeners' eye movements are different for ironic (Exp. 1) and opposite (Exp. 2A) speakers lends credence to this. In addition, listeners use information about speakers' language use to draw conclusions about the speakers themselves. Indeed, how listeners view a speaker is influenced by whether the speaker is ironic or uses opposites frequently (Exp. 2B). Finally, listeners track information about irony frequency. While this information may not be used immediately during irony comprehension (i.e., Exp. 1 magnitude effect), it is used eventually (Exp. 1 rate effect). Thus, listeners observe individual instances of each type of irony (criticisms and compliments), which must ultimately accumulate as probabilistic, speaker-independent information. Thus,

listeners have knowledge indicating the ironic criticisms are more frequent than compliments. When an ironic utterance is encountered, the listener brings to bear these various sources of information to interpret the utterance. This is consistent with constraint-satisfaction accounts of language comprehension in which multiple sources of information—linguistic and non-linguistic—are integrated simultaneously in real-time (Jurafsky, 1996; MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey-Knowlton, 1995).

Of course, how these communicative inferences are stored in memory is still a topic of debate (Brown-Schmidt & Duff, 2016). For example, early work by Clark and Marshall (1978) argued that interlocutors store diary-like memories including information about events and their participants. These memory structures are special-purpose and specific to individuals. More recently, researchers have argued that more general forms of memory support representations about interlocutors and their shared experiences (Brown-Schmidt & Duff, 2016; Horton, 2007; Horton & Gerrig, 2016). This latter approach seems best suited to account for the present findings. In particular, Brown-Schmidt and Duff (2016) argue that episodic memory may be a good candidate to store this kind of social communicative information. This memory system can track and integrate information across speakers, time, and circumstances. Indeed, patients with hippocampal damage (and consequently impairments to episodic memory) have difficulty adjusting their referring expressions for different speakers (Duff, Gupta, Hengst, Tranel, & Cohen, 2011). For example, rather than referring to a game as "the game" after being previously discussed, amnesic patients often continued using indefinite references with their interlocutors.

Thus, probabilities relating to different speaker goals and attitudes (and as a result, the frequency of different forms of irony) are tracked via episodic memory over the course of an individual's lifetime, and can be used as an additional cue in irony comprehension. The use of these probabilistic cues would be similar to other cases in which existing knowledge about speakers' goals and attitudes are integrated in real time (Katz & Pexman, 1997; Pexman & Olineck, 2002). For example, readers integrate information about a speaker's occupation immediately, alongside other types of relevant information (e.g., lexical, syntactic; Pexman, Ferretti, & Katz, 2000). Similarly, listeners integrate stereotype information about a speaker's gender, age, and socio-economic status as early 200-300ms after the onset of a critical word (Van Berkum et al., 2008). Thus, comprehenders can rapidly make use of social information—tracked and encoded over the course of a lifetime—to aid in real-time processing. I propose that listeners do the same for irony frequency information.

Implications for language comprehension

This dissertation may also speak to broader issues regarding language comprehension, word recognition, and sentence processing. As described in Chapter 1, much of the work on word recognition has focused on how context and frequency interact in real-time. For example, Swinney (1979) provided evidence for a two-stage model of word recognition. He showed that for homophones, both possible meanings are accessed initially, regardless of context. Then, a post-lexical access process integrates context and selects the appropriate meaning. Duffy et al. (1988) expanded on this work by considering the frequency of the homophone's two meanings. They showed that for non-equibiased homophones, context supporting the subordinate

meaning led to competition between the two possible meanings. This is similar to the findings observed here for ironic compliments: when context supported the less frequent ironic meaning, it led to delayed interpretation. However, Duffy et al. showed that for equibiased homophones, only the relevant meaning was activated initially. The findings in this dissertation seem to diverge here from the word recognition literature. When context was provided in support of ironic criticisms, there was still a delay in interpretation (Exp. 1). This was the case even when the context (i.e., speaker identity) perfectly predicted utterance interpretation. Furthermore, Experiment 3 revealed that even for this frequent form of irony, comprehenders experienced conflict during processing.

The use of context in real-time language has also been explored extensively with regard to syntactic processing. For example, Trueswell, Sekerina, Hill, and Logrip (1999) presented participants with sentences like (4):

(4) Put the frog on the napkin into the box.

Sentence (4) is temporarily ambiguous, because "on the napkin" could refer to the Destination of the putting event, or to a Modifier indicating which frog to move. These sentences were presented with one of two visual scenes. One scene supported the Destination interpretation (1-Referent context) and contained a frog on a napkin (target), a box (correct goal), an empty napkin (incorrect goal) box, and a horse (competitor). In the 1-Referent context, the modifier "on the napkin" would be unnecessary since there is only one frog. Thus, participants should initially interpret "on the napkin" as the destination. The other scene supported the Modifier interpretation (2-Referent context) and replaced the competitor object with another

frog that was not on a napkin. Because the comprehender in the 2-Referent context would need to know which frog to move, they should immediately interpret "on the napkin" as a modifier, rather than a destination. These conditions were also compared to conditions where the sentence was unambiguous, as in (5) below.

(5) Put the frog that's on the napkin into the box.

Tanenhaus et al. (1999) found that looks to the incorrect destination (the empty napkin) only increased in the 1-Referent ambiguous condition. Thus, in the 2-Referent context, the participants were able to rapidly use contextual information (the presence of two frogs) to determine that "on the napkin" was a modifier, not the destination. Similarly, participants made the most errors in the 1-Referent ambiguous condition. These data indicate that adults can make use of contextual information very early in syntactic processing. This seems to conflict with the present findings, where participants did *not* use context early. That is, even when provided with a perfectly reliable cue to irony (speaker identity), there were still fewer overall looks to the Target in the ironic conditions. This is in spite of the fact that comprehenders can make rapid use of explicitly-provided information about speaker tendencies during real-time processing (Arnold et al., in press; Grodner & Bergen, 2012; Katz & Pexman, 1997; Pexman & Olineck, 2002). Thus, as discussed above, it seems that comprehending irony necessarily involves accessing both the literal and ironic meanings, which then generates conflict and interpretation delays.

More recent work on syntactic parsing has focused on the cognitive mechanisms that enable listeners to resolve syntactic ambiguities. In particular, this work suggests that comprehending temporarily ambiguous sentences such as (4)

above and (6) below engages cognitive control (Hsu & Novick, 2016; January et al., 2009; Kan et al., 2013).

(6) The basketball player accepted the contract would have to be negotiated.

This is because during a garden path sentence, the two possible interpretations compete. In example (6) above, these two interpretations would be (a) the player agreed to a new contract and (b) the player acknowledged the need to negotiate the contract. While a listener would temporarily consider (a) to be the appropriate interpretation, the word "would" indicates that (b) is in fact the correct interpretation.

The activation of and competition between the two possible interpretations is supported by work using the conflict adaptation paradigm. This research indicates that completing a high-conflict cognitive control task prior to reading a garden path sentence leads to faster garden path recovery (Hsu & Novick, 2016). Furthermore, reading a garden path sentence facilitates performance on a subsequent cognitive control task (Kan et al., 2013). This bidirectional effect indicates that (a) processing a garden path sentence mitigates the Stroop effect on the next trial via sustained engagement of cognitive control, and (b) engaging cognitive control on a prior trial facilitates the subsequent revision of a temporarily ambiguous sentence.

This work raises an interesting possible explanation for the lack of adaptation observed in Experiment 4. It is possible that observing conflict adaptation in only one direction, as is the case here, could indicate that the literal interpretation is accessed prior to the ironic one. As described in Chapter 5, the listener might first activate the literal interpretation and then the ironic one. As a result, conflict would arise later during the ironic utterance, after both interpretations were already accessed. Thus,

the resulting engagement of cognitive control would carry over to the subsequent Stroop trial, but that adaptation would not occur in the opposite direction. This would result in adaptation from sentence to Stroop (Experiment 3), but not from Stroop to sentence (Experiment 4). It is important to note that the experiments in the present dissertation were not designed to test this specifically. Furthermore, there is still a lot that is unknown about the refractory period for cognitive control engagement, and it is hard to interpret null results. Nonetheless, it is an interesting possibility that could generate a new interpretation of the results in Experiments 3 and 4 of this dissertation.

As discussed in Chapter 5, the lack of adaptation observed in Experiment 4 may also be influenced by the setup of the experiment itself. There are a number of possible experimental causes for this null result, such as the use of only two possible referents in the visual scene and the increased memory demands compared to Hsu and Novick's (2016) study. In addition, Experiments 3 and 4 only include positive ironies, which are the most frequent types. It is possible that if these experiments included negative ironies, the conflict between the literal and ironic interpretations would be mitigated by prior cognitive control engagement. Because ironic compliments are more delayed than ironic criticisms (compared to their literal counterparts), ironic compliments might generate more conflict. Thus, effects of cognitive control engagement on interpretation might be more clearly detectable. Of course, this is an empirical question that should be tested in future research.

Limitations and future work

The experiments in this dissertation have a few limitations that are important to consider when interpreting the findings and planning future work. First, it is

possible that the observed delay for irony may be due to insufficiently strong context. Perhaps more implicit cues, such as prosody, would speed up ironic interpretations above and beyond the present, more explicit cue (speaker identity). In addition, it might be useful to give participants more time during the critical utterances to identify the speaker and their linguistic tendencies before the adjective is produced. Indeed, some work suggests that early in an utterance, speaker goal information might be inaccessible or too resource-intensive to use (Lin, Keysar, & Epley, 2010).

There are also several aspects of the conflict adaptation experiments that should be further investigated. First, it is possible that the conflict adaptation results in Experiment 3 could be driven by task difficulty. Specifically, it could be the case that comprehending irony is a difficult task (compared to comprehending literal utterances), which consequently leads to increased attention. This increased attention could then facilitate performance on a subsequent difficult task. Given that conflict adaptation was not observed from the Stroop task to the sentence task, this seems unlikely. However, a potential follow-up experiment could replace the Stroop task with a difficult, but non-conflict task and observe performance on that task as a function of the prior sentence trial type. If the present effects are the result of competition per se, then performance on the difficult, non-conflict task should not be modulated by the prior sentence type (literal or ironic).

There are also several modifications that could be made to Experiment 4 to test for conflict adaptation from the Stroop task to the sentence task. First, as described in Chapter 5, participants could be given additional practice trials. This would rule out the possibility that the lack of adaptation effects was driven by high

memory demands. If participants were having difficulty remembering the speaker identities, it could mask any adaptation effects. This possibility is slightly weakened by the fact that Experiment 3 did show adaptation, but the effect on sentence comprehension may simply be harder to detect (or, as described above, the timing of the conflict may play a role). In addition, it would be useful to run Experiments 3 and 4 with negative adjectives. This would reveal the extent to which conflict is experienced for the less frequent form of irony.

Conclusion and closing remarks

To conclude, the goal of this dissertation was to better understand how context and frequency interact during the real-time comprehension of ironic utterances. Across four experiments, several key findings have emerged. First, the overall interpretation of irony is delayed relative to literal utterances. In Experiment 1, the overall magnitude of Target looks was greater for literal utterances than ironic ones. This was the case regardless of irony frequency. Second, the speed of irony interpretation is modulated by frequency. The rate of increase in Target looks was greater for frequent ironic criticisms than infrequent ironic compliments. Third, the conclusions that listeners draw about ironic speakers differ from those made about "opposite" speakers. Opposite speakers were judged as being weirder and more confusing compared to ironic ones, and comprehenders processed opposite utterances differently from ironic ones. Finally, comprehending ironic utterances generates conflict, even for the more frequent form of irony. Compared to interpreting a literal utterance, interpreting an ironic utterance led to improved performance on a subsequent high-conflict Stroop trial.

Taken together, these results suggest a model in which the literal interpretation of an ironic utterance is activated before (or simultaneously with) the ironic interpretation. Upon hearing the utterance, the listener consults the context and determines that the literal interpretation is inappropriate. Thus, the listener generates a pragmatic inference to reach the ironic interpretation. This will be faster if the utterance is positive (ironic criticism) compared to if it is negative (ironic compliment), because frequency determines the strength of the cue. Finally, the listener engages cognitive control to inhibit or revise the literal interpretation and reach the final ironic interpretation.

Appendices

Appendix A

Instructions for Participants in Experiment 1

In this experiment, you're going to see two characters and listen while a narrator describes a scene involving the characters. Then you will hear a male or female speaker comment on one of the character's actions.

Lists 1-4 (literal female, ironic male)

The female speaker, **Literal Lucy**, always means what she says. Every time she says something, she always means it literally.

For example, if it is raining outside, she might say, "Wow, what a terrible day!"

The male 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!"

Lists 5-8 (literal male, ironic female)

*The male speaker, **Literal Luke**, always means what he says. Every time he says something, he always means it literally.*

For example, if it is raining outside, he might say, "Wow, what a terrible day!"

*The female speaker, **Ironic Irene**, has a quirky sense of humor. Every time she says something, she always means it ironically.*

For example, if it is raining outside, she might say, "Wow, what a beautiful day!"

All Lists
Your task is to use the mouse to select the character that the speaker describes.
(Show them mouse.)

Appendix B

Critical Items Used in Experiments 1 and 2A

Item	Vignette	Critical Utterance
1	Fred and Sally did their holiday shopping. Fred/Sally when shopping early and got all of the gifts he/she needed. Fred/Sally went shopping at the last minute and only got half of the gifts he/she needed.	What a fantastic/ horrible planner he is.
2	Fred and Sally went golfing. Fred/Sally got a hole in one. Fred/Sally hit the ball into a neighbor's window.	What a remarkable/ hopeless golfer she is.
3	Fred and Sally competed in a swim meet. Fred/Sally finished in first place. Fred/Sally finished in last place.	What an incredible/ abominable swimmer he is.
4	Fred and Sally went out to play darts with their friends. Fred/Sally got a bullseye on his/her first try. Fred/Sally missed the dartboard completely.	What an impressive/ atrocious thrower she is.
5	Fred and Sally drove home at night while no one else was on the road. Fred/Sally obeyed the speed limit the entire way home. Fred/Sally sped the entire way home.	What an excellent / abominable driver she is.
6	Fred and Sally gave a speech at a local school. At Fred's/Sally's speech, the audience hung on his/her every word. At Fred's/Sally's speech, the audience fell asleep.	What a magnificent/ incompetent speaker he is.
7	Fred and Sally babysat for their neighbor's baby. When Fred/Sally was babysitting and the baby cried, he/she soothed him until he fell asleep. When Fred/Sally was babysitting and the baby cried, he/she yelled and made the baby cry more.	What a wonderful/ horrendous babysitter she is.
8	Fred and Sally volunteered to take care of their neighbor's dog while he was away. Fred/Sally walked and fed the dog three times a day. Fred/Sally forgot to walk or feed the dog for several days.	What an upstanding/ coldhearted neighbor he is.
9	Fred and Sally decided to start a vegetable garden. Fred/Sally watered the plants every day and he/she grew many vegetables. Fred/Sally forgot to water the plants and they all died.	What an excellent/ terrible gardener he is.
10	Fred and Sally had to write a report for work.	What an exemplary/

	Fred/Sally stayed in his/her office until he/she was finished. Fred/Sally spent the day playing computer games instead.	awful worker she is.
11	Fred and Sally took pictures at their friend's baseball game. Fred/Sally got a picture of his/her friend hitting a home run. Fred/Sally only got pictures of the grass.	What an extraordinary/ atrocious photographer he is.
12	Fred and Sally decided to buy their children a pet. Fred/Sally bought his/her child a fish. Fred/Sally bought his/her child a wolf.	What a responsible/ horrendous parent she is.
13	Fred and Sally took care of their friend's house while he was away. Fred/Sally mowed the lawn and took in the mail every day. Fred/Sally forgot to mow the lawn and take in the mail the entire time.	What a responsible/ terrible friend she is.
14	Fred and Sally went on a weekend trip. Fred/Sally fir everything he/she needed in one suitcase. Fred/Sally filled up three entire suitcases.	What a terrific/awful traveler he is.
15	Fred and Sally's roofs were damaged in a storm and they decided to repair them. Fred/Sally repaired the roof so that it was as good as new. Fred/Sally tried to repair the roof but fell and broke his/her arm.	What a remarkable/ bumbling homeowner she is.
16	Fred and Sally worked at a newspaper and had to write many articles. Fred/Sally concentrated hard and wrote all of the articles before the deadline. Fred/Sally didn't write any articles by the deadline.	What a phenomenal/ worthless journalist he is.
17	Fred and Sally participated in a food drive. Fred/Sally donated an entire box full of food. Fred/Sally didn't even fill a single box with food.	What an outstanding/ self-centered citizen he is.
18	Fred and Sally constructed a tree house. Fred's/Sally's tree house lasted for several years. Fred's/Sally's tree house collapsed after one day.	What a fantastic/ incompetent builder she is.
19	Fred and Sally decided to do some baking. Fred/Sally baked three beautiful cakes. Fred/Sally only made one cake and he/she burned it.	What an amazing/ dreadful chef he is.
20	Fred and Sally played a game of soccer with their friends. Fred/Sally scored five goals. Fred/Sally kicked the soccer ball and his/her shoe fell off.	What a gifted/ hopeless athlete she is.

Appendix C

Instructions for Participants in Experiment 2B

Speaker Group A (literal and ironic)

Note: half of the participants in Group A saw Lucy and Ike, the other half saw Luke and Irene

This study consists of three parts. In Part 1, you are going to watch 8 brief videos involving different characters. Each character will perform an action. After the characters perform these actions, one of two speakers will describe one of the characters. One speaker, Literal Lucy/Literal Luke, always says what he/she means. The other character, Ironic Ike/Ironic Irene, is always ironic. Note that Literal Luke/Ironic Ike is male and Literal Lucy/Ironic Irene is female.

In Part 2 of the study, you're going to be asked several questions about the two speakers, Literal Lucy/Literal Luke and Ironic Ike/Ironic Irene. Please answer these questions based on the videos you watched in Part 1. Finally, in Part 3, you'll be asked a few optional demographic questions.

Speaker Group B (literal and opposite)

Note: half of the participants in Group B saw Lucy and Ollie, the other half saw Luke and Olive

This study consists of three parts. In Part 1, you are going to watch 8 brief videos involving different characters. Each character will perform an action. After the characters perform these actions, one of two speakers will describe one of the characters. One speaker, Literal Lucy/Literal Luke, always says what he/she means. The other character, Opposite Ollie/Opposite Olive, is always ironic. Note that Literal Luke/Opposite Ollie is male and Literal Lucy/Opposite Olive is female.

In Part 2 of the study, you're going to be asked several questions about the two speakers, Literal Lucy/Literal Luke and Opposite Ollie/Opposite Olive. Please answer these questions based on the videos you watched in Part 1. Finally, in Part 3, you'll be asked a few optional demographic questions.

Appendix D

Critical Items Used in Experiments 3 and 4

Item	Critical Utterance
1	What a splendid planner she is.
2	What a remarkable golfer he is.
3	What an incredible swimmer he is.
4	What an accomplished thrower he is.
5	What an excellent driver he is.
6	What a magnificent speaker she is.
7	What a compassionate babysitter she is.
8	What an upstanding neighbor she is.
9	What an admirable gardener she is.
10	What an exemplary worker she is.
11	What an extraordinary photographer she is.
12	What a responsible parent he is.
13	What a reliable friend she is.
14	What a terrific librarian he is.
15	What a savvy homeowner she is.
16	What a dedicated journalist he is.
17	What an outstanding citizen he is.
18	What a fantastic builder she is.
19	What an amazing chef he is.
20	What a gifted athlete she is.
21	What a wonderful artist she is.
22	What an effective salesperson he is.
23	What an exceptional performer he is.
24	What a commendable Bee Keeper she is.
25	What a lucky gambler he is.
26	What a graceful ice skater he is.
27	What an athletic hiker he is.
28	What a skilled dancer she is.
29	What a supportive teammate he is.
30	What a spectacular cook she is.
31	What a strong gym-goer he is.
32	What a knowledgeable teacher she is.
33	What an impressive bowler he is.
34	What a fast runner she is.
35	What a marvelous singer she is.

36	What a hardworking employee he is.
37	What an attentive spy she is.
38	What a talented surfer he is.
39	What an observant manicurist he is.
40	What a terrific folder he is.
41	What a successful farmer she is.
42	What a phenomenal fisher she is.
43	What a handy mechanic she is.
44	What a prepared host he is.
45	What a superb student he is.
46	What a qualified captain she is.
47	What a skilled magician he is.
48	What a responsible doctor she is.

Bibliography

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. http://doi.org/10.1006/jmla.1997.2558

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191-238.

Altmann, G., Garnham, A., & Dennis, Y. (1992). Avoiding the garden path: Eye movements in context. *Journal of Memory and Language, 31,* 685-712.

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). 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.

Arnold, J. E., Pancani, G. C., & Rosa, E. (under revision). Listeners perceived acoustic prominence differently for distracted and fluent speakers.

Attardo, S. (2001). Humor and irony in interaction: From mode adoption to failure of detection. In L. Anolli, R. Ciceri, & G. Riva (Eds.), *Say not to say: New perspectives on miscommunication* (pp. 165–185). IOS Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=PsiLjRHr1JQC&oi=fnd&pg=PA159&dq=Humor+and+irony+in+interaction:+From+mode+adoption+to+failure+of+detection&ots=GllbNjOOh4&sig=4Wo1UBWxWZhvEzaPJk968Gm-lwI

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor, 16*(2), 243-260.

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*(3), 1-43.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Listening, Memory, and Cognition, 38*(5), 1450-1460.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51*, 437-457.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652. http://doi.org/10.1037//0033-295X.I08.3.624

Boylan, J., & Katz, A. B. (2013). Ironic expression can simultaneously enhance and dilute perception of criticism. *Discourse Processes, 50*, 187-209.

Brown, P. (1995). Politeness strategies and the attribution of intentions: The case of Tzeltal irony. In E. Goody (Ed.), *Social intelligence and interaction* (pp. 153–174). Cambridge University Press.

Brown-Schmidt, S., & Duff, M. C. (2016). Memory and common ground processes in language use. *Topics in Cognitive Science*, *8*(4), 722–736.

Brown-Schmit, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition, 107*, 1122-1134.

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

Clark, H. H., & Marshall, C. R. (1978). Reference diaries. In D. L. Waltz (Ed.), *Theoretical issues in natural language processing*. Vol. 2 (pp. 57–63). New York: Association for Computing Machinery.

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

Cutler, A. (1976). Beyond parsing and lexical look-up: An enriched description of auditory sentence comprehension. In R. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 133-150). Amsterdam: North-Holland.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*, 317-367.

Dews, S., & Winner, E. (1995). Muting the meaning: A social function of irony. *Metaphor and Symbol, 10*(1), 3–19. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15327868ms1001_2

Dews, S., & Winner, E. (1999). Obligatory processing of literal and nonliteral meanings in verbal irony. *Journal of Pragmatics*, *31*, 1579–1599. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378216699000053

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? The social

functions of irony. *Discourse Processes, 19*(3), 347-367.

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.

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. http://doi.org/10.1037/a0035658

Filik, R., & Moxey, L. M. (2010). The on-line processing of written irony, *Cognition, 116*, 421-436.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 627-635.

Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General, 115*(1), 3–15. http://doi.org/10.1037//0096-3445.115.1.3

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and Symbol, 15*(1–2), 5–27. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/10926488.2000.9678862

Gibbs, R. W., Bryant, G. A., & Colston, H. L. (2014). Where is the humor in verbal irony? *Humor, 27*(4), 575-595.

Gibbs, R. W., Kushner, J. M, & Mills, W. R. (1991). Authorial intentions and metaphor comprehension. *Journal of Psycholinguistic Research, 20*(1), 11-30.

Gilovich, T., Medvec, V. H., & Savitsky, K. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology, 75*(2), 332-346.

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

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics, 8*(3), 183-206.

Giora, R., Fein, O., Laadan, D., Wolfson, J., Zeituny, M., Kidron, R., … Shaham, R. (2007). Expecting irony: Context versus salience-based effects. *Metaphor and Symbol, 22*(2), 119–146. http://doi.org/10.1080/10926480701235346

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic.

Grodner, D., & Sedivy, J. (in press). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter & E. Gibson (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.

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

Horton, W. S. (2007). The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and Cognitive Processes, 22*, 1114–1139.

Horton, W. S., & Gerrig, R. J. (2016). Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science, 8*(4), 780-795.

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. http://doi.org/10.1177/0956797615625223

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415. http://doi.org/10.1016/j.cogpsych.2008.09.001

Huang, Y. T., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, *26*(8), 1161–1172. http://doi.org/10.1080/01690965.2010.508641

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

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, *21*(12), 2434–2444.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Jorgensen, J. (1996). The functions of sarcastic irony in speech. *Journal of Pragmatics, 26*(5), 613–634. http://doi.org/10.1016/0378-2166(95)00067-4

Kan, I. P., & Thompson-Schill, S. L. (2004). Selection from perceptual and conceptual representations, *Cognitive, Affective, and Behavioral Neuroscience, 4*(4), 466-482.

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*, 637–651.

Katz, A. N., & Pexman, P. M. (1997). Interpreting figurative statements: Speaker occupation can change metaphor to irony. *Metaphor and Symbol, 12*(1), 19-41.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32-38.

Kreuz, R. J., Kassler, M. A., Coppenrath, L., & Allen, B. M. (1999). Tag questions and common ground effects in the perception of verbal irony. *Journal of Pragmatics, 31*, 1685-1700.

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

Kuchinke, L., Jacobs, A. M., Grubich, C., Võ, M. L. H., Conrad, M., & Herrmann, M. (2005). Incidental effects of emotional valence in single word processing: An fMRI study. *NeuroImage*, *28*(4), 1022–1032. http://doi.org/10.1016/j.neuroimage.2005.06.050

Kumon-Nakamura, S., Glucksberg, S., & Brown, M. (1995). How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General, 124*(1), 3-21.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2015). lmerTest: Tests in linear mixed effects models. R package version 2.0–29. https://cran.r- project.org/web/packages/lmerTest/index.html.

Lakoff, R. T. (1990). *Talking power: The politics of language*. New York, NY: Basic Books.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*, 551-556.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676-703.

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

Matthews, J. K., Hancock, J. T., & Dunham, P. J. (2006). The roles of politeness and humor in the asymmetry of affect in verbal irony. *Discourse Processes, 41*(1), 3-24.

Milham, M. P., Banich, M. T., Claus, E. D., & Cohen, N. J. (2003). Practice-related effects demonstrate complementary roles of anterior cingulate and prefrontal cortices in attentional control. *NeuroImage, 18*(2), 483-493.

Milham, M. P., Banich, M. T., Webb, A., Barad, V., Cohen, N. J., Wszalek, T., et al. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cognitive Brain Research, 12*(3), 467–473.

Nappa, R., & Arnold, J. E. (2014). The road to understanding is paved with the speaker's intentions: Cues to the speaker's attention and intentions affect pronoun comprehension. *Cognitive Psychology, 70*, 58-81.

Novick, J. M., Kan, I. P., Trueswell, J. C., & Thompson-Schill, S. L. (2009). A case for conflict across multiple domains: Memory and language impairments following damage to ventrolateral prefrontal cortex. *Cognitive Neuropsychology, 26*(6), 527–67. http://doi.org/10.1080/02643290903519367

Novick, J. M., Kim, A., & Trueswell, J. C. (2003). Studying the grammatical aspects of word recognition: Lexical priming, parsing, and syntactic ambiguity resolution. *Journal of Psycholinguistic Research, 32*(1), 57-75.

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

Pexman, P., & Zvaigzne, M. (2004). Does irony go better with friends? *Metaphor and Symbol*, *19*(2), 143–163. http://doi.org/10.1207/s15327868ms1902

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rayner, K.l, & Raney, G. E. (1996). Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review, 3*(2), 245-248.

Regel, S., Coulson, S., & Gunter, T. C. (2010). The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain Research*, *1311*, 121–35.

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of*

Psycholinguistic Research, 29(5), 483-495.

Sally, D. (2003). Risky speech: Behavioral game theory and pragmatics. *Journal of Pragmatics*, *35*(8), 1223–1245. http://doi.org/10.1016/S0378-2166(02)00170-4

Schacht, A., & Sommer, W. (2009). Time course and task dependence of emotion effects in word processing. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 28–43. http://doi.org/10.3758/CABN.9.1.28

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. http://doi.org/10.1080/10926488.2000.9678864

Searle, J. R. (1979). Metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 83–111). Cambridge, UK: Cambridge University Press.

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.

Spivey-Knowlton, M. J., & Sedivy, J. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition, 55*, 227-267.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-659.

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*, 1632-1634.

Thomson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role

of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *PNAS, 94*(26), 14792-14797.

Tomlinson Jr., J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, *69*(1), 18–35. http://doi.org/10.1016/j.jml.2013.02.003

Trueswell, J .C., & Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of Memory and Language, 29*, 102-103.

Trueswell, J. C., & Tanenhaus, M. K. (1991). Tense, temporal context and syntactic ambiguity resolution. *Language and Cognitive Processes, 6*, 303-338.

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.

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

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143.